UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Optimization and improvements in spatial sound reproduction systems through perceptual considerations

Doctoral Thesis

by

Pablo Gutiérrez Parera

Supervisor:

Prof. José Javier López Monfort

Valencia, Spain
March 2020

"Los sistemas monoaurales, los biaurales poli y perifónicos y principalmente los estereofónicos, tienden al retrato físico de los sonidos y a su relieve físico, a provocar la ilusión del espacio y sus desplazamientos en éste. [...] La línea que atraviesa nuestros oídos es el eje por donde se mueve nuestro tacto acústico. [...]
[Sin embargo] los sonidos [diafónicos] no se apoyan en la representación del espacio físico que rodea al espectador, todo queda reducido a una sensación externa del mundo donde vivimos convertida en contrapunto psíquico. La línea estereofónica es primaria, sensorial, física, apetecida por el pueblo joven. La línea diafónica subraya nuestra proyección sentimental, es flor del viejo mundo. [...]
Nos encontramos incorporados a un juego mecánico invisible espacio-temporal complejo biológico, caminando hacia la Unidad centrífuga. [...]
El que más da más tiene. Matemáticas de Dios. El que más da más tiene. El que más da, más tiene. Más tiene."

"The monaural systems, the binaural poly and peripheral and mainly the stereophonic ones, tend to the physical portrait of the sounds and their physical relief, to provoke the illusion of the space and its displacements on it. [...] The line that runs through our ears is the axis along which our acoustic touch moves. [...]
[However] the [diaphonic] sounds are not based on the representation of the physical space that surrounds the spectator, everything is reduced to an external sensation of the world where we live, converted into a psychic counterpoint. The stereophonic line is primary, sensory, physical, desired by the young people. The diaphonic line underlines our sentimental projection, it is a flower of the old world. [...]
We find ourselves incorporated into an invisible space-time mechanical biological complex game, walking towards the centrifugal Unity. [...]
The one who gives more has more. Mathematics of God. The one who gives more has more. The one who gives more has more. Has more."

Escritos de Técnica, Poética y Mística.
José Val del Omar

Abstract

The reproduction of the spatial properties of sound is an increasingly important concern in many emerging immersive applications. Whether it is the reproduction of audiovisual content in home environments or in cinemas, immersive video conferencing systems or virtual or augmented reality systems, spatial sound is crucial for a realistic sense of immersion. Hearing, beyond the physics of sound, is a perceptual phenomenon influenced by cognitive processes. The objective of this thesis is to contribute with new methods and knowledge to the optimization and simplification of spatial sound systems, from a perceptual approach to the hearing experience. This dissertation deals in a first part with some particular aspects related to the binaural spatial reproduction of sound, such as listening with headphones and the customization of the Head Related Transfer Function (HRTF). A study has been carried out on the influence of headphones on the perception of spatial impression and quality, with particular attention to the effects of equalization and subsequent non-linear distortion. With regard to the individualization of the HRTF a complete implementation of a HRTF measurement system is presented, and a new method for the measurement of HRTF in non-anechoic conditions is introduced. In addition, two different and complementary experiments have been carried out resulting in two tools that can be used in HRTF individualization processes, a parametric model of the HRTF magnitude and an Interaural Time Difference (ITD) scaling adjustment. In a second part concerning loudspeaker reproduction, different techniques such as Wave-field Synthesis (WFS) or amplitude panning have been evaluated. With perceptual experiments it has been studied the capacity of these systems to produce a sensation of distance, and the spatial acuity with which we can perceive the sound sources if they are spectrally split and reproduced in different positions. The contributions of this research are intended to make these technologies more accessible to the general public, given the demand for audiovisual experiences and devices with increasing immersion.

Keywords: Spatial sound, perception of sound, perceptual test, headphones, binaural, HRTF, Wave-field Synthesis, VBAP.

Resumen

La reproducción de las propiedades espaciales del sonido es una cuestión cada vez más importante en muchas aplicaciones inmersivas emergentes. Ya sea en la reproducción de contenido audiovisual en entornos domésticos o en cines, en sistemas de videoconferencia inmersiva o en sistemas de realidad virtual o aumentada, el sonido espacial es crucial para una sensación de inmersión realista. La audición, más allá de la física del sonido, es un fenómeno perceptual influenciado por procesos cognitivos. El objetivo de esta tesis es contribuir con nuevos métodos y conocimiento a la optimización y simplificación de los sistemas de sonido espacial, desde un enfoque perceptual de la experiencia auditiva. Este trabajo trata en una primera parte algunos aspectos particulares relacionados con la reproducción espacial binaural del sonido, como son la escucha con auriculares y la personalización de la Función de Transferencia Relacionada con la Cabeza (*Head Related Transfer Function* - HRTF). Se ha realizado un estudio sobre la influencia de los auriculares en la percepción de la impresión espacial y la calidad, con especial atención a los efectos de la ecualización y la consiguiente distorsión no lineal. Con respecto a la individualización de la HRTF se presenta una implementación completa de un sistema de medida de HRTF y se introduce un nuevo método para la medida de HRTF en salas no anecoicas. Además, se han realizado dos experimentos diferentes y complementarios que han dado como resultado dos herramientas que pueden ser utilizadas en procesos de individualización de la HRTF, un modelo paramétrico del módulo de la HRTF y un ajuste por escalado de la Diferencia de Tiempo Interaural (*Interaural Time Difference* - ITD). En una segunda parte sobre reproducción con altavoces, se han evaluado distintas técnicas como la Síntesis de Campo de Ondas (*Wave-field Synthesis* - WFS) o la panoramización por amplitud. Con experimentos perceptuales se han estudiado la capacidad de estos sistemas para producir sensación de distancia y la agudeza espacial con la que podemos percibir las fuentes sonoras si se dividen espectralmente y se reproducen en diferentes posiciones. Las aportaciones de esta investigación pretenden hacer más accesibles estas tecnologías al público en general, dada la demanda de experiencias y dispositivos audiovisuales que proporcionen mayor inmersión.

Palabras Clave: Sonido espacial, percepción del sonido, test perceptual, auriculares, binaural, HRTF, Wave-field Synthesis, VBAP.

Resum

La reproducció de les propietats espacials del so és una qüestió cada vegada més important en moltes aplicacions immersives emergents. Ja siga en la reproducció de contingut audiovisual en entorns domèstics o en cines, en sistemes de videoconferència immersius o en sistemes de realitat virtual o augmentada, el so espacial és crucial per a una sensació d'immersió realista. L'audició, més enllà de la física del so, és un fenomen perceptual influenciat per processos cognitius. L'objectiu d'aquesta tesi és contribuir a l'optimització i simplificació dels sistemes de so espacial amb nous mètodes i coneixement, des d'un criteri perceptual de l'experiència auditiva. Aquest treball tracta, en una primera part, alguns aspectes particulars relacionats amb la reproducció espacial binaural del so, com són l'audició amb auriculars i la personalització de la Funció de Transferència Relacionada amb el Cap (*Head Related Transfer Function* - HRTF). S'ha realitzat un estudi relacionat amb la influència dels auriculars en la percepció de la impressió espacial i la qualitat, dedicant especial atenció als efectes de l'equalització i la consegüent distorsió no lineal. Respecte a la individualització de la HRTF, es presenta una implementació completa d'un sistema de mesura de HRTF i s'inclou un nou mètode per a la mesura de HRTF en sales no anecoiques. A mès, s'han realitzat dos experiments diferents i complementaris que han donat com a resultat dues eines que poden ser utilitzades en processos d'individualització de la HRTF, un model paramètric del mòdul de la HRTF i un ajustament per escala de la Diferència del Temps Interaural (*Interaural Time Difference* - ITD). En una segona part relacionada amb la reproducció amb altaveus, s'han avaluat distintes tècniques com la Síntesi de Camp d'Ones (*Wave-field Synthesis* - WFS) o la panoramització per amplitud. Amb experiments perceptuals, s'ha estudiat la capacitat d'aquests sistemes per a produir una sensació de distància i l'agudesa espacial amb que podem percebre les fonts sonores, si es divideixen espectralment i es reprodueixen en diferents posicions. Les aportacions d'aquesta investigació volen fer més accessibles aquestes tecnologies al públic en general, degut a la demanda d'experiències i dispositius audiovisuals que proporcionen major immersió.

Paraules Clau: So espacial, percepció del so, test perceptual, auriculars, binaural, HRTF, Wave-field Synthesis, VBAP.

Acknowledgements

This thesis work was carried out at the Institute of Telecommunications and Multimedia Applications ITEAM of the Universitat Politècnica de València, Spain. The research leading to these results has received funding from the Spanish Ministry of Economy and Competitiveness under the training grants for research staff programme (State Training Subprogramme call 2013) with the grant awarded BES-2013-065034, as well as with funds for mobility (State Mobility Subprogramme calls 2015 and 2016) with the grants awarded EEBB-I-16-11288 and EEBB-I-17-12527.

I would like to express my gratitude to many people I have met over the years in the realization of this thesis. With them I have learned, suffered, enjoyed and shared many experiences and moments. Probably I will forget to name some people in this list but I am equally grateful to all of them.

First I want to thank my director José Javier López Monfort for the trust he has placed in me, his guidance, collaboration and patience during these years. This work would not have been done without his support.

I am also thankful to the pre-evaluators of the thesis manuscript, Felipe Orduña Bustamante of the Universidad Nacional Autónoma de México and Marcos F. Simón Gálvez of the University of Southampton, for their prompt reviews and gentle comments.

The Audio and Communications Signal Processing Group (GTAC), in which I belonged, made our workplace a very pleasant and rewarding environment. My gratitude extends to all the people that passed from it during my stay. I would like to thank especially Laura Fuster (the boss), Emanuel Aguilera, José Antonio Belloch, Marian Simarro, Christian Antoñanzas, Juan Estreder, Gabi Moreno, Fabián Aguirre, Enrique Palazón, Vicent Moles, as well as the professors Gema Piñero, Miguel Ferrer, María de Diego and Germán Ramos, and also the former member and continuous collaborator from the Universitat de València Máximo Cobos.

I also want to thank professor Ville Pulkki from Aalto University in Finland for hosting me, opening a door for me to learn about spatial sound and Finnish culture, and for singing me traditional songs. Thanks to the people from the Communication Acoustics research group and related groups I knew there: Archontis Politis, Symeon Delikaris-Manias, Alessandro Altoè, Ilkka Huhtakallio, Catarina Mendoça, Mika Iivonen, Leo Mccor-

mack, Marko Hiipakka, Fabián Esqueda, Sofoklis Kakouros, and of course the professors Vesa Valimaki and Lauri Savioja. Javier Gómez Bolaños deserves a special mention for all that he taught me with dedication and patience. I usually consider that the work of my thesis kicked off from the stay I spent with all of them.

My stay in the EURECAT Multimedia Technologies group in Barcelona was possible thanks to Adán Garriga who gave me the opportunity and confidence to exchange knowledge. I learned a lot with Julien de Muynke, Umut Sayin, Niklas Reppel, Toni Farran, Tim Schemele to whom I am grateful, and especially to Andres Pérez López for his participation in some of the experiments presented in the thesis and his generosity to share ideas. The experience I had with them to fluently auralize the result of my measurements and processes gave me back my confidence in binaural sound.

It is very necessary to recognize here the help in the last experiment and the encouragement in the completion of the thesis of Diego Larios, Enrique Personal and Javier Mora from the University of Seville. They have given unexpected support and generous friendship.

All the people (more than a hundred) who have volunteered to take the various perceptual tests and measurements contained in this work are also an essential part of it.

My parents, brother and sister have supported me unconditionally all these years and I am very grateful to them for that. And finally, I am most grateful to my wife Berta who stood by my side throughout this work and my one-year-old daughter Mara. They are the muses that give meaning to this work.

Pablo Gutiérrez Parera
Valencia, March 2020

Contents

List of Figures

1 Introduction

As listeners we are used to perceive sound with its natural spatial field characteristics. These provide us with one of the basic functions of the sense of hearing which is the perception of the surrounding environment. According to the classical Greek enumeration of the senses (sight, hearing, taste, smell and touch), hearing and vision are the only two senses that give us spatial information beyond the limits of our own body. So we are very sensitive to the spatial dimension especially with the sense of hearing, because it is able to provide us with continuous omnidirectional information. Adding a spatial dimension to sound reproduction generates a great sense of reality since it reconstructs or simulates a fundamental characteristic of sound which is the spatial field condition [1, 2]. Thanks to this we are able to perceive the surroundings and our spatial position within it. Therefore, the reproduction of sound with spatial attributes results in a dramatic enhancement of the listening experience. Although popularly this concept is not clear, probably because of the lack of direct experience with spatial sound systems, the audiovisual industry has been working on this idea for decades.

For almost a century, sound reproduction technology has been developing spatial sound, also known as 3D sound. In 1931 Alan Blumlein invented the stereophony as we know it today [3], which opened up a new dimension

on the listening experience and revolutionized the music and sound production industry. Shortly thereafter, cinema took over the development of spatial sound to incorporate the sensation of stereophony in theaters. Disney with its *Fantasound* was a practical and conceptual milestone, which together with the development of recording and playback technologies led to the blooming of a variety of multichannel sound formats. Since the 1950s, every new sound system for cinemas incorporated substantial technical improvements, including a greater number of audio channels (Cinerama, Cinemascope, Perspecta sound, Todd-AO and others). Important advances were also made in the 1960s and 1970s with the popularization of two-channel stereo and the failed attempt of Quadraphony in the domestic scene. In cinemas, Dolby, DTS and Sony continued the multi-channel development until the introduction of digital sound in the 1990s. In this technological escalation [4] there was always a motivation to improve the richness of sound reproduction with special emphasis on the spatiality of sound. Not surprisingly, this process is similar to which the history of music has followed. Throughout the development of Western music many composers brought forth innovation through the introduction of new instruments and different orchestra configurations, with attention also to the spatiality of sound. In this process a parallelism or even a continuation can be seen in the techniques of recording and reproduction of spatial sound. Nothing better than the following quote from Tomlinson Holman to expose this idea: "I often ask my students to think about the following progression: Mozart, Beethoven, Mahler, *Terminator 2*. This progression makes sense in the following way: each step increases the use of frequency range, dynamic range, and especially spatialization of sound. From Mozart to Beethoven we see larger and louder forces at play, but still seated as a conventional orchestra. Mahler breaks the front orchestra mode with off-stage sound. *Terminator 2* carries this thrust on, with even greater emphasis on surround sound" [5].

The innovation in this field has not only concerned the speaker configuration of the system, but also the reproduction techniques. This has led to the development and improvement of spatial positioning algorithms, such as the three-dimensional generalization of amplitude panning with Vector Base Amplitude Panning (VBAP) [6]. Sound field synthesis is also the subject of new studies, especially Wave Field Synthesis (WFS) [7, 8] and recording and reproduction techniques using decomposition in spherical harmonics (Ambisonics) [9, 10]. On the other hand, in the field of spatial sound cod-

ing, the concept of sound objects (SAOC) [11] has been introduced, which allows to break free from the channel sound mixing paradigm by directly coding each sound source and adding spatial positioning metadata.

Headphone listening has also had a great development in spatial sound, being currently one of the most innovative fields, both in research and in commercial products [12, 13]. Binaural systems based on headphones have attracted much interest both because of the privacy of listening they provide in any type of environment and because of the proliferation of mobile devices that make use of headphones. Binaural sound reproduction through headphones uses the principles of the human hearing system based on two ears. It is founded on the following principle: if we are able to reproduce in the listener's ears through headphones the same sound pressures that the listener would experience in reality, then we can simulate a realistic acoustic immersion [2]. In addition, the reproduction of spatial sound through headphones can also benefit from the technologies discussed above, such as Ambisonics or sound object coding.

Higher definition resolutions and 3D image, mobile and wearable devices are fostering the development and implementation of spatial sound technologies [14]. There is a trend towards an increasing immersion [15, 16] that comes to cover new needs beyond the audiovisual industry. Telepresence, Virtual Reality or Augmented Reality systems demand a recreation of reality in conjunction with interactivity, which makes the application of spatial sound technologies essential [12, 17]. For these reasons, the techniques for capturing, synthesising and reproducing spatial sound are still being developed today to provide users with more complete experiences.

1.1 Motivation

In the technological evolution of sound systems, understanding the mechanisms of perception has always been important. On the one hand to build systems that are more suitable for human listening and on the other hand to understand the human perception of sound itself. If we consider the study of sound, physics is not the same as perception. Sound is a physical phenomenon whereby vibrations are transmitted through a medium by pressure waves. However, our experience of that physical phenomenon is something else. The perception of sound is the result of the excitation of

our auditory system by these pressure waves. The physics of sound and the perception of sound are not the same thing. Our perception of sound is influenced by a complex cognitive process for which there is no completely consensual model. The auditory system introduces its own peculiarities, not described in the physical acoustics. Since sound reproduction systems are mainly constructed to be heard by ourselves, it is interesting to approach sound reproduction not only from a physical but also from a perceptual point of view. This is the idea that has motivated the investigation of this work, with the conviction that attending to the perception of sound it is possible to improve and optimize the spatial sound systems, both to obtain a better reproduction and to improve the efficiency or accessibility of these systems. Hence, this work is oriented with a perceptual approach to the listening experience and particularly focused to the spatial impression of the acoustic experience.

This dissertation deals in a first part with some particular aspects related to the binaural spatial sound reproduction, such as listening with headphones and the personalization of the Head Related Transfer Function (HRTF). In a second part about loudspeaker reproduction, certain questions concerning distance perception and spatial acuity are studied.

Binaural reproduction, despite having a great potential, presents a series of relatively important problems for its implementation. The listening experience through headphones is influenced by the response of each actual headphone model, generating undetermined effects on spatial perception and therefore difficult to correct. As listening through headphones is becoming popular and expanding considerably, the perceptual effects that can be introduced not only by high-end headphones but also by consumer models are of increasing interest to popularize the binaural spatial sound. Additionally, the most challenging problem of binaural sound is the individualization of the HRTF. Each individual has learned to hear with their own morphology (that determines their own HRTF) and by listening using a different HRTF than their own, the spatial cues received change, generating an incongruent perception for the individual that translates into different types of errors. Although the use of generic HRTF can generate a certain spatial impression, individualized HRTFs are needed for accurate recreation of reality. Obtaining such individualized HRTFs for each person either by measurement or adaptation of a generic HRTF involves many difficulties.

Spatial sound reproduction using loudspeakers has different problems from those of binaural sound, some of which are the reduced optimal listening area, the perception of distance or the large number of speakers that may be needed. Distance perception is considered a key issue in spatial sound and different techniques of spatial sound reproduction through speakers can generate different perceptions. Also, while WFS can expand the listening area with respect to VBAP and Ambisonics, it has the problem that it requires many speakers. It is interesting to study the effect that a limitation on the number of loudspeakers in a complex system may introduce. Manufacturers propose systems with speakers spaced arbitrarily or by trial and error, without knowing for sure the effects on the final quality and the artifacts that may appear.

The aim of helping to resolve these different problems by focusing on a perceptual perspective motivated the research developed in this thesis.

The research was carried out following the main goals set in the projects SoundAction (Procesado de sonido para la interacción hombre-máquina en entornos acústicos complejos, TEC2012-37945-C02) and PersonalSound (Técnicas de optimización y personalización de sonido inmersivo para el gran público, TEC2015-68076-R) which were both supported by the Spanish Ministry of Economy and Competitiveness.

1.2 Objectives

Taking into account the above context, the main objective of this thesis is as follows:

To contribute new methods and knowledge for the optimization and simplification of spatial sound, both for reproduction through headphones and by means of loudspeakers, in order to make these technologies more accessible to the population. To test different sound reproduction systems and techniques from a perceptual point of view, in the search for key features that can be improved for a natural and realistic spatial sound perception. With the acquired knowledge, propose and implement new solutions and post-processing methods that can on the one hand improve the perception or on the other hand simplify the complexity of spatial sound systems.

Some particular aims emerge from this main scope, which are presented

as follows:

– To study the capacities, suitability and problems of consumer headphone models for the proper reproduction of spatial perception.

– To implement a viable HRTF measurement system, overcoming or proposing solutions for inherent difficulties such as long measuring time, precision or room reverberation, that allows for the study of real individual HRTF measures.

– To evaluate the perceptual differences of the HRTF between different people and to propose new methods for the individualization of the HRTF.

– To examine the capacities or limitations of some loudspeaker reproduction systems to generate perceptual attributes such as distance or spatial acuity, in order to propose specific uses or simplifications of the reproduction methods.

1.3 Organization of the thesis

The body of this thesis describes the research that has been undertaken to develop the aims stated above. Since the contributions of this thesis are framed within two different approaches to sound reproduction, which are binaural systems using headphones and spatial reproduction using arrays of loudspeakers, the content of this dissertation has been structured in two parts. Note that this two-part division has not been applied to the introductory, background and the concluding chapters.

The chapters are then organized and presented as follows:

– Chapter 2 - Background: In this chapter, an overview of spatial sound reproduction systems and techniques are provided, as well as the principles of spatial hearing on which they are based. Stereo, multichannel surround systems, binaural and other advanced techniques based on sound field rendering are presented, with special emphasis in the fundamentals of binaural technology.

Part I: Headphones and Binaural systems

– Chapter 3 - Effects of Headphones in perception: This chapter studies the influence of headphones on the perception of spatial impression and

quality, taking into account medium and low quality range headphones. Special attention is given to the effects of equalization and the perception of non-linear distortion.

– Chapter 4 - HRTF measurements: The purpose of this chapter is first to provide an overview of HRTF measurement systems, at the same time that a complete implementation of a real built measurement system is presented, and then to introduce a new method for HRTF measurement in non-anechoic rooms.

– Chapter 5 - HRTF individualization tools: Two different and complementary experiments are presented here to approach the individualization of the HRTF, one that manipulates the magnitude of the HRTF and other about adjustments of the ITD. As a result, two tools have been obtained that can be used in HRTF individualization processes.

Part II: Loudspeaker based systems

– Chapter 6 - Distance perception comparison between WFS and VBAP: This chapter presents an evaluation of the perception of distance with Wave Field Synthesis and Vector Base Amplitude Panning methods, with the aim to explore the capacities of these different rendering techniques to generate the sensation of depth.

– Chapter 7 - Perceptual spatial acuity of spectrally divided sound sources: The problem of the number of speakers and the acuity to perceive different sound sources is investigated here through a series of experiments that present sound sources split into two different frequency bands that are reproduced spatially separated. The findings can be applied in a reduction and optimization of the quantity of loudspeakers in complex systems.

– Chapter 8 - Applications and future work: Finally, the conclusions obtained throughout this thesis are presented, including some guidelines for future research lines. A list of published work related to this thesis is also given.

Since the orientation of the work is based on the analysis of auditory perception, the basic work tool used is the perceptual test [18]. In each chapter different experiments based on perceptual tests are presented to investigate the different questions.

The main background and contextual references are presented in chapter

2, but due to the wide variety of experiments and specific topics covered in this thesis, a motivation and a specific state of the art are presented in the introduction of each chapter or section. This arrangement of the information is made for the sake of clarity and ease of understanding and reading of the whole text. For the same reason, each main section includes its own conclusions and future work part.

2 Background

The objective of spatial sound systems is to accurately recreate the acoustic sensations that a listener would naturally perceive from the surrounding environment, for example the certain acoustic properties inside a particular room or in any other place. This concept implies a series of physical and technological difficulties that are a current research issue in sound engineering. Two channel stereo sound systems, considered as the simplest approximation to spatial sound, have been utilized throughout the last 60 years as an added value in sound recordings, specially for music material. Together with the entertainment industry, spatial sound evolved to multichannel surround sound systems, which provide a better sensation by using more reproduction channels. Nowadays, the most promising systems for spatial sound reproduction are those based on providing the binaural sound signals directly to each ear, or the ones based on acoustic field synthesis. In this chapter, the basics underlying all these audio systems are described, as well as the essentials of spatial hearing perception.

2.1 Principles of spatial hearing

Humans and other animal species have the remarkable ability to identify the direction of a sound source originating from any point in three-dimensional space, as well as the sound spatial attributes of the environment. This ability of humans listeners to perceive the location (direction and distance), the spaciousness of sound sources, and the acoustic properties of the environment is called "spatial hearing" [1, 2]. Spatial hearing helps to interact with the surroundings more effectively and also allows the listener to focus and concentrate on an individual audio source at a particular position in the space. The human auditory system is very sophisticated and therefore capable to analyze and extract most of the spatial information related to a sound source using two ears. Besides, the process of locating a sound source is dynamic and is often aided and complemented by other sensory inputs.

Past studies suggest that an average human listener can localize a sound source in space very precisely and accurately. The precision and resolution with which the sound sources can be localized are different for the horizontal and elevation planes. For example, in the horizontal planes, a human listener can distinguish between the direction of arrival for a sound signal with a resolution of 1° to 3° on average. On the other hand, localization of the sound source in the elevation plane is more difficult, and depending on the sound source properties and stimuli, humans can only localize a sound source in elevation plane with an average resolution of 4° (for white noise) and 17° (for speech stimuli) [19].

Spatial hearing does not only provide punctual localization positions of sound sources. It also provides with a spatial impression that collects information about the size, shape and type of environment and source. The concept of spatial impression can be decomposed into particular spatial attributes to which we are sensitive to (presence, envelopment, scene, depth and more) [20].

According to [21], the spherical coordinate system used in spatial hearing is represented in Figure 2.1, where ϕ represents the azimuth angle, θ the elevation angle, and r the radius that describe the position of sound sources relative to the egocentric position of the listener (0, 0, 0), who is looking to the positive direction of the X axis. The ear that is closest to a source is called ipsilateral and the opposite ear is called contralateral.

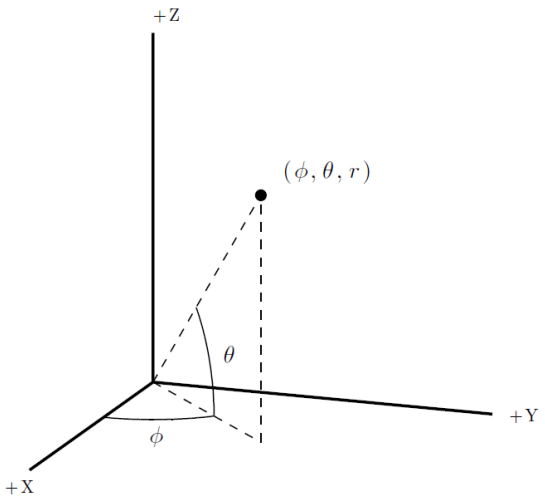


Figure 2.1. Spherical coordinate system for spatial audio

The ability of spatial hearing is the result of the spatial cues generated by the interaction of the sound field with the anatomy of the listener and surroundings on its way to the listener's eardrums. The following is a brief introduction to these cues.

2.1.1 Interaural Differences

For a given sound field, the sound signals reaching both ears of the listener are different. These differences between what each ear hears are essential for spatial listening. One of the basic binaural processing mechanisms involves the comparison between the time of arrival of the sound to the left and right ears. This difference is commonly known as *Interaural Time Difference* (ITD). If we assume that the average distance between human ears is about 18 cm [22], the ITD has a maximum value of around 0.75 ms. Notice that the ITD will not uniquely determine the direction of a sound source since there will always exist ambiguity with respect to the front and back hemispheres.

Another consequence of the presence of the head is that higher frequencies are attenuated or shadowed by the head as they reach the contralateral ear. This attenuation produces an *Interaural Level Difference* (ILD) which

also plays a major role in lateral localization, especially at high frequencies.

The ITD and ILD are considered to be the primary cues for the perceived azimuth angle of a sound source, as proposed by Rayleigh in what is known as the "Duplex theory" [23]. In principle, knowledge of the ITD and ILD would allow to the listener to estimate the azimuth angle, and hence to constrain the location of the source to a particular cone of confusion. Localization in elevation involves other auditory cues as described next.

2.1.2 Spectral cues

In the median plane (i.e. $\phi = 0°$), the bilateral symmetry of the body implies that both the ITD and the ILD vanish. However humans are still able to localize sound in the median plane by what is known as monaural cues, which are related to the spectral changes introduced by the outer ears (i.e. pinnae) at high frequencies and other body structures like the torso at low frequencies [24]. Some studies have shown that these cues help listeners with complete hearing loss in one ear to localize the azimuth direction of a source with relatively high accuracy. However, this was not the case for fully-binaural subjects with a blocked ear [25].

Spectral cues are also used to discriminate the front from the back when the sound source has sufficient high-frequency energy (above 3 kHz). These cues are introduced by the front/back asymmetry of the pinna, which results in a pinna shadow for sound sources arriving from the back. In the absence of this cue, head rotation is necessary to resolve front/back ambiguity [26]. In fact, effective localization of unfamiliar sources in the median plane can only be achieved with head motion.

2.1.3 Distance cues

Perception of the distance of a sound source relies on many factors and is strongly dependent on the familiarity of the listener with the source sound itself. In the range of distances encountered in regular rooms, the inverse-distance law of sound pressure level attenuation is probably the most prominent distance cue. However, level attenuation with distance is a relative cue and possible the only one if the listener has an auditory reference distance for the source.

In absence of a reference, the auditory system seems able to infer dis-

tance from a variety of additional cues. For close ranges below about 1 m, ITD and ILD become distance-dependent, due to near-field effects of the sound source that are otherwise negligible at longer ranges. Significant ILDs occur even at low frequencies at such distances [27]. These supplementary binaural cues seem to be utilized by the auditory system for nearby sources [27, 28].

Otherwise, at regular distances an important factor for distance perception is the direct-to-reverberant ratio (DRR) of the source power over the reverberant sound power reflected and diffused by the room [29]. Reverberation itself is not a primary cue, but it affects jointly a number of distance cues. Recent research shows that externalization and distance perception may originate from short-term statistics of ILD [30, 31] or ITD fluctuations [32] that occur in reverberant conditions. Furthermore, there is evidence of monaural cues that depend on reverberation [33], such as information due to absorption and filtering of the reflected sound with regard to the direct sound.

At long distances, after a few tens of metres, propagation effects become perceptible, mainly due to the kinetic energy of the wave dissipating as thermal energy in air, reducing high frequency content [34].

2.1.4 Dynamic cues

Dynamic cues are extremely useful to resolve the ambiguities that static cues cannot handle. Localization of an unknown sound source in anechoic conditions, especially across cones of confusion, can be a daunting task in the absence of other assisting cues if the listener and the source are immobilised. Free-field localization can improve considerably through relative motions between the listener and the source. Small head rotations especially can jointly modify all major localization cues and resolve front-back confusions or ambiguous elevation effectively [26, 35, 36, 37]. These experiments have shown that listeners evaluate interaural differences at the same time as they move their head in relation to the direction of the source. All cues need to be consistent to produce the correct perception, including other non-auditory cues (e.g. visual cues) that carry information [26].

Another common dynamic cue for distance perception is the Doppler effect of a moving source or listener at high speeds, with its characteristic frequency shift [38].

2.2 Spatial sound systems classification

Different techniques or systems can be used for spatial sound reproduction. These are founded on the different exploitation of some of the perceptual cues described above. The use of the perceptual cues will be limited by the medium used for reproduction (loudspeakers or headphones) and will also be influenced by the application of the spatial sound.

Spatial sound systems can be classified into three main groups: those based on amplitude panning, those based on binaural reconstruction and those based on acoustic field synthesis.

All three basic approaches of spatial sound reproduction have a number of different physical and perceptual properties. None of the systems seems to provide superior properties in all aspects. When it comes to the design of the appropriate reproduction technique for a given purpose, several aspects need to be considered: the timbral fidelity, spatial fidelity, listening area, spatialization parameters and scene representation, number of loudspeakers, computational complexity, etc. From these aspects, some preferences for the choice of a spatial sound reproduction technique can be derived. Beyond a single choice, developments can lead to systems based on a combination of the basic approaches.

In order to address the difficult problem of sound reproduction, it is becoming more important to further understand and compare the different methods in terms of perception; and also to gain a better understanding of human auditory perception and the reproduction of complex auditory scenes. Progress in both areas is expected to support the development of optimized and novel methods for the upcoming trends.

In the following, the three groups of spatial sound systems are briefly introduced before moving on to a more detailed description of some of their particularities.

2.2.1 Based on amplitude panning

Reproduction using two-channels (popularly known as stereo) is the common way most people know to convey some spatial content in sound recording and playback, and this can be considered the simplest approach to spatial sound. In addition, multichannel surround sound systems evolved in the 1970s to provide a better sense of stereophony to a large audience in

movie theaters. They subsequently entered into homes providing a more complete sound experience to a large part of the population. Multichannel sound mixes create a spatial sensation that does not always seek precision in source localization. For example, multichannel mixes intended to be used with video projections often introduce back channels that help to generate an overall diffuse spatial impression with ambient sound or reverberation.

There are generalizations of the panning techniques for an arbitrary number of loudspeakers considering also three-dimensional positions of loudspeakers. The most famous is the Vector Base Amplitude Panning (VBAP) method [6].

All amplitude panning based systems have an optimal listening position, known as the "sweet spot". This optimum listening area is quite limited by concentrating on the central point in the loudspeaker set-up and the spatial impression is considerably degraded outside this central area [39]. Acoustic simulations of complex loudspeaker setups are able to predict the behavior of a spatial sound system with any room geometry [40].

2.2.2 Based on binaural reconstruction

Another strategy consists of reproducing directly in the listener's ears, usually by means of headphones, the signal that would be perceived in the acoustic environment to be simulated. This strategy is known as binaural reproduction. The signals to be reproduced with the headphones can be recorded with an acoustic dummy head or can be artificially synthesized using a measured Head Related Transfer Function (HRTF) [41].

There are some issues to be resolved regarding the variability of HRTF between different subjects and active lines of research focus on this aspect of binaural reproduction.

In addition, incompatibility in the reproduction of binaural signals through loudspeakers is another classical problem: the reproduction of binaural signals through loudspeakers introduces crosstalk, where the left channel signal intended for the left ear will also be heard by the right ear and vice versa. This unwanted effect can be partially eliminated by pre-filtering the binaural signal with an inverse system called crosstalk cancelling filter [42].

2.2.3 Based on acoustic field synthesis

Sound field rendering or synthesis methods use a large number of loudspeakers to reproduce a sound field not only at the ears of one listener, but in a larger space that includes several listeners. The goal is to correctly reproduce the sound field generated by a set of virtual sources. Unlike amplitude panning techniques, psychoacoustic effects play a minor role here, since it is assumed that the listeners respond to the synthesized sound field in the same way as to the original.

Ambisonics is a technique proposed in the early 1970s [9, 43], which allows the codification of three-dimensional sound fields, either by recording or synthesis. These encoded sound fields can be reproduced through different arrays of speakers, which is known as Ambisonic decoding. An advantage of Ambisonics playback is that it is based on solid mathematics. The accuracy of sound field reproduction is given by the Ambisonic order, which is related to the order of a spherical harmonic decomposition of the sound field. The zero order, would use a single channel and corresponds to the reproduction in mono. Ambisonics starts to work as a spatial system from the first order, known as B format, which uses a four-channel coding corresponding to the first spherical harmonics of the sound field decomposition. Loudspeaker signals are derived by using a linear combination of these four channels, where each signal depends on the actual position of the speaker relative to the center of an imaginary sphere, the surface of which passes through all available speakers. From order two it is considered Higher Order Ambisonics (HOA), and with each new order it progressively includes more and more channels corresponding to the series of spherical harmonics that comprise the decomposition of the sound field [44]. With a higher order the regeneration of the sound field is more and more precise. Although Ambisonics is based on sound field synthesis, correct listening also depends on a particular small sweet spot. However, the sweet spot can be extended with large speaker configurations and high orders of ambisonics. Ambisonics can also be used in conjunction with binaural headphone sound, binaurally synthesizing virtual speakers into the desired positions for Ambisonics decoding [45]. In spite of the importance that Ambisonics is gaining in the reproduction of spatial sound, this technique will not be explained in depth in this work because it has not been used directly in the experiments carried out.

Wave Field Synthesis (WFS) is a spatial sound field rendering technique

based on the Kirchhoff-Helmholtz integral theorem, which states that the acoustic field inside a source-free region can be completely determined by the values of the pressure and pressure-gradient on a surface that encloses it [8, 46, 47]. The method employs loudspeaker arrays to synthesize the wave field based on virtual sound sources at some position in the sound stage behind the loudspeakers or even inside the listening area. By reducing the problem to two dimensions and with a few simplifying steps, WFS approximates an acoustic field over a large area with a dense distribution of loudspeakers on a boundary which can be either completely surrounding or covering only partially the area. The method derives sets of digital filters for the loudspeaker array that can recreate plane waves and point sources over the valid region of reconstruction. WFS, like certain variants of high-order ambisonic panning, physically recreates the sound field, thus delivering appropriate cues to all listeners inside that region. Complex scenes can be created by superimposing multiple sources. Additionally, it is possible to generate spherical wavefronts with their virtual origin inside the reproduction region, an effect known as focused source. With WFS every listener experiences the feeling that the origin of the sound is actually in the position of the virtual sources, regardless of their own position in relation to the system. Furthermore, the synthesized wave field is correct for an extended listening area, not depending on a "sweet spot" like amplitude panning systems. As a drawback, WFS has an operating frequency range for which the valid sound field reconstruction is limited by the practical spacing of the loudspeaker distribution. For wavelengths below half that distance, spatial aliasing occurs, rendering reconstruction inaccurate and perceptually incorrect [48]. Limitations of practical WFS systems and some of their perceptual effects are studied and presented in detail in [49].

2.3 Panning with loudspeakers

In the past, space sound reproduction systems have been based only on the phantom effect to try to simulate a sound environment around the listener. In this section a review of the most important systems of this category will be made, making previously a small revision on the phantom effect in which they are based.

2.3.1 The phantom effect

The simultaneous use of two separate loudspeakers for the reproduction of a single sound source allows the creation of a *phantom* source, which is perceived as a substitute for the real sources. This effect is called *summing localization* [3] and is assumed to create binaural signals very similar to those created by the actual sources. There are objections to this explanation. In his *association model* [50], Theile argues that overlapping signals from different speakers do not create localization, but rather that the signals from the two speakers give two different localization stimuli that melt together into a phantom source after a complex psychoacoustic process. Leaving aside the open questions about the nature of the phantom sources, it is a fact that the laws of stereo panning and recording techniques with stereo microphones have been widely used to achieve spatial localization in all kinds of good quality products.

The most commonly used feature to generate the phantom effect is the ILD. The virtual position of the source is achieved by varying the amplitude or level of sound pressure with which a sound is reproduced between two loudspeakers. If the two speakers maintain the same signal level, the ILD perceived by the listener should be zero, which is interpreted as the sound coming from the centre between the speakers. In the case that one of the speakers has a higher level the ILD will be biased to that side, which will be perceived as the source being located towards this direction.

2.3.2 Stereophony and multichannel sound

The word stereo has its etymological origin in the Greek term stereós ($\sigma\tau\epsilon\rho\epsilon\acute{o}s$) which means solid. Applied to the sound, that is, stereophony, the aim is to add to the sound the characteristic of the solid, in other words, sound with volume and therefore spaciousness. Thus, by saying stereophony or stereo sound, we mean "solid sound", or sound that has spatial characteristics of volume and depth. The popular use of the term stereo to refer to two-channel stereophony is due to the fact that the first commercial system using stereophony employed two channels, even though the word stereo refers to a broader concept.

Practical experience and a variety of formal research claim that the optimal configuration for two speaker stereo is an equilateral triangle with the listener located at one of the vertexes, as seen in Figure 2.2. If the

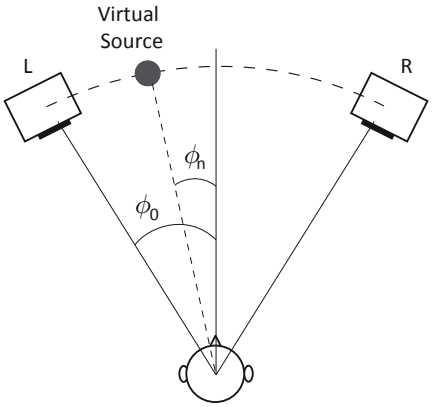


Figure 2.2. Two channel stereo set-up

amplitudes of the two channels are controlled properly, it can produce a resultant phase and amplitude differences of continuous sounds that are very close to those experienced with natural sources, thus giving the impression of virtual or phantom images anywhere between the left and right speakers [51]. This is the basis of the Blumlein's stereo system, invented in 1931. It is often assumed that the mixing coefficients used in stereo synthesis relate to the angle ϕ_n of the virtual source perceived by the panning law of the tangent [3]:

$$\frac{\tan\phi_n}{\tan\phi_0} = \frac{a_{ln} - a_{rn}}{a_{ln} + a_{rn}} \tag{2.1}$$

where ϕ_0 is the angle of separation of the speakers, ϕ_n is the angle of the position of the virtual sound source n, and a_{ln} and a_{rn} are the gain coefficients for the left and right loudspeakers.

Although two-channel stereo sound was a major breakthrough for consumers in the 1950s and 1960s, it has some limitations. The difference between the left and right channels was overemphasized on some recordings, and there were not enough mixing elements in the "phantom" center. Also, although the sound is more realistic, the lack of ambient information, such as background reflections or other elements, leaves the two-channel stereo sound with a "wall effect" where everything comes from the front. Therefore, it lacks the natural sound of reflections from the back wall or other acoustic elements.

In order to improve the spatial impression of the sound played in the

entertainment industry, many film production companies proposed to play the sound tracks using multiple audio channels. This was the birth of multichannel sound systems in the 1970s. The intention was to provide stereophonic sound to the entire audience in a movie theater, partially relieving them of the extremely rigid "sweet spot" position of two-channel stereo. Multichannel sound systems are popularly known as *surround* systems, as it was the term employed to advertise them commercially. As some loudspeakers are placed at the back and surrounding the listeners, these systems can produce ambience effects that "surround" the listener, but they can also generate more precise source positions and spatial impressions than two-channel stereo.

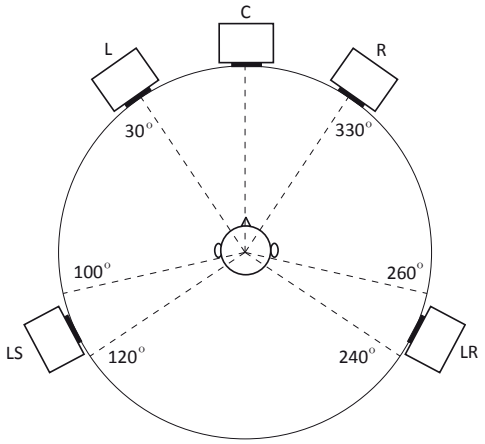


Figure 2.3. 5.1 format according to the ITU-R BS.775 standard

The most widely used multichannel format is 5.1, which allows the provision of stereo effects or room environment to accompany a sound stage mainly oriented to the front. In essence, the three front channels are intended for a conventional three-channel stereo sound image, while the rear/side channels are primarily used to generate a supportive environment, effects or "room impression". In the early 1990s, the 5.1 configuration, introduced into their systems by both Dolby Laboratories (Dolby Digital) and Digital Theater Systems (DTS), became the "de facto" standard for speaker arrangements for multichannel systems, especially in the home. The channel configuration can be seen in Figure 2.3, according to the recommendation ITU-R BS. 775-3 [52]. The ".1" in 5.1 refers to a

channel dedicated to low frequency effects (LFE) or a subwoofer channel and has a limited bandwidth. Later, more channels were gradually added in other systems, such as the 7.1 Sony Dynamic Digital Sound or the 6.1 Dolby Digital Surround EX and DTS-ES. A good review of the evolution of all these early multichannel surround systems can be found in [53] and [54].

The next generation of multichannel systems incorporated height channels to improve the immersive listening experience. Different configuration were proposed: 2 height loudspeaker systems (2+2+2 and THX 10.2); 4, 5, and 6 height loudspeaker systems (AURO-3D 9.1, 10.1, 11.1 or 13.1); and a 9 height loudspeaker system (NHK 22.2) [4, 55]. The distribution of the channels of these systems follows the recomendations ITU-R BS.2051 [56] and ITU-R BS.2159 [57]. Research indicates that at least three or four height loudspeakers are required to deliver an enhanced spatial quality that is better than the conventional 5-channel (horizontal only) configuration [58].

Discrete channel-based methods for height sound reproduction face the challenge of how and where to install height speakers. The position of height speakers has a significant influence on the perceived spatial attributes. To overcome this challenge, a number of methods were proposed that can reconstruct height information, as originally intended by the designer or sound producer, for virtually any speaker configuration. The most versatile and robust method is called object-based audio [59, 60], and provides a scalable and dynamic sound mix for any speaker configuration. The introduction of Dolby ATMOS (a hybrid method using channel and object based audio) brought object based audio to the consumer market. The formats DTS:X, AuroMax and MPEG-H can also employ object-based audio [61, 62].

The multichannel systems with height loudspeakers have benefited from the generalization of the three-dimensional loudspeaker panning law described in the following section.

2.3.3 Vector Base Amplitude Panning VBAP

Vector Based Amplitude Panning (VBAP) is a method for positioning virtual sources to arbitrary directions using a setup of multiple loudspeakers, and was formulated by Pulkki [6, 63]. A great advantage of VBAP repro-

duction is that the number of loudspeakers needed is arbitrary and they can be also positioned in an arbitrary 2-D or 3-D arrangement. VBAP is based on amplitude panning, so the same sound signal is applied to a number of loudspeakers with appropriate amplitudes. For 2-D setups, VBAP is a reformulation of the existing pairwise panning method [64]. However, it can be generalized for 3-D loudspeaker setups as a triplet-wise panning method [65]. A sound signal is then applied to one, two, or three loudspeakers simultaneously.

With loudspeaker systems that also include elevated loudspeakers, the pair-wise paradigm is not appropriate. Triplet-wise panning can be formulated for such loudspeaker configurations. The loudspeakers in a triplet form a triangle from listener's view. The listener will perceive a virtual source inside the triangle, depending on the ratios of the loudspeaker amplitudes.

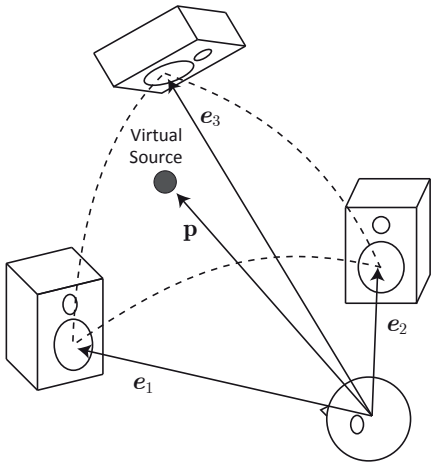


Figure 2.4. Three-dimensional panning VBAP

In three-dimensional VBAP, a loudspeaker triplet is formulated with vectors as in Figure 2.4. The Cartesian unit-length vectors e_1, e_2 and e_3 point from the listening position to the loudspeakers. The direction of the virtual source is presented with a unit-length vector $\mathbf{p}$. Vector $\mathbf{p}$ is expressed as a linear weighted sum of the loudspeaker vectors:

$$\mathbf{p} = g_1 e_1 + g_2 e_2 + g_3 e_3 \tag{2.2}$$

Here, g_1, g_2, and g_3 are the gain factors of the respective loudspeakers. The

gain factors can be solved as

$$\mathbf{g} = \mathbf{p}^T \mathbf{L}_{123}^{-1} \tag{2.3}$$

where $\mathbf{g} = [g_1\ g_2\ g_3]^T$ and $\mathbf{L}_{123} = [e_1\ e_2\ e_3]$ ($[\cdot]^T$ means transpose). The calculated factors are used in amplitude panning as gain factors of the signals applied to respective loudspeakers after suitable normalization, e.g. $||\mathbf{g}|| = 1$. If more than three loudspeakers are available, a set of non-overlapping triangles are formed of the loudspeaker system before run time. There can be several virtual sources applied to one triplet and the triangularization can be performed automatically using the method presented by Pulkki in [66].

2.4 Binaural sound

Binaural is a term that describes the natural hearing process involving the use of two ears and cognitive processing in the auditory cortex. Binaural hearing allows, among other things, the perception of sound with full spatial characteristics. *Binaural sound* is a technical concept that makes use of the binaural hearing by focusing on the specific sound signals that reach the right and left eardrums of a listener, including all the modifications that sound waves undergo due to the environment and the morphology of the listener. With this sound reproduction technique, the listener's perspective is different from that of normal loudspeaker reproduction. Classic loudspeaker reproduction (stereo, multichannel or sound field synthesis) tries to simulate or generate the field that the listener goes into, while binaural reproduction tries to directly feed the sound signals of a supposed sound field into the listener's ears. One of the main differences is that in the case of loudspeaker reproduction the presence of the listener physically alters the sound field generated before it reaches the eardrum, whereas in the case of binaural sound the effect of the listener on the sound field must be included in the sound signals to be reproduced.

The main use of binaural sound is to reproduce spatial sound with headphones, since headphone listening does not include the physical effect of the listener on the sound field and is capable of feeding different sound signals to each ear, without crosstalk between them.

Binaural sound can be recorded directly using a binaural microphone,

also known as "dummy head", a mannequin that resembles the human anatomy (shoulders, head and ears) with microphones inside the ears. These mannequins produce the same physical effect on the sound as a real person would do, and the sound that includes this effect is recorded by the microphones. There is also the possibility of synthesizing the binaural sound by artificially introducing the morphological effects of the listener on the sound by means of signal processing. This process is done by using the HRTF.

2.4.1 The HRTF

In an anechoic environment, as the sound propagates from the source to the listener, the different structures of the listener's own body will introduce changes in the sound before it reaches the eardrums. The effects of the listener's body are captured by the *Head Related Transfer Function* (HRTF), which is the transfer function between the sound pressure present at the center of the listener's head when the listener is absent and the sound pressure experienced in the listener's ear [2, 67]. HRTF is a function of direction, distance and frequency. The inverse Fourier Transformation of the HRTF, i.e. the same information in the time domain, is the Head Related Impulse Response (HRIR), which is a function of direction, distance and time. In the time domain, the ITD is encoded in the HRIR as differences in the arrival time of the sound between the ipsilateral and contralateral sides. Close to the median plane, the arrival time of the wavefront is similar for both ears. However, as the azimuth angle increases, the time of arrival to the contralateral ear progressively exceeds that of the ipsilateral, thus increasing the ITD. The ILD is encoded as the level differences observed in the magnitude responses of the HRTF. The level difference is small near the median plane, and increases with lateral angle. In the median plane, both the ITD and ILD are very small, nearly zero, but there are strong spectral variations (i.e., monaural cues) that change with elevation.

While the HRTFs of most humans share many similarities, a close examination reveals differences determined primarily by disparities in the subjects' body shape and size. These morphologically-dependent differences play an important role in accurate spatial location and perception. Only the use of our own HRTF can result in realistic and accurate binaural audio, as has been demonstrated in several experiments [68]. Some research groups have investigated the effects of synthesizing spatial audio

using distorted versions of the measured HRTF. Kulkarni and Colburn [69] conducted experiments with progressively smoothed versions of HRTFs. The results showed that the HRTFs could be drastically smoothed, and still generate a convincing spatial sound. However, these spatializations can produce localization errors and lack the feeling of transparent realism.

2.4.2 Problems of binaural sound

One of the main problems of binaural sound is that although it works straight with headphones, it is not directly compatible with loudspeakers reproduction. If the two-channel stereo signal was encoded with the HRTF of the listener, then delivering the signal on each channel of the stereo recording to the ipsilateral ear, and to that ear only, would ideally ensure that the listener's ear-brain system receives the cues it needs to perceive accurate three-dimensional reproduction of the recorded sound field. Since, with playback from two speakers, each of the cues is also heard by the contralateral ear (crosstalk), accurate reproduction of binaural audio through loudspeakers requires an effective cancellation of this unwanted crosstalk. Without such crosstalk cancellation, the ITD and ILD cues will inevitably be corrupted [54]. Even though that crosstalk cancellation has its own problems [70], there are systems that exploit this concept to reproduce binaural sound [71, 72], but this dissertation will focus on the reproduction of binaural sound with headphones, as this is the main and direct application of binaural spatial sound.

The most challenging problems of binaural sound are those related to the individualization of the HRTF. As previously commented, the HRTF depends on the morphology of the subject, with spectral and temporal differences between people. Although the basic similarities among generic HRTFs are included in a certain range and therefore can generate a certain generic spatial impression, the individual differences are used for the binaural hearing process to properly identify a spatial scene and locate sound sources. After all, each individual has learned to hear with their own morphology (their own HRTF) and by listening using a different HRTF than their own, the spatial cues received change, generating an incongruent perception for the individual that translates into different types of errors.

The perceptual changes that occur when using a non-individual HRTF can be classified into timbre irregularities and localization errors [73, 74], although both perceptions are obviously related. Because the HRTF acts

as a filter that modifies the spectrum of sound, each directive filter (the HRTF is dependent on the position of the source/listener) introduces coloration, a timbre variation, which the brain compensates for by interpreting it as a spatial position variation. Even if listening using two different HRTFs for the same individual resembles similar spatial positions, the timbral differences of a non-individual HRTF will manifest as an unnatural and unrealistic feature of the sound scene. This non-individual timbral perception can lead to a large and undetermined variety of errors in spatial perception.

Incorrect sound localization due to the use of a non-individualized HRTF can also be grouped into different types of positioning errors: front-back confusion (also called reversals), in-head localization (or internalizations), and errors in elevation perception. Front-back confusion occurs when the listener is unable to distinguish whether the sound source is located in the front or rear hemisphere of the listener. Although front-back confusion may occur in everyday listening, the ratio increases for listening with headphones, especially with non-individualized HRTFs [75, 76]. The phenomenon of in-head localization refers to the special case in which spatial perception places the sound source inside the head, which normally occurs over headphones [77]. Elevation perception (or perception of sound sources in height, either up or down) is usually unprecise in binaural listening with non-individualized HRTF [68], making the elevation perception very dependant of the shape of pinnae [78].

Head-tracking systems can add information about the movements and position of the listener's head. This supplementary information can be used to dynamically correct the binaural sound, so that the sound scene is perceived as static and anchored to the environment, rather than anchored to the listener's head. The information obtained from the tracker about the position of the listener is used to adapt the HRTFs applied to synthesize a binaural environment in real-time. In comparison to a static listening mode, an interactive system involving head tracking has a significant benefit not only in creating a more realistic and immersive listening environment, but can also improve localization accuracy and externalization, and decrease front-back and up-down confusions [79, 80].

Achieving even partial externalization of sound sources can help to resolve many cases of front-back confusion and also to localize better the elevation. However, accurate elevation perception remains highly dependent on the spectral cues defined by the individual HRTF [80, 81].

Apart from the use of individualized HRTF and tracking, the addition of early realistic reflections and diffuse field is also reported to improve location performance and perceived realism [80].

2.4.3 HRTF individualization techniques

Using non-individualized HRTFs in binaural reproduction can lead to an unconvincing experience due to a distorted spatial impression suffering from poor localization, an increase in front-back and up-down confusions, and decreased externalization. Many studies have confirmed that individualized HRTFs lead to a better spatial impression and a more realistic listening environment. The improvements include better localization accuracy [82], including fewer errors along the cone of confusion, and an improved externalized image. Significant efforts are being made to improve the quality and accuracy of the spatial image by providing listeners with personalized, or close to personalized, HRTFs in order to provide a more natural listening environment over headphones. There are a number of methods available to approximating a listeners individualized HRTFs.

These methods include individualized acoustical HRTF measurement, adjustment or modeling through antropometric data, and selection or customization with perceptual feedback.

There exists a trade off between the quality of the spatial auditory image and the complexity of obtaining HRTFs. Figure 2.5 represents this trade off and continuum of the spatial quality of the auditory image as a function of the method of obtaining HRTFs. Typically, HRTFs that are easily obtained and require no listener participation, like a generic HRTF measured on a mannequin, will generally result in a lower virtual source spatial image quality. A generic HRTF dataset will usually lead to a virtual sound source with poorer localization accuracy, has increased front-back confusions, and may lead to artifacts such as spectral distortions or an unnatural sound, or an image that is perceived inside the head, or any combination of these. On the other end of the spatial auditory image quality spectrum are the individually measured HRTFs. These require highly specialized equipment, facilities, and significant time, but result in an accurate spatial impression of the virtual auditory space. Between these two extremes can be found levels of customized or adapted filter sets. Adding a level of customization through either input from the listener or individual measurements will result in an improved quality image.

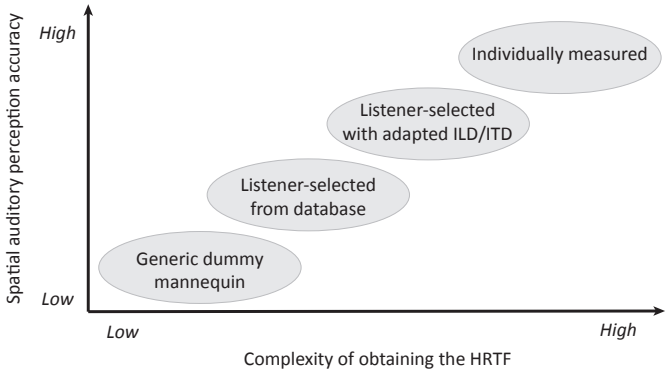


Figure 2.5. Spatial perception accuracy vs. complexity of obtaining the HRTF (Adapted from [54])

Acoustical measurements

The most straightforward individualization technique is to actually measure the individualized HRTFs for every listener at different sound positions [83, 84]. This is the most ideal solution but it is extremely tedious, involves highly precise measurements and complex, prototypical installations. These measurements also require the subjects to remain motionless for long periods, which may cause the subject's fatigue. Zotkin et al. developed a fast HRTF measurement system using the technique of reciprocity, where a microspeaker is placed into the ear and several microphones are placed around the listener [85]. Other researchers developed a continuous 3D azimuth acquisition system to measure the HRTFs using a multichannel adaptive filtering technique [86]. However, all these techniques to acoustically measure the individual HRTFs require a large amount of resources and expensive setups.

Anthropometric data

Individualized HRTFs can also be modeled as weighted sums of basis functions, which can be performed either in the frequency or spatial domain. The basis functions are usually common to all individuals and the individualization information is often conveyed by the weights. The HRTFs are essentially expressed as weighted sums of a set of eigenvectors, which can be derived from PCA or ICA [84, 85]. The individual weights are derived from

the anthropometric parameters that are captured by optical descriptors, which can be derived from direct measurements, pictures, or a 3D mesh of the morphology [85]. The solution to the problem of diffraction of an acoustic wave with the listeners body results in individual HRTFs. This solution may be obtained by analytical or numerical methods, such as the boundary element method (BEM) or the finite element method (FEM) [84, 85]. Other methods used include multiple linear regressions [84], multiway array analysis [87], and artificial neural networks [84]. The inputs to these methods can be a simple geometrical primitive [88] (for example a sphere, cylinder, or an ellipsoid), a 3D mesh obtained from a magnetic resonance imaging (MRI) machine or laser scanner or a set of two-dimensional (2D) images [85]. Another technique used a simple customization technique, where an HRTF is selected by matching certain anthropometric parameters [89]. One of the major challenges today to numerically model the HRTF is the very high resolution of imaging techniques required for accurate prediction of HRTFs at high frequencies. The required resolution of the mesh imaging depends on the shortest wavelength, which is around 17 mm at 20 kHz [85]. Obtaining these accurate 3D models can be done with expensive laser or MRI scanners. However, good results have also been obtained with more affordable structured-light 3D scanners [90].

Perceptual feedback

Several attempts have been carried out to personalize HRTF from a generic HRTF database using perceptual feedback. Subjects select the HRTFs through listening tests, where they choose the HRTFs based on the correct perception of frontal sources and reduced frontback reversals [85]. Listeners can also adapt to the non-individualized HRTF by modifying the HRTFs to suit his or her perception. Middlebrooks observed that the peaks and notches of HRTFs are frequency shifted for different individuals and that the extent of the shift is related to the size of pinna [91]. Listeners often tune the spectrum until they achieve a satisfactory and natural spatialization [85]. Other techniques involve active sensory tuning [84] and tuning the PCA weights [92] to individualize the HRTFs. These perceptual-based methods are much simpler in terms of the required resources and effort compared to the individualization methods using acoustical measurements or anthropometric data. However, these listening sessions can sometimes be quite long and result in listener fatigue.

2.4.4 Headphones equalization

The specific spectral cues on which binaural listening depends can be distorted when the effect of the headphones is added to the sound to be played. The response of the headphones themselves introduces coloration into the signal that can result in a poor spatial image. Ideally, the signals in the eardrum reproduced by the headphones should be the same as those generated by sources in the natural environment. The reality is that the acoustic transmitter of the headphones introduces its own frequency response. In addition, due to the physical configuration, resonances are created between the headphones and the cavities of the listener's pinna. This can also add significant and unnatural spectral coloration. If this coloration is reduced to a minimum, studies have shown that externalization can be improved even with non-individualized HRTF [93, 94].

The frequency response characteristics of the headphones and the individual morphology of the listener generate a combined effect that can be characterized by the *Headphone Transfer Function* (HpTF). HpTF describes both the response of the headphones and the coupling to the ear of the listener. It is important to note that the coupling of the headphone with the individual's ear generates a response modification that is individualized, since it is dependent on the morphology of the individual. HpTFs can be measured using a dummy head or by placing microphones in an individual's ears, and individual equalization can be performed to improve binaural listening. The coupling to the individual's ear is especially variable between 4 and 10 kHz, as this is the band most dependent on the morphology of the pinna [95].

To correct the effect of the headphones, the HpTF can be inverted and result in a flat playback method that eliminates the coloration of the headphones and their coupling with the individual's ear. With this inverted response, a filter is generated that compensates for the HpTF, making the reproduction chain as transparent as possible, including the individual listener's effect. These inverted filters are usually calculated by means of manually adjustable adjustment methods [96]. Gomez Bolaños [97] has proposed a frequency-dependent regularization method that takes into account the perceptual peculiarity of the ear to large spectral peaks, automatically generating more reliable perceptual filters.

The HpTF is also influenced by the position of the headphones in the

listener's ears, as variations in the coupling between the headphones and the ear occur each time the headphones are repositioned. The spectral repositioning effect is low to moderate in the low frequencies, but can result in significant color differences in the high frequencies [96]. The repositioning effect is more noticeable in supra-aural than in circum-aural headphones. To minimize the effect of headphone repositioning, inverse compensation filters are usually made from the average of several HpTF measurements [98].

The impact of the headphone response and its individual coupling on the individual can be so important in binaural listening (and even in general when listening with headphones) that headphones capable of self-calibration have been proposed [99].

2.4.5 Conversion to binaural

Any channel based sound mix (two-channel stereo or multichannel) can be converted to binaural just convolving the sound signal of each channel with the corresponding HRTF of the spatial position of the respective loudspeaker. The HRTF then works as a virtualizing filter for the loudspeakers, to which binaural reverberation or other effects such as tracking can also be added to enhance the experience.

SAOC and Ambisonics

Object-based audio and Ambisonics spatial coding systems also have applications in binaural reproduction with headphones.

Spatial Audio Object Coding (SAOC) is one of the recent standardization activities in the MPEG audio group [61, 100, 101], and is also being successfully used commercially in Dolby ATMOS [102] and 360 Reality Audio by Sony [103]. SAOC is a multi-object parametric encoding technique that allows great flexibility in decoding during reproduction. The principle of SAOC is based on encoding audio objects (independent pieces or sound sources) as separate elements together with additional information (metadata) about their spatial position and behaviour with respect to the rest of the sound mix. It is designed to transmit a number N of audio objects in an audio signal comprising K downmix channels, where $K < N$ and K is typically one or two channels. Along with this backwards compatible downmix signal, the object metadata is transmitted simultaneously to the decoder

via a dedicated SAOC bitstream. Although this object metadata grows linearly with the amount of objects, the amount of bits needed to encode this data in a typical scenario is negligible compared to the bit rate needed for the downmix channels encoded. In this way, spatial sound reproduction is not dependent on any particular speaker configuration, instead the decoding is scalable depending on the configuration of the playback system. This makes it also fully compatible with binaural headphone sound. In fact, the freedom offered by an object-based encoding system makes it possible for a binaural headphone playback system to dynamically reproduce mixes that respond to the movements of the listener through tracking, providing individual immersive experiences.

Ambisonic decoding for headphones can be done in a similar way to loudspeaker decoding, with the difference that the signals from the loudspeakers are transmitted to the headphones by convolution with the HRIRs of the corresponding playback directions. This headphone decoding approach classically uses a small set of so-called virtual loudspeakers [104, 105]. However, as pointed out in some research papers [104, 106], these approaches have in common that the low-order Ambisonic synthesis is problematic. By inserting a dense grid of virtual speaker HRIRs, Ambisonic smoothing attenuates the high frequency in the front and back directions. The solution for a long time has been to use a coarse grid of virtual speaker HRIRs that does not attenuate the high frequencies, but still produces a spatial quality that depends largely on the particular arrangement or orientation of the grid [106]. Different solutions have been proposed to overcome the problems of Ambisonic to binaural conversion at high frequencies or due to low order material, and the derived complications [45, 107]. In any case, playing Ambisonics via binaural with headphones has some advantages. The sound field reconstruction performed by Ambisonics is dependent on the sweet spot (the spatial point where the effective sound field reconstruction takes place and the only place where listening is correct). However, when Ambisonics is played with headphones the listener's position is always correct, since the sound is synthesized binaurally for the sweet spot, even when tracking is added to the system. In fact, adding rotation of the sound scene is a simple and efficient task in Ambisonics [108], which is perfect for binaural tracking systems that have an egocentric reference system, such as head-mounted displays for virtual or augmented reality.

The compatibility of spatial sound formats (especially through SAOC and Ambisonics) with binaural sound has multiplied the interest in listening through headphones. For the next generations of space sound systems, listening with headphones will be of particular importance.

2.5 Wave-Field Synthesis principles

WFS was introduced by Berkhout [109] as a concept for sound reproduction without the sweet-spot restrictions inherent in common multichannel systems, as illustrated in Figure 2.6. In two- (or multi-) channel stereo playback the spatial properties of the reproduced field are determined by the characteristics of the loudspeakers. The source localization is correct only in a small area between the loudspeakers, shown by dashed lines in Figure 2.6 (a). In WFS the wave patterns of the sources to be reproduced are correctly synthesized in time and space by an array of closely spaced loudspeakers such that their localization is correct for all listeners in the audience area, as depicted in Figure 2.6 (b).

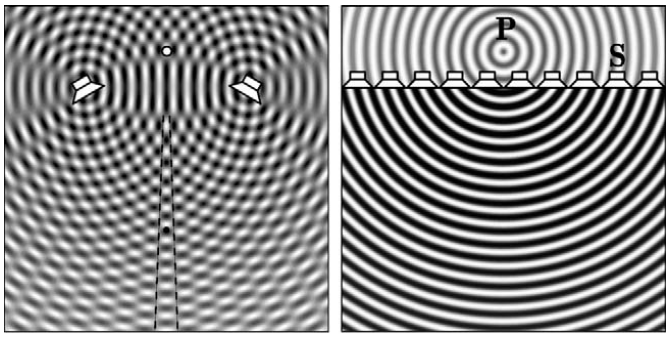


Figure 2.6. Illustration of basic difference between two-channel stereo and WFS. (a) Two-channel playback. Proper sound localization is shown between two dashed lines where sweet spot is located. (b) Wave field synthesis. P - primary source; S - secondary source

Based on the Huygens's principle, WFS reproduces an acoustic field inside a volume from the stored signals recorded in a given surface. Huygens's principle states that the wave front radiated by a source behaves like a dis-

tribution of sources that are in the wave front, named secondary sources, together creating the next wave front. In WFS the synthetic wave front is created by loudspeaker arrays that substitute the individual secondary sources. The ideal situation would be an area completely surrounded by loudspeakers and fed with signals that create a volumetric velocity proportional to the particle velocity normal component of the original wave front. The application of planar loudspeaker arrays, as prescribed by the Huygens's principle, would involve a very high number of loudspeakers and reproduction channels. However, the recreation of a true natural wave field can only be fulfilled with certain restrictions. Huygenss principle needs to be discretized in practice, which means that an infinite continuous secondary source distribution is replaced by a number of finite arrays of equidistant discrete loudspeakers. Therefore, practical WFS systems employ linear loudspeaker arrays that synthesize the field of 3D sources in the ear plane of the listeners, as depicted in Figure 2.7. The lack of continuity leads to a maximum usable frequency, known as spatial aliasing frequency, whereas the finiteness of the array causes some truncation effects. For example, a typical loudspeaker distance in the technical literature is 18−20 cm, which gives an aliasing frequency of about 1 kHz. A detailed description of these drawbacks within a listening room can be found in [110, 111].

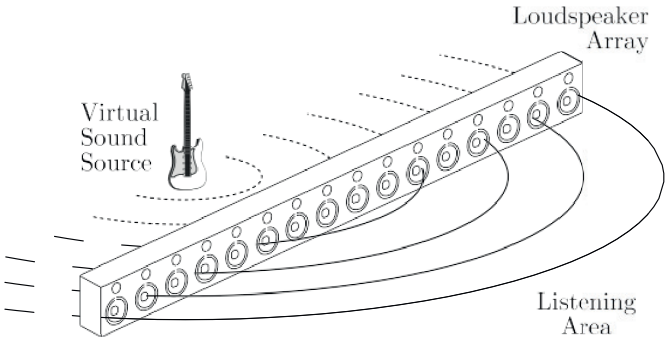


Figure 2.7. In WFS a linear loudspeaker array synthesizes a virtual sound source in the horizontal plane inside the listening area

The main advantage of these systems is that the acoustic scene has no sweet spot since it recreates the wavefront of the virtual sources. When listeners move inside the listening area, the spatial sound sensation changes

also in a realistic way according to its relative position to the virtual source. In addition to virtual sources behind the loudspeaker array, it is also possible to synthesize sources inside the listening area, the latter known as focused sources. A comprehensive derivation of the underlying mathematics can be found in several studies, such as [46, 109].

Part I

Headphones and binaural systems

Effects of Headphones in perception 3

According to Global Market Insights Inc. [13] and Futuresource Consulting [112], audiovisual content is increasingly consumed through headphones, resulting in a remarkable expansion of headset-based audio playback in recent years. Some of the reasons for this popularity are the private hearing they provide in any type of environment as well as the widespread use of smart-phones and mobile devices. Headphones are commonly employed to reproduce stereo-channel-based material, but new formats that employs object coding or ambisonics have also focused on the reproduction with headphones [61, 101]. Besides, immersive technologies such as augmented and virtual reality are also employing headphones to reproduce 3D audio with binaural techniques [113], adding a greater sense of realism and presence to multimedia content.

There are many different types of headphones, which have their advantages and disadvantages in differing listening environments. In addition, different applications have their own requirements for the headphones, such as size, isolation from or aperture to ambient noise, sound quality and even aesthetic properties. The ITU Telecommunication Standardization Sector (ITU-T) classifies headphones into four groups: circum-aural, supra-aural, intra-concha, and in-ear/insert headphones [114]. Figure 3.1 shows the classification of ITU-T headphones, in which the different types can be

described as: a) circum-aural headphones, which are placed completely around the ear against the head; b) supra-aural headphones, which are placed at the top of the auricle (also called on-the-ear headphones); c) intra-concha headphones (button type), which are placed loosely in the concha cavity; and d) in-ear headphones, which are inserted tightly into the ear canal (similar to earplugs).

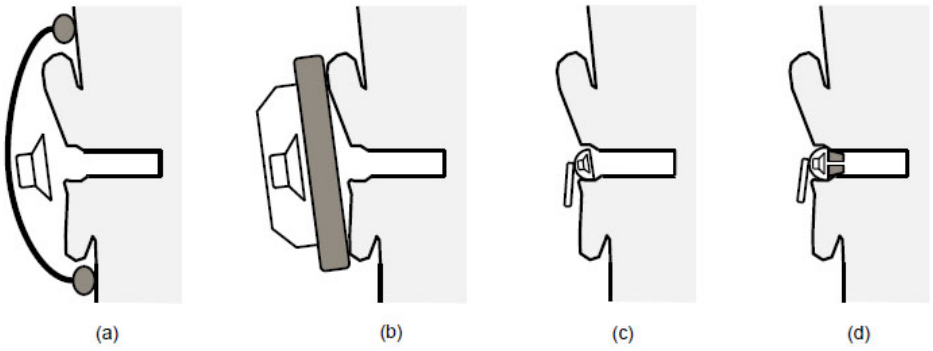


Figure 3.1. Categorization of headphones according to ITU-T Recommendation P.57, (a) circum-aural, (b) supra-aural, (c) intra-concha, and (d) in-ear/insert headphone. Adapted from [114]

Møller and his team at the Aalborg University did a variety of work studying the influence of headphone to ear coupling, taking into account different types of headphones. They also developed procedures to make precise measurements addressing specific problems of headphone listening (different and individual coupling, imprecision of position, leakage...) [67, 115, 116]. Their studies were aimed at specifying the design of headphones also taking into account the use of binaural spatial sound.

In recent years, the design criteria for commercial headphones have undergone significant development. At Harman International Industries, Olive et al. investigated the best target responses for designing headphones based on the listeners preference for the most natural sound, with special focus on in-ear headphones and the consumer market [117, 118, 119, 120, 121].

Other previous works have studied headphones listening attending to perceived quality [122, 123] and listeners perception [124, 125, 126].

In this investigation we focus on the influence of the headphones in the perception of spatial sound image, with a view to the final products that consumers receive and how they influence their listening experience. To have an accurate sense of spatial immersion, a high quality recording chain should be employed, including microphones or acoustic mannequins. In addition, high quality headphones should be used for playback. However, low-end headphones are widely used in most cases, either for economic reasons or simply because they are included with mobile devices. It is generally known that low cost headphones usually provide a poorer sense of immersion, but the degrading factors that cause such a loss in quality and perceived spatiality has not been sufficiently studied, as well as the level of their effects. With the accessibility of spatial sound in mind, consumer headphones of different price ranges and supposed qualities have been studied in this work.

This Chapter includes some experiments structured as a series of perceptual tests, to explore the question of the perception of sound with headphones, from the point of view of the user and his experience. In the Section 3.1 the perception of sound quality and the sensation of immersion produced with consumer headphones of medium and low quality range are studied. Since this experiment showed that frequency response is especially relevant to the perception of spatial sensation, equalization is suggested as a method to modify the response of the headphone. Therefore, Section 3.2 addresses the perception of non-linear distortion produced in particular by a process of equalization of the headphone.

3.1 Perception of quality and immersion

3.1.1 Introduction and motivation

High quality headphones can generate a realistic sound immersion reproducing binaural recordings. However, many people commonly use consumer headphones of inferior quality, as the ones provided with smartphones or low-priced ones.

Listening with headphones is becoming increasingly popular but there is little specific knowledge of how the characteristics of consumer headphones affect the listening perception. As the market is at a time when

more experiences are to be introduced through headphones, it is interesting to know how they will be experimented through the entire existing pool of headphones of different qualities. Moreover, the sensation of spatiality and sound immersion varies from one headphone model to another. Given the interest in producing spatial sound through headphones, the study of these perceptual sensations becomes more relevant.

The spatial sound image or spatial immersion (that a headphone can produce) is considered here as a generic capacity of each headphone model to provide to the user with sensations of sound spatiality. All sound material (stereo and even mono) produces a certain spatial sensation in the listener. The binaural sound would be a specific case of spatial immersion which pursues high precision and realism. Here we focus on studying the generic capacity of the headphones to generate the spatial immersion, since we consider that therefore it will also influence the reproduction of binaural sound.

Different factors can affect the perception of the spatial sound image. Here is going to be considered as hypothesis that there main factors could be responsible for the spatial sensation degradation: the frequency response, the distortion and the disparity between the left-right transducers, especially in low-cost headphones. To determine if this is true and how they may affect, a series of perceptual tests [18] are proposed to particularly study these factors.

A virtual headphone listening test methodology was implemented to carry out the subjective tests in a rigorous way isolating each studied factor. So firstly this technique, the measurements performed and the headphones used in the study are described. Then four perception tests and their results are presented. The first test studies the influence of the sensitivity disparity between left and right transducers and establishes the degree to which perception of the sound source position in the azimuth is affected. Although in high quality headphones manufacturers match transducers with similar sensibilities, low-cost headphones have different sensibilities due to broader manufacturing tolerances. The second experiment focuses on the effect of the frequency response in the perception of quality and spatial impression with headphones. As frequency response is the one of the factors that more prominent varies among different headphones, this test is of particular interest to better understand how frequency response affects the spatial sound impression. The third test analyzes the effects of harmonic

distortion in listening with headphones. Distortion can be considerable if high-dynamic sound and high reproduction levels are employed. It has been employed a Volterra kernels scheme for the simulation of the distortion using convolutions. Finally, the fourth test studies the relation of the frequency response with the accuracy of localization in the horizontal plane. The capacity of a headphone to generate a good spatial immersion can be different from its capacity to generate precise source locations. To explore this point, azimuth localization with binaural sound is tested for different types of headphones.

3.1.2 Measurements and virtual headphone simulation

It is well known in loudspeaker testing that visual cues play an undesirable role in the results provided by test subjects. Similarly, when testing headphones, tactile cues can also influence results. Consequently, it can be challenging to conduct a double-blind comparative listening test for headphones. It is difficult to hide the possible influencing variables, such as brand, design or price. In addition, the manual substitution of different headphones on the subject's head can be disruptive and introduces useless fatigue on the subject [127]. Moreover, the fitting and tactile sensations are impossible to remove, making them an important bias factor [122].

In order to avoid these effects, it is appropriate to use a virtual headphone simulation to perform the listening tests [124, 128]. This method employs one reference headphone to simulate the different headphones under test. In this way, listeners can evaluate the simulated versions of the different headphones wearing just the reference headphone, therefore avoiding the manual change of headphones and removing the visual and tactile biases. Some other advantages are obtained with this virtual method: listeners can have immediate access to the different headphones, and the procedure test becomes more flexible, transparent, controlled and repeatable.

The reliability of this virtual simulation method has been previously studied, finding good correlation between standard listening tests using real headphones and the virtual simulation method. However, in some cases, some discrepancy related to a specific model or sound signal [124] has been found due to the visual and tactile bias present in the standard test [129].

Due to the great advantages of a virtual test over a standard one,

this study used a virtual headphone listening test methodology. This will remove the strong bias that would appear in this study due to the great difference in appearance and fitting characteristics among the consumer headphones and high quality ones used in this test.

Headphones selection

Different headphones were selected in order to represent a range of commercial and readily-available headphones. According to this principle and the scope of the study described in previous sections, seven different headphones were selected plus a high quality reference one. A Sennheiser HD800 was chosen as the reference headphone (REF). The reason for this selection is due to its great fidelity, response, low distortion and accurate timbral reproduction. The other seven headphones were selected to cover a wide range of possible common uses. The brands and models of the rest of the headphones are provided for reference purposes, as no advertising is intended and neither brands are fully represented in this study, and of course they are not necessary for the result analysis.

The headphones used in the study are classified in Table 3.1.

	acronym	**definition**	**model**	**characteristics**
(a)	REF	Reference	Sennheiser HD800	open and circumaural
(b)	HQop	High Quality open headphone	Sennheiser HD545	open and circumaural
(c)	MQcl	Medium Quality closed headphone	Sennheiser HD429	closed and circumaural
(d)	BDso	Big Diaphragm semi-open headphone	Superlux HD668B	semi-open and circumaural
(e)	LCmul	Low-Cost multimedia headphone	Genius HS-04SU	supra-aural
(f)	AirL	Airline headphone	Airfrance	supra-aural
(g)	Woh	Wireless open headphone	Sony MDR-RF800R	open and circumaural
(h)	LCmul2	Low-Cost multimedia headphone 2	Woxter i-Hph 780	closed and supra-aural

Table 3.1. Headphones used in the study about quality and spatial impression

Some of the headphones employed can be seen in Figure 3.2.

Figure 3.2. Some of the headphones employed in the perceptual experiments

The reference headphone was the only one that participants used, saw and had contact with during the tests. The rest of the headphones were simulated through the reference one. Then, all of the participants performed the test using the same high quality reference headphone (REF, Sennheiser HD800). The resulting signals for the rest of the headphones (used in Tests 2, 3 and 4 and described in their sections) were simulated by means of signal processing algorithms and heard through the reference headphone.

Headphones frequency responses measures

To measure the response of the different headphones, a exponential sweep method was employed [130] using a Head and Torso Simulator (HATS) or "dummy head" model Brüel & Kjær (B&K) Type 4100. (Figure 3.3). This technique gave us both the frequency response, as well as the first and second distortion harmonics needed for the simulation of the different headphones.

To avoid differences in the amplitude level of the measures, the selected criterion was to achieve the same equivalent power between 100 Hz to 10 kHz for all of the headphones (for calibration, band-pass pink noise from 100 Hz to 10 kHz was employed instead of 20 Hz to 20 kHz in order to minimize the influence of roll-off in low and high frequencies in low quality headphones). This decision allowed us to measure all of the headphones in the same reproduction conditions and to achieve the same reproduction level in this band of frequencies. The reproduced pressure level for all of the headphones was equivalent to 69 Sound Pressure Level dB (dBSPL) of

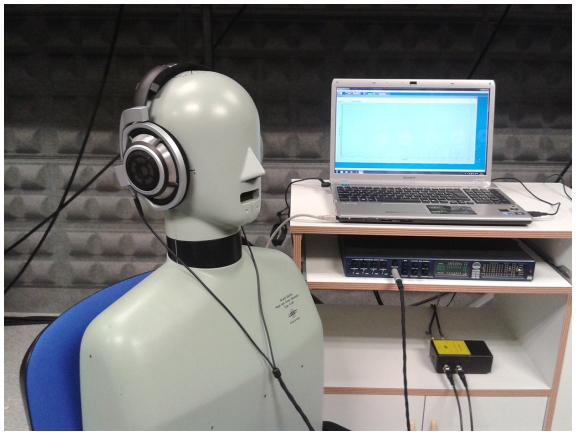


Figure 3.3. Set-up for measuring the headphones with the Head and Torso Simulator (HATS)

pink noise in the reference headphones. This level was selected in informal tests as a pleasant listening level. Besides, this level allowed the measurement of the different headphone models without any saturation distortion in equivalent conditions.

Each of the headphones, including the reference one, were measured with the mentioned exponential sweep method. The resulting impulse responses ($h_i[n]$) were truncated to 50 ms (2205 samples for a $44100Hz$ sampling frequency) and windowed with a half Hamming window. This length provides good resolution in low frequencies until 20 Hz. To minimize errors related to headphone positioning on the ear of the HATS simulator, five resets of the headphones were done and measured. The curves shown in Figure 3.4 are based on the average of those measures.

The first curve corresponds to the reference headphone (a)-REF, which shows a smooth response and flat below 3 kHz. The next three, (b)-HQop, (c)-MQcl, (d)-BDso headphones, were chosen as good mid-quality range with different characteristics: open, closed and semi-open. Their frequency responses below 6 kHz are quite flat, with the exception of some irregularities in the (c)-MQcl curve and a peak down at 4.5 kHz that decreases to -14 dB. There is another peak up in the curve (d)-BDso at 6 kHz of 15 dB. The next curves (e to h) represent the frequency responses of the multimedia (e)-LCmul, airline (f)-AirL, wireless (g)-Woh and another multimedia (h)-LCmul2 headphones, that were chosen to be an example of mid- and

poor quality headphones. Their frequency responses have important peaks and valleys that affect the sound. Curve (e)-LCmul has a reinforcement in frequencies around 1.5 kHz and a big dip in 3.5 kHz, and curve (f)-AirL has a strong peak in 140 Hz, as well as other distortions up to 4.5 kHz. Curve (g)-Woh is flatter in the mid frequencies with a small reinforcement in 1.5 kHz and a decay around 4.5 kHz. In the case of curve (h)-LCmul2, it is important to note the rapid decline above 3 kHz and the lack of proper high frequency beyond 5 kHz. All of these headphones are intended to be a small representation of quality range in commercial headphones.

Headphones frequency response simulation

The seven headphones under study were simulated to be reproduced with the reference headphones ((a)-REF-Sennheiser HD800). The simulation of each headphone was done by filtering with its frequency response, but compensating the effect of the reference headphone using its inverted frequency response. Equation 3.1 shows the process for the simulation, where $H_i(\omega)$ is the measured response of the headphone to simulate, $H_{HD800}(\omega)$ is the measured response of the reference headphone and $H_{i\,corrected}(\omega)$ is the response of the simulated headphone, which is applied to the corresponding stimulus.

$$H_{i\,corrected}(\omega) = \frac{H_i(\omega)}{H_{HD800}(\omega)} \tag{3.1}$$

These virtual headphone equalizations include not only the magnitude response, but also the phase of the headphone measured. Although it is generally accepted that phase does not seem to affect the perceived accuracy of the simulations [96], especially if the stimuli material is a typical music program, it can be noticed with pink noise stimuli. All of the impulse responses of the headphones measured, the correction of the reference headphone and its application convolving with the stimulus, respect and keep the original phases. Moreover, accurate phase processing guaranties that our filtering will not alter in any way the Interaural Time Difference (ITD) between left and right transducers.

The filter implementation of Equation 3.1 was carried out in MATLAB in the time domain, using Equation 3.2; where $h_{i\,corrected}[n]$ is the response for the simulation of the virtual headphone, $h_i[n]$ is the impulse response of the headphone to simulate and $h^I_{HD800}[n]$ is the inverted impulse response

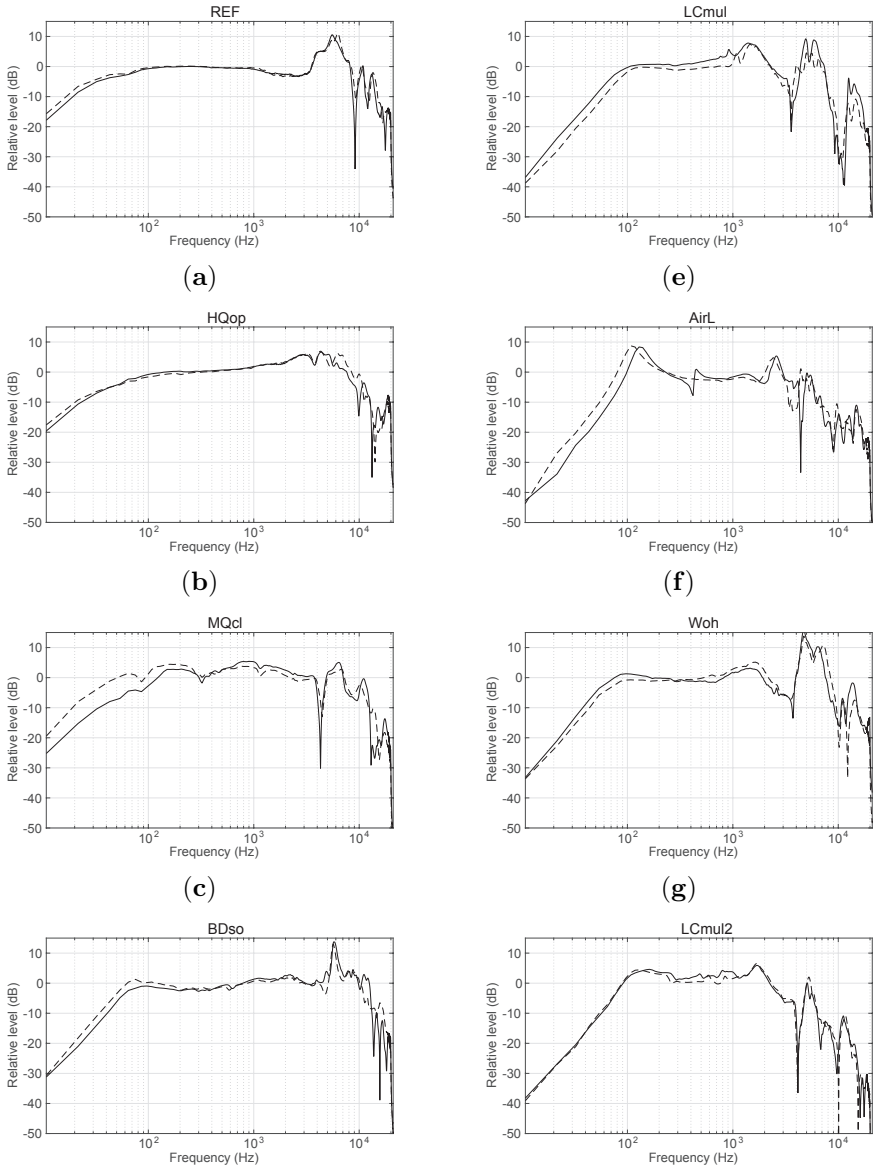


Figure 3.4. Frequency response of the headphones used in the study. Solid curve, left channel; dashed curve, right channel. (**a**) REF, Reference headphone; (**b**–**h**) headphones under study. (**b**) HQop, High Quality open headphone; (**c**) MQcl, Medium Quality closed headphone; (**d**) BDso, Big Diaphragm semi-open headphone; (**e**) LCmul, Low Cost multimedia headphone; (**f**) AirL, Airline headphone; (**g**) Woh, Wireless open headphone; (**h**) LCmul2, Low Cost multimedia headphone 2

of the reference headphone.

$$h_{i\,corrected}[n] = h_i[n] * h^I_{HD800}[n] \tag{3.2}$$

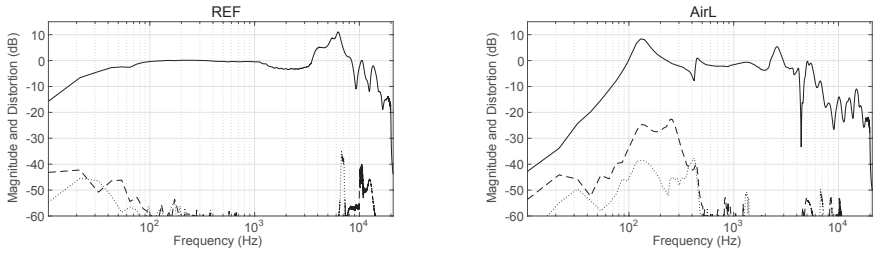
To obtain $h^I_{HD800}[n]$, firstly the impulse response of the reference headphone h_{HD800} was recorded with 2205 sample points (50 ms at fs = 44100 Hz). Secondly, the Fast Fourier Transform (FFT) of the response was computed, with zero padding up to a size of 4096, which guaranties a spectral resolution of 10 Hz. This is low enough to see details of the frequency response. Thirdly, the resulting FFT was inverted, taking into account a boost limitation of +15 dB. This limitation was included to avoid an excess of boost at a couple of very narrow notches of the h_{HD800} response (see Figure 3.4 (a)), assuring that final signals are inside the reproducible dynamic margin and free from artifacts. Lastly, the inverted and limited response was then used to properly compute the inverse FFT and next Hamming windowed to obtain the $h^I_{HD800}[n]$. This process guaranties the avoidance of undesirable effects, such as circular convolution or others.

Finally, the different headphones were simulated applying the simulation filter $h_{i\,corrected}[n]$ to the sound materials for each test, obtaining the different stimuli. This was the procedure used for Tests 2 (Section 3.1.4) and 4 (Section 3.1.6).

Non-linear distortion simulation

As commented on before, the exponential sweep method employed to measure the frequency response of the headphones provides the frequency response and the non-linear distortion harmonics simultaneously. Figure 3.5 shows the frequency response and the second and third order non-linear distortions of the reference ((a)-REF) and the airline ((b)-AirL) headphones. Both of these headphones are a good example of low (a) and high non-linear distortion (b).

To simulate the non-linear distortion of each headphone, a Diagonal Volterra kernels model was used along with a series of linear convolutions as described in [131, 132]. With this method, the transfer function of a non-linear system is estimated by means of a truncated Volterra series. The output signal of a non-linear system can be represented as an infinite sum of convolutions of the Volterra kernels with power series of the input signal. These Diagonal Volterra kernels are computed as a linear combination of each of the infinite orders of distortion impulse responses.



(**a**) (**b**)

Figure 3.5. Frequency response with non-linear distortion of two headphones (left channel). Solid curve, magnitude; dashed curve, second order distortion harmonic; dotted curve, third order distortion harmonic. (**a**) REF, Reference headphone; (**b**) AirL, Airline headphone

The non-linear distortion produced by headphones is in general low and decreases rapidly with the order, making the 4^{th} and subsequent distortion orders almost negligible compared to the 2^{nd} and 3^{rd}. Simplifying the equations in [131] with the previous consideration, Equation 3.3 is obtained,

$$\begin{cases} H_1(\omega) = H_1^{'}(\omega) + 3H_3^{'}(\omega) \\ H_2(\omega) = -2\hat{H}_2^{'}(\omega) \\ H_3(\omega) = -4H_3^{'}(\omega) \end{cases} \tag{3.3}$$

where $H_1^{'}$, $H_2^{'}$, $H_3^{'}$ are the first three harmonic of the impulse response ($H_1^{'}$ the linear part and $H_2^{'}$, $H_3^{'}$ the two first distortion orders), and H_1, H_2, H_3 are the Diagonal Volterra kernels ($\hat{}$ represents Hilbert transform).

Therefore, the second and third order non-linear distortions can be simulated by convolution, as shown in Equation 3.4, where $x(n)$ is the input signal and M is the number of samples of the kernel:

$$y(n) = \sum_{i=0}^{M-1} h_1(i) \cdot x(n-i) + \sum_{i=0}^{M-1} h_2(i) \cdot x^2(n-i) + \sum_{i=0}^{M-1} h_3(i) \cdot x^3(n-i) \tag{3.4}$$

To simulate the linear part of the system response, only the first harmonic $H_1^{'}$ is employed (i.e. $H_2^{'} = H_3^{'} = 0$). Applying this technique to a sound stimulus, it is possible to simulate the effect of a frequency response with and without the measured non-linear distortion.

More details of this technique can be found in [131, 132]. This procedure was followed for Test 3 (Section 3.1.5), and also in the subsequent experiment described in Section 3.2.

Binaural Room Impulse Responses measurements

In order to produce sound sources with spatial characteristics, some Binaural Room Impulse Responses (BRIR) [133] were measured with a HATS B&K Type 4100.

Reverberation is an influential factor for spatial localization [1, 39], and because of this, BRIRs with natural reverberation were recorded instead of using dry responses from a library. The impulse responses were recorded in a rectangular room with a volume of 132 m^3 and a reverberation time of about 0.7 s. Nine different azimuth angles were recorded (0°, 30°, 60°, 90°, 135°, 225°, 270°, 300°, 330°) in the horizontal plane at 1.5 m of distance.

These measures were used to simulate binaural sound source positions in Test 4 (Section 3.1.6).

3.1.3 Test 1. Sensitivity disparity between left-right transducers

Test Description

The idea of this test is to evaluate how sensitivity disparity between the left and right transducers affects the perception of the source azimuth. To do that, a subjective perceptual test was carried out applying some volume level variations to different binaural sounds and checking how this affects the accuracy of horizontal localization.

In this test, participants had to listen, wearing headphones, to some binaural recordings obtained with a HATS on specific angles in the horizontal plane. Different variations of the original level between left and right transducers were applied to these sounds and then presented to the listeners. Participants should then indicate the direction of arrival, marking the angle in a Graphical User Interface (GUI).

The volume level variations applied were 0 (no modification), 1, 2 or 4 dB more on the left channel than the right one. Four different angles of direction of arrival were chosen, −30°, 0°, 65° and 90° of azimuth in the horizontal plane. Besides, the influence of different types of sounds was also studied.

These sounds were specifically recorded for this test using a binaural mannequin (B&K Model 4100) at the specific angles under study. A 44100-Hz sampling frequency was employed, obtaining full audio band recordings. The mannequin was in a semianecoic room, and sources were placed around it at 1 m apart. Four different sounds were recorded: a timbal drum hit, voice, a whistle and pink noise. The impulsivity of the timbal hit is an interesting characteristic regarding sound localization, also interesting for its low frequency content. Both voice and whistle are easily recognizable common sounds, which make them useful for the test. Moreover, the reduced spectral content of the whistle can be an interesting feature that can affect the test. The voice signal was the syllables "ba-be-bi-bo-bu", pronounced by a male voice. This sound has diverse vocalic contents and bilabial consonantal phoneme /b/, which produces impulsive sound. Pink noise was employed to evaluate a wide spectrum signal. All of these sounds were reproduced by the Sennheiser HD800 reference headphones.

According to the different types of sounds described above, the total number of stimuli presented to each participant in this test was: 4 angles $\times$ 4 types of sounds $\times$ 4 level variations = 64 stimuli. These stimuli were randomly presented, and the participant could listen to each of them as many times as he or she wanted.

During the test, participants also had the possibility of hearing a reference stimulus at any time, choosing between $-90°$, $-45°$, $0°$, $45°$ and $90°$ of azimuth.

To perform the test, a simple Graphical User Interface (GUI) was developed in MATLAB that brings the user full control of the test. The participant could select the perceived sound source direction angle in an arc of $-90°$ to $90°$ of azimuth (with a $5°$ resolution). It was also possible for the subject to freely control and listen to the reference stimulus (Figure 3.6).

The test was performed by 20 people, 10 men and 10 women (21 to 45 years, with an average age of 32). The average runtime of the test was 9 min. Every participant did a training session before taking the test, so all could listen to all of the stimuli and become familiar with the GUI and the assigned task.

Figure 3.6. Participant performing the test 1

Results

Figure 3.7 (a) shows the average of the perceived angles (for all of the level variation cases) (95% Confidence Intervals, CI) according to the reproduced target angle. The average of the perceived angles (answers) has a deviation to the left-hand side. This is to be expected since the variations (0, 1, 2, 4 dB) always gave more level to the left channel than to the right one.

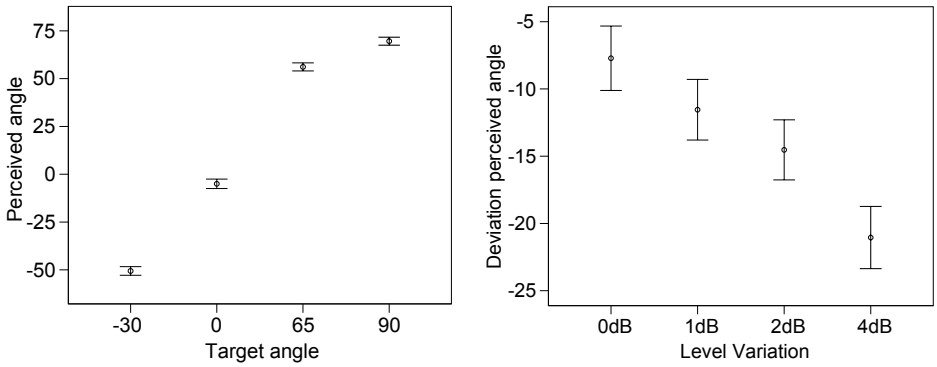


Figure 3.7. (**a**) Average of the perceived angles *versus* target angles (degrees); (**b**) average of the deviation of the perceived angles (degrees) *versus* level variation (dB). (95% CI)

The tendency of this angle deviation to the left can be seen in Figure 3.7 (b), considering the level variation applied (0, 1, 2, 4 dB).

An Analysis of Variance (ANOVA) indicates that the level variation has a very significant influence ($F = 27.338$, $df = 3$, $p < 0.001$) over the deviation in the perception of the angles of sound.

If we consider just the central angles used in the experiment (0° and 65°), a smaller average deviation can be seen (Figure 3.8). This leads us to believe that listeners tended to divert the location of the sounds perceived on the sides more, which means that the introduced level variations made the lateral angles disperse more than the central ones.

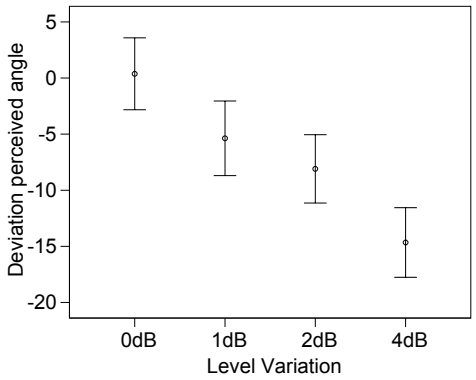


Figure 3.8. Average deviation of the perceived angles (degrees) *versus* level variation (dB), considering only the angles 0° and 65°. (95% CI)

On the other hand, the influence of the type of sound (timbal, voice, whistle or pink noise) on the deviation in responses can be seen in Figure 3.9 (a). Voice and pink noise have lower deviation than timbal and whistle sounds, especially in cases of 0 and 1 dB of deviation. Besides, voice stimuli and pink noise manifest a more separate and clear deviation at varying levels.

The influence of the type of sound over the deviation of the perceived angles is significant ($F = 4.409$, $df = 3$, $p = 0.004$) according to an analysis of variance. The sound angle reproduction has a very significant influence ($F = 54.932$, $df = 3$, $p < 0.001$) over the deviation of the answers. In Figure 3.9 (b), the deviation of the answers for each perceived sound angle reproduction is represented. Angles 0° and 65° present less deviation to the left. The biggest deviation of the answers corresponds to the angle $-30°$, and it could be due to the fact that it was the only angle on the left side.

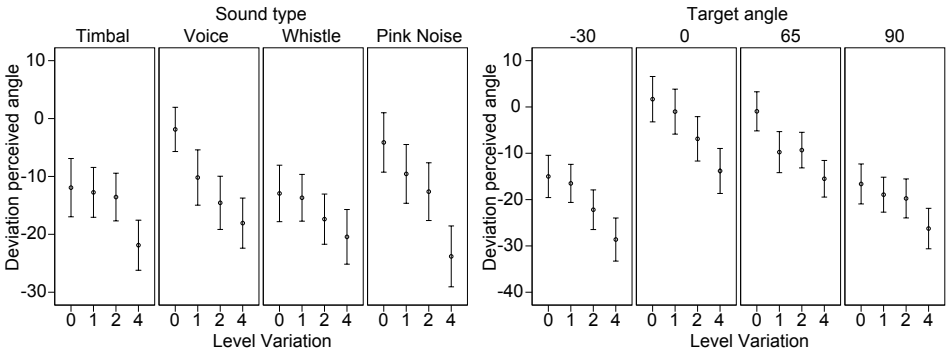


Figure 3.9. Average deviation of the perceived angles (degrees) *versus* the level variation (dB): (**a**) considering the type of sound; (**b**) considering the reproduced target angle. (95% CI)

3.1.4 Test 2. Frequency response on Quality and Spatial Impressions

Test description

In this test, participants listened to some excerpts of sound with headphones and rated their quality and their sound spatial image. These different headphones were simulated as described in Section 3.1.2 by means of the convolution of their frequency responses with the stimuli sounds, and all of them were reproduced with the reference headphones.

Due to the fact that different frequency responses produce noticeable effects, the perceptual test was designed according to the recommendation International Telecommunication Union, recommendation by Radiocommunication sector (ITU-R) 1534-3 [134], which describes the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) perceptual test. This kind of test describes a method to assess intermediate quality audio systems and also all of the requirements needed to accomplish the test with rigor. Besides, this test sets a zero to 100 continuous scale (zero – bad; 100 – excellent) to evaluate quality and other parameters of sounds and systems, always using a reference sound. All systems are compared to a reference of maximum quality, and the different systems are also compared between them.

Two different tasks were evaluated during the test by the participants.

The first task was to indicate the quality of the sound with respect to the reference. The second task was to evaluate the spatial impression (locations, sensations of depth, immersion, reality of the audio event) [20] with respect to the reference.

Five different excerpts of audio (12 to 14 s) were employed as source material (see Table 3.2), and all of them were reproduced simulating the different headphones under study. All of these sound fragments were chosen by their spatial, stereophonic and timbral attributes.

Artist	**Track**	**CD**	**Description**
Bettina Flater	*Haugebonden*	Women en Mi	female voice and guitar
Paco de Lucía	*Zambra Gitana*	Canción Andaluza	male voice and guitar
Jerry Glodsmith	*Night Boarders*	OST The Mummy	high dynamic orchestral
The Chad Fisher Group	*Basin Street Blues*	live	jazz (binaural)
Smashing Pumpkins	audience sound	live	audience and drums (binaural)

Table 3.2. Music program used for listening Tests 2 and 3

In this test, five headphones simulations were done, corresponding to headphones (b)-HQop, (c)-MQcl, (d)-BDso, (e)-LCmul and (f)-AirL (described in Section 3.1.2, with frequency responses in Figure 3.4). Each of the five sound excerpts previously mentioned were reproduced by the virtual headphone simulation described in Section 3.1.2. A virtual headphone simulation for each sound was presented randomly in series to the listeners, as well as a hidden reference ((a)-REF) and also two anchor signals. The first Anchor signal (ANC1) was a 7-kHz low pass filtered version of the sound (according to the mid-quality anchor of the ITU recommendation 1534-2 [134]), and the second Anchor signal (ANC2) was a monaural version of the sound. This second anchor was determined to set a reference for the spatial impression question.

To perform the test, a GUI was developed in MATLAB according to the recommendation [134], which allowed participants to freely listen to each of the sounds and to the reference, as many times as they wanted (Figure 3.10). The different sound fragments were presented randomly as a series with all of the different headphone simulations, to compare to the reference sound. Once the participant had scored all of the simulations of a series, a new sound excerpt was presented to be evaluated. This process was repeated twice, once for each question of the test (the first about quality and the second about spatial impression), with a pause in between.

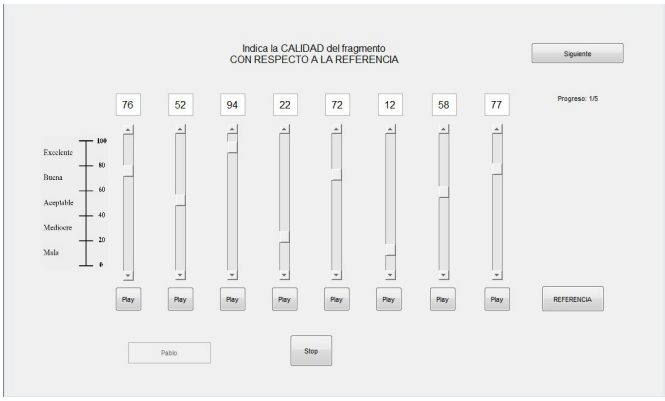


Figure 3.10. GUI of test 2

The number of stimuli of this test was: (5 headphones simulations + 1 hidden reference + 2 anchor signals) $\times$ 5 sound excerpts = 40 stimuli, presented in five series of eight stimuli plus the reference. As commented before, these 40 stimuli were presented twice in a different random order, to answer the two different questions.

The test was performed by 11 people, seven men and four women (21 to 37 years, with an average age of 30). As the test had two different questions, they were separated into two parts with a rest pause in the middle. The average runtime of the test was 22 min for the first part and 16 min for the second. Every participant did a training session before preforming the actual test, so all of them could listen to all of the stimuli and become familiar with the GUI and the assigned tasks.

Results

Figure 3.11 (a) shows the average of the normalized (zero to 100) answers about quality perception for the hidden reference, all five headphones simulated and the two anchors. As shown, the reference has been properly identified in most cases. The three supposedly good quality headphones have high scores; meanwhile, the two supposedly poor quality ones have the lowest scores. Both anchors remain in the middle of the scores of these two groups.

An analysis of variance confirms that the headphones have a very significant influence ($F = 58.33$, $df = 7$, $p < 0.001$) over the quality perceived.

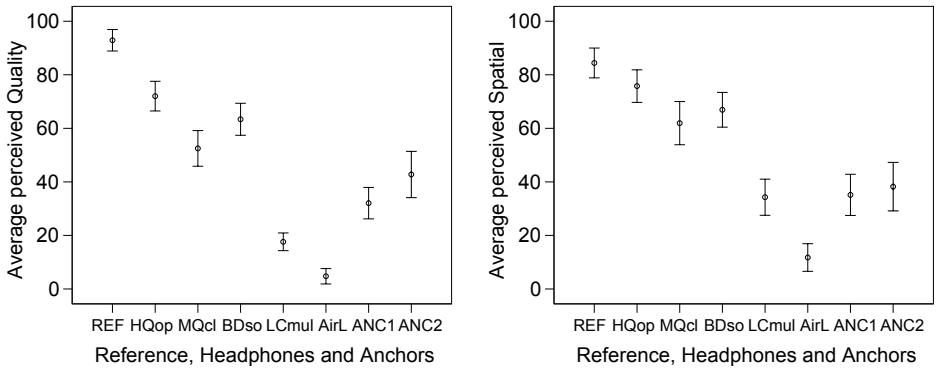


Figure 3.11. (**a**) Average perceived quality *versus* reference, headphones and anchors; (**b**) average perceived spatial impression *versus* reference, headphones and anchors. (95% CI)

Figure 3.11 (b) shows the average of the normalized (zero to 100) answers about spatial impression perception for the hidden reference, the five headphones simulated and the two anchors. The results present similarities with the answers about quality, with a high correlation of $r = 0.648$. Nevertheless, in this case, the confidence intervals are a bit wider, and the scores have some differences. The three supposedly good quality headphones have high scores again, but the confidence intervals do not separate them very much. There is a bigger difference between the two supposedly poor quality headphones, and the low cost multimedia ((e)-LCmul) ones are in the same range as both anchor signals. It is also noticeable that the Anchor Signal 2 (ANC2) as a monaural signal does not have a lower score.

In any case, an ANOVA confirms that the headphones have a very significant influence ($F = 58.33$, $df = 7$, $p < 0.001$) over the perceived spatial impression. No significant influence of the type of sound has been detected, even though some of them were binaural recordings.

3.1.5 Test 3. Non-linear distortion

Test description

The objective of this test is to evaluate the effect that the non-linear harmonic distortion present in each headphone model may have on the spatial impression.

Stimuli simulating the headphones with and without their direct non-linear harmonic distortion were presented to the participants who had to score their perception.

The effect of these distortions is very subtle. For that reason, the perceptual test was designed according to the recommendation ITU-R 1116-2 [135], which describes a method to assess small impairments in audio systems. This recommendation also establishes rigorous requirements of room, equipment and other arrangements. A continuous scale from one to five (1 – very annoying; 5 – imperceptible) is used to evaluate degradations with respect to a reference signal. The recommendation proposes an ABC test in which two stimuli, A and B, are presented to be compared against a known reference. One of these two stimuli, A or B, is always a hidden reference (in our case the simulation without non-linear distortion), and the other a degraded signal (in our case the simulation with the non-linear distortion).

One single question was presented to the participants: "What degradation of quality and spatial impression do you hear with respect to the reference?"

The same five audio excerpts previously described in Test 2 were used here (see Table 3.2), as well as the same five virtual headphone simulations (b)-HQop, (c)-MQcl, (d)-BDso, (e)-LCmul and (f)-AirL (described in Section 3.1.2, with frequency responses in Figure 3.4). No anchors beyond the proposed scale were used this time.

Two different versions of the headphones simulations were presented in this test. One with and the other without the distortion simulated with the method described in Section 3.1.2. These two versions of the same stimulus were presented each time to the participants. They have then to rate the distorted against the not distorted version of the same sound in a double-blind manner (A *vs.* B). In each trial, there was always a non-distorted version sound that acted as the known reference (C sound), which according to the recommendation [135] has to be compared to the A and B sounds.

The number of stimuli of this test was then: 5 headphones simulations $\times$ 2 versions (with and without distortion) $\times$ 5 sound excerpts = 50 stimuli, presented in twenty five series of two stimuli plus the reference. All of these pairs were presented randomly to each participant.

To perform the test, a GUI was developed according to the recommendation, which allowed participants to freely listen to each of the sounds to evaluate and the reference, as many times as they wanted.

The five headphones under study were simulated (including distortion) to be reproduced with the reference headphones ((a)-REF, frequency response in Figure 3.4).

This test was performed by the same 11 people of the previous Test 2; seven men and four women (21 to 37 years, with an average age of 30). The average runtime of the test was 16 min. Every participant did a training session before preforming this test, so all of them could listen to all of the stimuli and become familiar with the GUI and the assigned task.

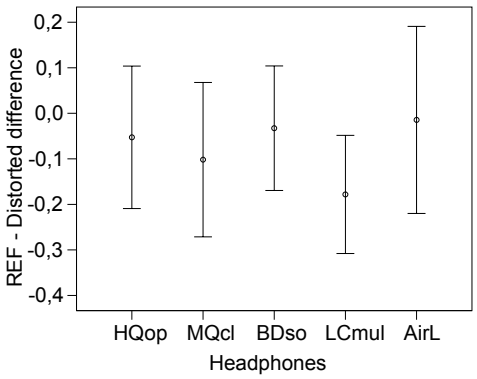


Figure 3.12. Difference between hidden reference and distorted signals *versus* headphones. (95% CI)

Results

According to the recommendation [135], the difference between the score of the hidden reference and the score of the degraded signal is analyzed. Figure 3.12 shows these differences for each of the headphones simulated.

No significant effect has been found. Then, the direct non-linear distortion produced by the headphones can be considered as imperceptible. Therefore, it has none effect in spatial perception, at least with the fixed level used to simulate all the headphones (69 dBSPL). In order to produce perceptible non-linear distortions, headphones should probably play at much higher levels, working at the limits of their linear behavior.

Later, in Section 3.2, another test is described that revisits nonlinear distortion, this time in a different scenario that takes into account equalization of the frequency response and different levels of reproduction.

3.1.6 Test 4. Frequency response on binaural azimuth localization

Test description

The results obtained in Test 2 are significant, but do not provide information about the accuracy in the localization of sources. For that reason, a test to evaluate the influence of frequency response on this accuracy was carried out.

Attempts to describe different spatial attributes have been a constant pursuit in the field of spatial audio [20, 136, 137]. The diffuse term employed in Test 2 to ask about spatial characteristics (spatial impression) was intended to relate in a simple way the perception of quality with the feeling of spaciousness. A more specific study of spatial attributes is then necessary to better evaluate the performing of the different headphones. In this direction, the localization accuracy in azimuth is one of the most studied spatial attributes [138, 139, 140, 141] and therefore a good anchor point to contrast the previous Test 2 with a localization experiment. Therefore, this test tries to establish a relation of the influence of the frequency response on the binaural azimuth localization in the horizontal plane.

As commented on in Section 3.1.2, to simulate the position of the sound sources in the horizontal plane, recordings of BRIRs in a medium-sized room were done. Nine different azimuth angles, 0°, 30°, 60°, 90°, 135°, 225°, 270°, 300° and 330°, were used.

Four types of sound were employed: door, voice (female), guitar and pink noise. A closing door is an impulsive sound with quite low frequency content, which can be useful for sound localization. The guitar sound was composed by various impulsive sounds in different main frequencies, one for each chord. Voice is an easily-recognizable common sound, and female was chosen to have some energy in high frequencies. The words "*estímulo sonoro*" (*sound stimulus* in Spanish) were employed. They present the repeated fricative phoneme /s/ with high frequency content and the phoneme /t/, a occlusive articulation that generates impulsive sound. Pink noise was employed to evaluate a wide spectrum signal.

For this test, seven different headphones plus a hidden reference were simulated (Section 3.1.2). Besides these, an additional anchor auralization (low pass filtered (LPF) sounds at 7 kHz) for each angle was employed (ANC1).

Therefore, the number of stimuli in this test was: 9 angles $\times$ 4 types of sound $\times$ (7 headphones simulation + 1 hidden reference + 1 anchor auralization) = 324 stimuli. These stimuli were presented in random order in two parts of 162 stimuli, with a rest in between.

To perform the test, a GUI was developed in MATLAB (Figure 3.13), which allowed participants to freely listen to the stimuli from a random list as many times as they wanted. Participants should indicate the perceived angle of the sound source. The GUI consists of a circle of points, which represents the top view of the listener, with a 5° resolution. Additionally, it included a parallel control to freely listen to a reference sound (pink noise) in the angles of 0, 45, 90, 135, 180, 225, 270 and 315 degrees.

The test was performed by 16 people, 10 men and 6 women (21 to 36 years, average age of 30). The average runtime was of 21 and 17 min for each part.

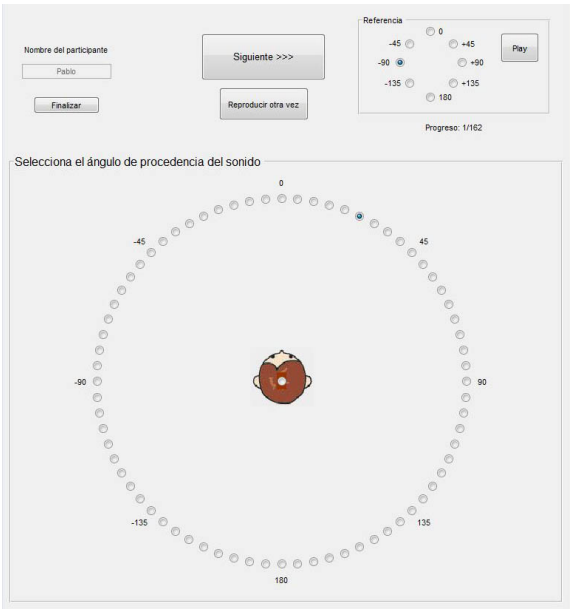


Figure 3.13. GUI of test 4

Results

A Cronbach's alpha analysis over the answers has been performed giving a value of $\alpha = 0.982$, which shows a high internal consistency.

A one-way ANOVA showed a significant influence between the headphones and the deviation of the perceived angle (deviation = perceived angle − target angle) ($F = 2.399$; $df = 8$; $p = 0.014$).

A first exploration of the participants' answers reveals that several front-back confusions [76, 142] occur. For this reason, an evaluation of the amount of front-back confusions was performed for each of the headphones simulated. An ANOVA showed that there is a very significant influence of the type of headphones on the number of front-back confusions ($F = 46.307$; $df = 8$; $p < 0.001$). In Figure 3.14, we can see that headphones (f)-AirL and (h)-LCmul2 produce an average of nearly 50% of front-back confusions. This can be logical, as both headphones are supposed to be in the low quality range. However, the (c)-MQcl headphone stands out in the group of high quality ones, as it has 30.2% of front-back confusions, more confusions than the (e)-LCmul headphone, with a significant difference. A comparison of the frequency response of the headphones that produce more front-back confusions ((f)-AirL, (h)-LCmul2 and (c)-MQcl) reveals that they share in common strong irregularities in the band of 100 to 1600 Hz. On the other side, other headphones of medium and low quality ranges that have less front-back confusions do not present these strong irregularities in that four-octave band. Because of that, we suspect this can be an affecting factor disturbing the front-back discrimination.

There is no significant influence of the type of sound crossed with the headphones. The sound guitar is the only one that produces slightly less front-back confusions for all of the headphones.

Due to the strong front-back confusion, the analysis of the deviation of the perceived sound with respect to the target reproduced sound will produce large angle errors with complicated analysis of the results. A front-back confusion produces a bigger error for sources in the median plane than lateral sources, avoiding an analysis of the deviation angle (perceived angle − reproduced angle) with respect to the source position.

To overcome this setback, it is proposed here a modified analysis of the error consisting of a preprocessing of the listener responses based on reflecting to the correct semi-plane the ones that have front-back confusion,

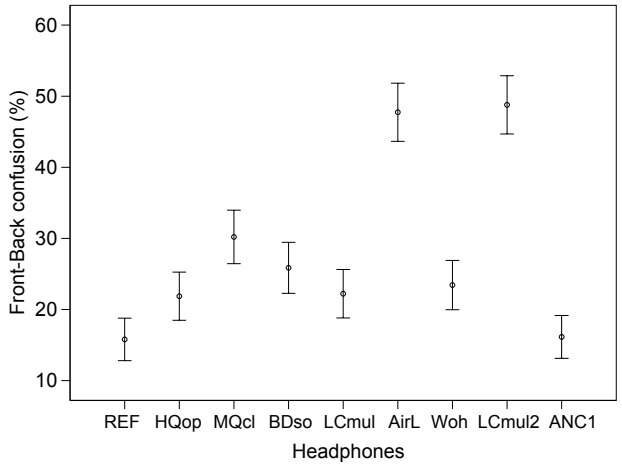


Figure 3.14. Percentage of front-back confusions for the reference, headphones and the anchor. (95% CI)

leaving untouched the ones that do not. This correction eliminates big jumps in the deviation, focusing the experiment in the performance analysis of the headphones reproducing correctly the main spatial cues as ITD and the low frequency part of Interaural Level Difference (ILD). The high frequency part is more related to the pinna effect that is not considered with the reflection applied.

Taking into account the strong front-back confusion, the analysis of the deviation of the answers from the target reproduction angle was performed introducing the correction of the front-back confusion. Therefore, a symmetric image of the responses in the back (90° to 270°) is brought to the front.

Figure 3.15 shows the deviation of the perceived angles with respect to the target reproduction angle of the sounds, both of them front-back corrected. We can see that the deviations are quite uniform across the different headphones, except for the angles 90° and 270° in the cases of (f)-AirL and (h)-LCmul2. Looking at Figure 3.4, it is easy to see that the frequency responses of these two headphones present irregularities and deep level drops between 4 and 7 kHz. It is noticeable that the anchor LPF 7-kHz sounds auralized in the different angles (ANC1) are not affected by this problem, supporting the suspicion that the commented band is important for sources located in lateral positions.

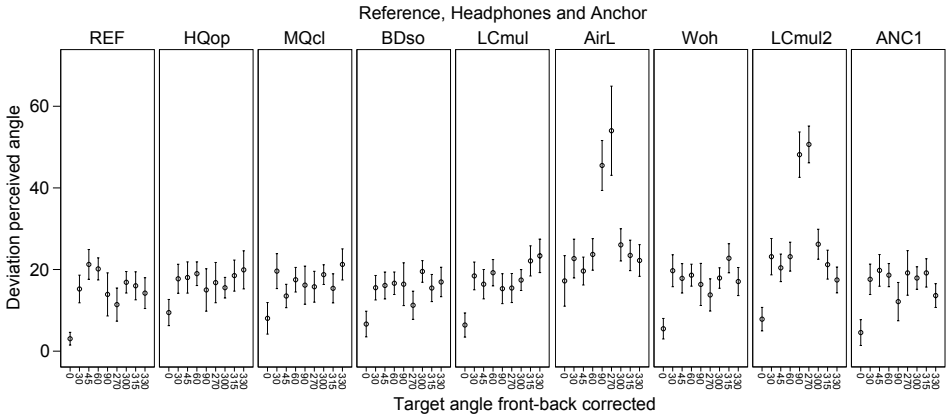


Figure 3.15. Deviation in degrees of the perceived angles with respect to every target reproduced angle of sound. The reference, headphones under testing and anchor are represented. (95% CI)

3.1.7 Conclusions

This study investigates the influence of different headphone parameters in the context of spatial sound reproduction. Four different perceptual tests have been done to analyze: (1) the effects of the sensitivity disparity between the transducers; (2) the influence of the frequency response over the perception of quality and the spatial impression; (3) the effects of non-linear distortion; and (4) the influence of the frequency response over binaural azimuth localization.

The following main conclusions can be drawn:

1. The sensitivity disparities between left and right transducers affect the localization of sound sources, starting from level differences of 1 dB.

2. The quality and uniformity of the frequency response have an important influence in the *spatial impression*.

3. Additionally, the *spatial impression* has a high correlation with the subjective *perceived quality*.

4. The binaural recordings do not obtain significant better results for the parameter *spatial impression* compared to two-channel stereo mixes.
5. The direct non-linear distortion introduced by consumer level low quality headphones does not affect the perception of the spatial sound image.
6. It has been ratified that much front-back confusion is produced, both for high and low quality headphones.
7. Irregularities of the frequency response in the band of 100 to 1600 Hz seem to especially affect the front-back discrimination.
8. A poor response in the band of 4 to 7 kHz has been found to degrade the accuracy in lateral position localization.

All of these conclusions have been supported with statistical and ANOVA analysis. Some other interesting comments and clarifications about these conclusions can be added:

In addition to Conclusion 1, the angles chosen in the disparity test are a determining factor, whereby the more lateralized the angle, the larger the deviation. An increased number of angular positions may be of interest in later studies.

In relation to Conclusions 2 and 3, it is worth remarking that the mono anchor signal (ANC2) has obtained equal or even better results for *spatial impression* than some headphones ((e)-LCmul, (f)-AirL) and the stereo LPF anchor (ANC1). This fact seems to be in relation to a deficient high frequency reproduction and the general listening sensation, as evidenced by the high correlation statistics obtained with the parameter *perceived quality*.

In relation to Conclusion 4, the binaural recordings employed do not seem to significantly produce a better ranked spatial impression than stereo mixes, probably because the binaural recordings employed were made with a generic dummy head and not synthesized with each individual's HRTF.

In relation to Conclusion 5, other works, such as [143], have not found significant perception of the distortion. However, this earlier study used high quality headphones, while the study presented here does so also with low quality consumer headphones, and also analyzing the influence on spatial reproduction.

Finally, taking into account these three characteristics, *perceived quality*, *spatial impression* and accuracy in *azimuth localization*, it has been concluded that the first two are highly correlated. Surprisingly, and contrary to how it might seem *a priori*, there is virtually no correlation between spatial impression and accuracy in localization, because the strong influence that the subjective perceived quality has over the spatial image perception. An illustrating example can be seen with the (f)-LCmul headphone. It would be interesting to deepen this relationship in future work.

Based on the results of this study, some general guidelines for the design of headphones suitable for spatial sound reproduction can be suggested. A sensitivity difference between left-right transducers less than 1 dB should be assured in the manufacturing process to avoid azimuth localization errors. A flat frequency response between 100 to 1600 Hz is desirable to reduce front-back confusion. Finally, a frequency response with enough resolution in the band 4 to 7 kHz would guarantee a good accuracy in the localization of lateral sources.

Many of these requirements can be addressed by careful equalization of the frequency response. Correct equalization would help to eliminate the disparities between left and right transducers, and to reduce the negative influences of frequency response on quality perception and spatial impression. In addition, with binaural sound the equalization is presented as needed to correct the response resulting from the coupling between headphone and ear. With this frequency correction much of the front-back confusion is eliminated and the externalization of binaural sound is improved [144]. The problem that arises is what degree of equalization can be applied to consumer headphones without non-linear distortions being noticed due to forcing transducers. How much non-linear distortion does a strong equalization produce in a consumer headphone? Is the non-linear distortion generated in this case audible? At what playback levels? This problem is studied in the next section with another perceptual experiment.

3.2 Perception of non-linear distortion caused by equalization

3.2.1 Introduction and motivation

The interest in headphone technology is fostering new techniques for headphone products. A clear possibility to enhance the perceived quality of a headphone is the use of active signal processing. Previous works suggest that frequency response is a dominant factor in the perceived quality [145, 146], this fact has focused the studies of subjective user preferences on different target frequency responses [128]. Some recent headphones and software for headphones put the attention in the individualization of the listening experience by shaping or tailoring the frequency response [99, 147]. These works and new products are showing that with precise measurements and equalization it is also possible to mimic the frequency response of a specific headphone with a different headphone model [124, 128]. A virtual headphone emulation by equalization has some limitations, as transducers could not have capacity to reproduce specific frequencies, for example, the lowest or highest frequencies cannot be reachable for certain consumer headphone models. Besides, a strong equalization can force a transducer to work out of its linear condition creating non-linear distortion effects. Despite non-linearities are difficult to perceive due to masking effects, if some equalization is applied audible distortion can be greatly increased, which can degrade the final perceived quality.

The objective of the experiment presented in this section is to examine the perception of non-linear distortion when applying equalization over different consumer headphone models. To this end, a combination of techniques has been used by mixing a simulation of the non-linear distortion measured in each of the headphone models, together with a virtual headphone listening test methodology.

In order to test the effect of equalization, the frequency response of a high-end headphone has been chosen as target response, and other mid- and low-end headphones have been made to emulate this target response by equalization. The distortion produced by each headphone after the equalization must then be measured, along with the linear response obtained. With these parameters, a virtual headphone listening test has been performed, comparing samples of the equalized headphones with and without non-linear distortion.

3.2.2 Measurements, equalization and non-linear distortion simulation

Headphones selection

All the measurements of the headphones and the reproduction of sounds were done with a Head and Torso Simulator (HATS) model B&K Type 4100 and a MOTU Traveler sound card. A Sennheiser HD800 was chosen as the reference headphone because of its high quality, using its frequency response as the target to be emulated in the different headphones under test. Besides, this reference headphone was used to perform the virtual listening test. To emulate the target frequency response, eleven headphones of different qualities and prices were selected. They were intended to cover a wide range of possible common uses. The brands and models of the rest of the headphones are provided in Table 3.3 for reference only. No advertising is intended and besides these brands are not fully represented in this study.

	model	**characteristics**
(Ref)	Sennheiser HD800	open and circumaural
(1)	Sennheiser HD650	open and circumaural
(2)	AKG K7XX	open and circumaural
(3)	Beyerdynamics DT990	open and circumaural
(4)	Sennheiser HD429	closed and circumaural
(5)	August EP650B	closed and supra-aural
(6)	Superlux HD668B	semi-open and circumaural
(7)	Nubwo NO-3000	closed and circumaural
(8)	Woxter i-Hph 780	closed and supra-aural
(9)	Genius HS-04SU	supra-aural
(10)	McDonalds gift	closed and supra-aural
(11)	Airfrance (provided on-board)	supra-aural

Table 3.3. Headphones used in the study about non-linear distortion

It is very important to emphasize here that the number assigned to the headphones used is a consequence of the conclusions of the experiment. The headphones are listed in this order throughout the description of the experiment only to be consistent and to facilitate their review in the conclusions obtained.

Calibrated measurements with non-linear distortions after equalization

Non-linear distortion depends on the signal level applied to the headphones at each frequency. For this reason, all the headphone measurements carried out in this work have passed a calibration process. The calibration takes into account all the measure chain, including the HATS microphones sensitivity (mV/Pa) and the electrical full scale value (mVFS) of the sound card. The IEC 61672-1 [148] recommendation for A weighting pressure levels was employed.

In a first stage, the linear responses of the headphones were measured and a filter was calculated to model the target response. A logarithmic sweep of 5 seconds was used, covering 20 to 20000 Hz without pre-ringing effects. Both, left and right transducers were measured with a sampling rate of 48 kHz. The following steps were implemented to obtain filters that emulates the target response over each tested headphone:
1 - Frequency response smoothing of 1/6 octave for every headphone.
2 - Calculation of the headphone's inverse filter by direct inversion of the measured response, from 20 to 20000 Hz.
3 - Spectral product of the inverted frequency response of the tested headphone and the target frequency response.
4 - Limitation of the filter gain to +20 dB to avoid excessive boost at certain frequencies.
5 - Calculation of the equivalent minimum phase filter.

In a second stage, the individual target response filters were applied over each headphone and the non-linear distortions generated by these response corrections were measured. This is the key point of this experiment, because it allows to evaluate the behavior of a poor quality response headphone in terms of distortion, when it is corrected by equalization. The steps to accomplish these measurements were:
1 - The synchronized swept-sine [149] employed for measurement is filtered by the equalization filter for each headphone.
2 - The resulting test signal is reproduced for each headphone at six differ-

ent calibrated reproduction levels of 70, 80, 85, 90, 95 and 100 dBA. These levels have been calibrated in the HATS ears for 20-20000 Hz pink noise. 3 - With the recorded responses, second and third-order nonlinear distortions generated by the headphone were computed using the method described in [149].

Following this equalization procedure, each of the tested headphones were forced to try to emulate the same target frequency response at the same acoustic pressure levels, which means that they were compelled to work in the same conditions. Figure 3.16 shows the linear magnitude responses achieved for each headphone after the equalization and the first two harmonics of distortion generated. To represent the second and third distortion orders for the six reproduction levels clearly at a glance, they have been drawn as two shaded areas where the lowest limit corresponds to 70 dBA and the highest to 100 dBA, therefore intermediate levels fall into the shaded area. The linear responses were almost identical for all six reproduction levels, so the average is represented.

Reposition of the headphones is a convenient procedure for the frequency response measurements [98]. However, the described measurements were done without reposition. In this experiment we want to focus on the distortion generated during the emulation of a target frequency response, therefore, the correction of a mean frequency response would generate an unreliable measure of the distortion, as the final measurement of the emulated response with its distortion also depends on a fixed position. As the subjective test performed with these measures concentrates in the perception of the distortion, the possible variations in the linear frequency response were considered not relevant here.

Virtual simulation of the non-linear distortion

To simulate the non-linear distortion of each headphone obtained after the equalization, the method described in [131, 132] was chosen, which uses a Diagonal Volterra kernels model and a series of linear convolutions.

This methodology and the equations employed have been previously described in Section 3.1.2. Following this method, the output signal of a non-linear system can be estimated by means of a truncated Volterra series. With Equation 3.3 the Diagonal Volterra kernels can be computed for the simulation of the main response and the first and second order distortion harmonics. Then, Equation 3.4 describes the truncated linear convolution

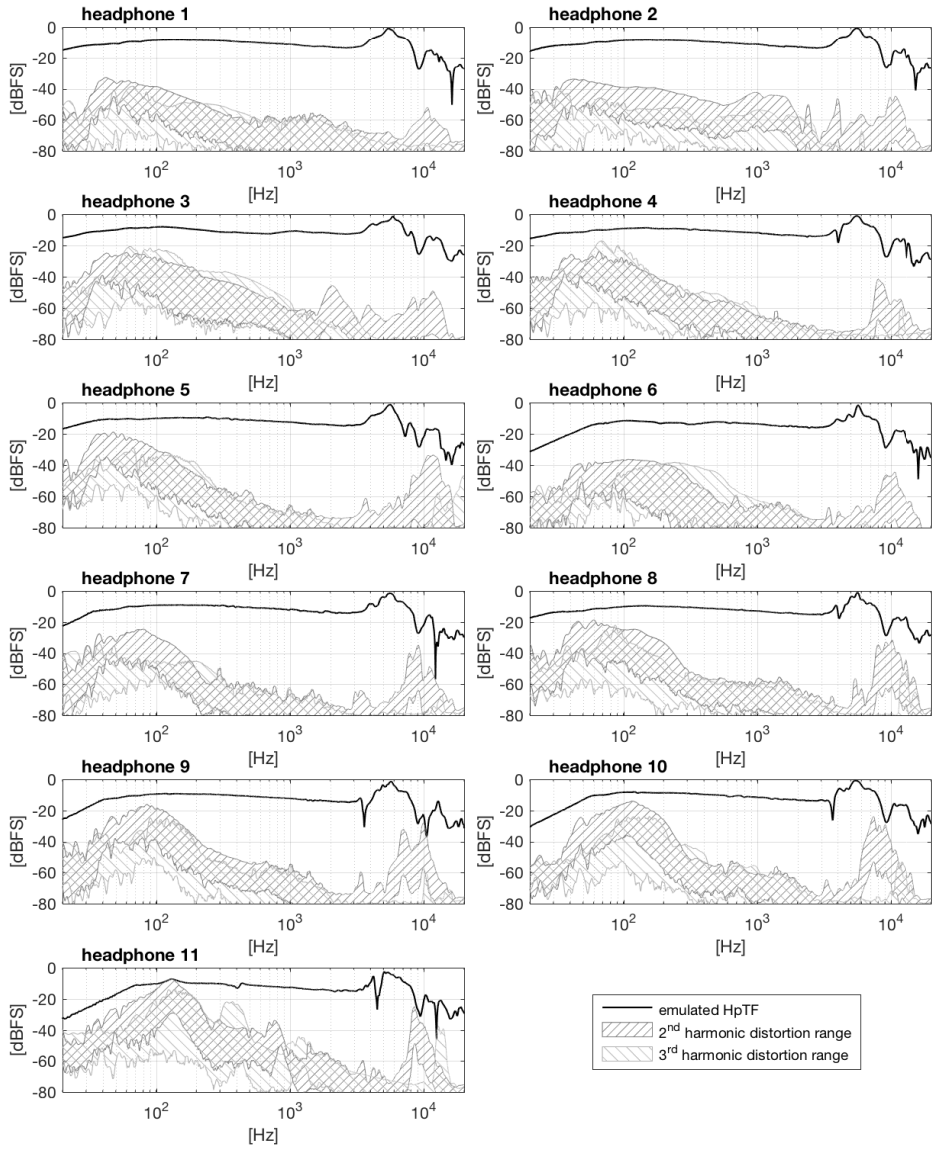


Figure 3.16. Linear frequency responses achieved for each headphone after the equalization, and the first two non-linear distortion harmonics generated (range from minimum to maximum levels measured). (Left channels)

series to simulate the second and third order non-linear distortions. The linear part of the system response can be simulated independently using only the first Diagonal Volterra kernel.

Applying this technique to a sound stimulus, it is possible to simulate the effect of a frequency response with and without the measured non-linear distortion.

3.2.3 Perceptual test

The purpose of this test is to verify whether the distortion produced by each equalized headphone at different reproduction levels can be perceived or not. To avoid visual and tactile biases, all the different headphone emulations and their distortions measured were simulated through the reference headphone, in a virtual simulation listening test. Wearing just the reference headphones, the subjects performing the test can have immediate access to the different headphones and the procedure of the test becomes more flexible, transparent, controlled and repeatable [129]. This methodology is desirable due to the differences in appearance, fitting and range of qualities of the headphones employed.

The process of generating the stimuli for the virtual headphone listening test consists of using the method described in Section 3.2.2 and applying the compensation filter of the reference headphone response. This filter was obtained with an automatic regularized method for the inversion of the frequency response, which produces perceptually better equalization than the regularized inverse method with a fixed factor [97]. In this case, the mean of five repositioned measurements of the reference headphone response was used.

To evaluate the perception of the non-linear distortion the stimuli generated with and without distortion were presented to ten expert listeners [135] by means of an ABX test. The different emulations achieved for each headphone were presented randomly. Positive detections of the distortion were considered at a significance level of $\alpha \leq 0.02$ [150]. The sound clip employed to generate the stimuli of the test is accessible online [151]. It was selected because of its rich low frequency content and large dynamic range. The reproduction level of the stimuli during the test was fixed at 85 dBA (slow, [148]). With this ABX test, subjects identified the minimum level at which they detected the non-linear distortion for each of the em-

ulated headphones. The number of stimuli generated were 11 headphones x 6 levels x 2 with-without distortion = 132 stimuli. Absolute ecological validity is achieved just for the fixed reproduction level. This procedure has been employed in other works [143], suggesting that for the rest of the levels, if anything, this should result in an increased sensitivity of subjects to audible distortion.

Results

The results of the ABX test can be seen in Figure 3.17. Mean of the minimum reproduction levels with distortion detection are shown for each headphone. In addition, prices of the headphones were determined with the average of the retail price ($USD) during the previous 24 months. Thanks to this, headphone models are sorted by the retail price, with number 1 as the most expensive and 11 as the cheapest. A high correlation has been found between the minimum detection level of the distortion and the retail price ($r = 0.91$ $p < 0.001$). Besides, three groups of headphones can be identified: Group A - headphones 1 and 2, priced above $200; Group B - headphones 3 to 7, priced from $200 to $20; and Group C - headphones 8 to 11, priced at less than $20. These groups have been found to correspond with different detected distortion levels, group A with an interval from no distortion detected to some detections at 100 dBA, group B presents a range of distortion detected from 95 dBA to 80 dBA and group C have a range of distortion detected from around 80 to 70 dBA.

It is interesting to point that in group B most of the headphones have a pricing range from $100 to $20, except for headphone 3 that exceeds $100. This model is a bit out of the distortion detection trend, with slightly worse results according to its price.

The main idea that emerges from these results is that for most of the headphones the distortion is not noticeable at comfortable listening levels up to 80 dBA. Group A headphones with only a few detections at 100 dBA can be considered to produce an almost undetectable distortion, as headphone users would rarely listen at levels over 95 dBA.

3.2.4 Conclusions and future work

The equalization of the frequency response of headphones can produce audible non-linear distortion, depending on the reproduction level.

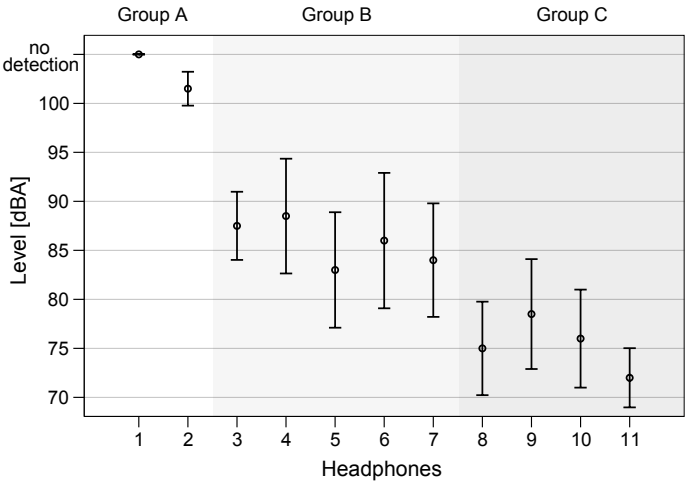


Figure 3.17. Mean of the minimum reproduction levels with detected distortion for each headphone (95% Confidence Intervals). Headphone models are sorted by the retail price

In this experiment, a method has been implemented to measure and simulate the non-linear distortion produced by the emulation of a target frequency response (Sennheiser HD800) by equalization over diverse consumer headphones. This approach has allowed to simulate the frequency response achieved with the equalization, with and without the non-linear distortions generated. Six different reproduction levels were analyzed on eleven headphone models.

An ABX test was performed by ten expert listeners to evaluate the audibility of the non-linear distortion generated at the different reproduction levels. High correlation has been found between the level of reproduction at which distortion is detected and the retail price of the headphones, with negligible detections in expensive models and a gradually increasing perception of the distortion as the price is reduced.

Some studies indicate that frequency response and retail price of headphones have no correlation [145], but this experiment suggests that retail price can have a direct correlation with perceived non-linear distortion caused by equalization. Despite this, the frequency response equalization is shown to be a viable technique that does not produce disturbing distortion at moderate listening levels with medium quality headphones and not

noticeable non-linear distortion in the case of high-end headphones.

Future work

As a main conclusion we have seen that in most cases, even strong equalization can be used to modify the response of the headphones without the non-linear distortion being appreciated. This ensures the usefulness of equalization for the correction of the acoustic coupling of headphones for binaural sound, and also for the design of headphones with active signal processing. In fact the multimedia industry is showing great interest in this type of products.

Other acclaimed studies based on numerous perceptual tests have proposed a statistical model to predict the listeners' preference ratings of headphones [119, 120], but without taking into account possible nonlinear distortion. It would therefore be interesting to propose a model that predicts the audibility of the non-linear distortion of a headphone after equalization. For this purpose, the methodology proposed here could be used and a psychoacoustic model could be added to indicate whether or not the nonlinear distortion is perceptually masked. This model would predict the audibility of the non-linear distortion of a headphone without the need of its evaluation by means of costly perceptual tests.

4 HRTF measurements

4.1 Introduction and motivation

Binaural sound employs *Head Related Transfer Functions* (HRTF) to generate the spatialization of sounds to listen with headphones. The HRTF captures the effects that a source in free-field experiences to the ear canals of a subject. The contribution of the head and torso and significantly the outer ear, are registered in the HRTF [83]. Due to the strong influence of the anthropometric characteristics of the listener (size of the head, position of the ears, shape of the pinna, etc.) the HRTFs present tailored features that make them specific for each subject. Then, for a better experience, the binaural sound must be individualized for each subject through the use of their personal HRTF. Individualized HRTF provide a better immersive and natural listening experience [85]. It is possible to accurately simulate virtual sound sources at any position in space by simply convolving the source signal with the HRTF and listening to them through headphones.

The time-domain equivalent of the HRTF is named *Head Related Impulse Response*, HRIR. The definition of HRTF/HRIR in particular relies on the free-space, that is, anechoic acoustic conditions to separate the human receiver characteristic from other room acoustic characteristics. If an

acoustic enclosure is meant to be involved in the HRIR, this is indicated by using the term *Binaural Room Impulse Response*, BRIR.

There are different techniques that try to obtain personalized HRTFs: measuring directly the subjects HRTF in an anechoic chamber, through anthropometric data (including synthesis calculation with numerical methods), based on subjective perception, etc. [12]. The traditional method for measuring HRTFs was described by Blauert in [1]. It uses a single loudspeaker mounted in a positioning system that measures the acoustic transfer path between the loudspeaker and the two microphones inserted in the subject ears. Using the positioning system, the loudspeaker is moved to different points from a virtual sphere around the subject, and the acoustic paths from these locations are measured. Despite the minimum audible angle is around 1°-2° for frontal sources [152], the HRTF set resolution should be around 5° in the horizontal plane and 10° in the vertical plane for common applications. According to this rule, the number of HRTF measurement points should be bigger than 1000. However, different interpolations techniques have been proposed, where using less points, produce successfully results in practice.

In any case, because of the large number of measurement points (from hundreds to thousands) the total time for a personal HRTF recording session could extend to even more than one hour with the traditional method were only one loudspeaker or just a few are employed. This causes fatigue and discomfort in subjects because they should not be moved to avoid measurement errors.

To reduce the measurement time, the most obvious method is to install multiple loudspeakers in multiple positions to save the time required for positioning system movements. However, one loudspeaker in each measuring position is impractical, because it would require about a thousand units. Therefore, hybrid combinations have been used, using multiple loudspeakers to cover a polar angle and a single-axis positioning system to cover the other. The most common hybrid method uses as many loudspeakers as there are elevations to be measured, installed in an arc around the listener. In this way, the single-axis positioning system rotates the arc structure around the listener or rotates the listener seated on a turntable [153, 154].

There are also methods for simultaneous measurement of different HRTFs using multiple loudspeakers. A *Multiple Exponential Sweep Method* (MESM) was proposed in [154] and refined in [155], which allows the si-

multaneous playing of sweep signals through various loudspeakers, saving even more time.

The HRTF refers to a free field measurement, so the measurement inside a room will introduce reflections that are not part of the HRTF. Anechoic conditions are then generally necessary for the HRTF measurement environment. The measurement systems described above should be installed in an anechoic chamber, setting up a complex, very expensive and not everywhere available installation, restricting these measurements to research laboratories and keeping these technologies away from the general public.

Some of organizations and research centers have chosen to provide their measurements as publicly available databases to the community. In the following, a few are listed in chronological order:

– KEMAR, the MIT Media Lab HRTF Database [156]: This early database for the Knowles-Electronics Mannequin for Acoustic Research, KEMAR, represents an extensive, but non-individual recording. 710 different positions were sampled at elevations ranging from $-40°$ to $+90°$ in $10°$ increments with regard to the horizontal plane and roughly $5°$ azimuth spacing per elevation at 1.4m distance between loudspeaker and KEMAR.

– AUDIS, the AUDIS Catalog of Human HRTFs [157]: This database was one of the output results of an European Union funded project on auditory displays. Here, measurements were done at 2.4m distance to the loudspeaker, $10°$ spaced elevations from $-10°$ to $+90°$, and an azimuth spacing of $15°$. The total measurement then comprised 122 directions for each of about 20 individuals. Moreover, round-robin tests have been performed with four contributing partners to analyze differences in the data across different laboratories.

– CIPIC, the CIPIC Lab HRTF Database [158]: 45 individuals' HRTFs were measured at high spatial resolution, including KEMAR with large and small pinnae. The spatial sampling is mostly uniform with $5°$ spacing in both elevation and azimuth, resulting in 1250 sampling points on the 1m radius auditory sphere. The database also includes a set of individual anthropometric measurements for each subject. Additional documentation and Matlab utility programs are provided with the database.

– LISTEN, the IRCAM HRTF Database [159]: Again developed in an EU project, this database contains blocked-meatus HRTFs for about 50 in-

dividuals, at elevations from $-45°$ to $+90°$ in $5°$ increments with roughly $15°$ azimuth spacing, resulting in 187 positions in total. It provides raw HRTF measurements, optional diffuse-field compensation and morphological data.

– ARI, the Acoustics Research Institute HRTF Database [160]: The number of measured subjects has been growing over time and lately it comprises high-resolution HRTFs of more than 200 individuals. Most of them were measured using in-ear microphones, but for a few further ones behind-the-ear microphones placed in hearing-aid devices were employed. 1550 positions were then measured for each listener, including the full azimuth-circle (with $2.5°$ spacing in the horizontal plane) and elevations from $-30°$ to $+80°$. Anthropometric data of 60 individuals are also provided.

– FIU, the Florida International University DSP Lab HRTF Database [161]: It contains HRTF data from 15 individuals at twelve different azimuths and six different elevations, at an unusual sampling frequency of 96kHz. It further includes 3D images of the persons pinnae and related anthropometric measures of various parts of the pinnae.

– RIEC, the Research Institute of Electrical Communications (Tohoku University) HRTF Database [162]: It includes the HRTFs of 105 subjects, measured with a double circular array of speakers. The number of source directions per subject ear is 865, distributed at $5°$ intervals along the azimuth and $10°$ intervals of elevation from $-30°$ to $+90°$ in spherical coordinates. Besides, anthropometric data of 39 subjects are also provided.

– HUTUBS, the Huawei - Technical University of Berlin - Sennheiser HRTF Database [163]: This database consists of measured and simulated HRTFs, measured HpTFs of two headphone models, 25 anthropometric measurements per subject, and 3D surface meshes of 96 subjects. A grid of 440 spatial points is provided for each subject. The data were cross-evaluated showing a good agreement between repeated measurements and between measured and simulated data.

All the described databases are of tremendous value to research, as they provide public data of sophisticated measurements done in very expensive prototypical installations. The personal HRTF of the different persons measured can be used to study the individualization of the HRTF, besides that these measurements can be use to synthesize non-individual binaural sound. But if you want to study the individualization of HRTF taking into account individual perception, and using subjective tests incorporating

individual HRTF, it is necessary to have the individual HRTF of people who can subsequently listen to them. It is therefore imperative to have a measurement system of HRTFs with which to obtain personalized measures that can then be used with the same measured subjects. Otherwise, it would not be possible to test and evaluate individual HRTFs.

According to the above, the aim of the work described in this chapter is the construction of a system for real HRTF measurements taking into account two objectives:
1 - to obtain HRTFs in a more accessible way, less tiring for the measured person and with a more affordable installation, i.e. in a faster way and without an anechoic chamber.
2 - the possibility of using personal measurements to carry out research on individualization of HRTF, being able to make perceptual tests that include the individual HRTF of the subjects.

Following these premises, a complete system for HRTF measuring was constructed from scratch, including room conditioning, loudspeaker configuration, adapted microphones, and measurement and post-processing software. The system actually measures BRIR and *quasi* HRIR are obtained by post-processing the measurements.

4.2 Constructed HRTF measurement system

4.2.1 Room conditioning

The room where the measurement system was built is located in the CPI research building that belongs to the Universitat Politècnica de València. It has a size of $9.88 \times 4.90 \times 2.65$ m, with an original reverberation time [164] of $RT_{60} = 698$ ms. The highly reflective and parallel surfaces of the largest dimension of the room produced particularly annoying reflections known as "flutter echoes" [165]. An acoustic treatment with physical panels was carried out to partially reduce reverberation, and to eliminate the main reflections from the walls, including the "flutter echoes".

More than forty acoustic absorbent panels were built to modify the acoustics of the room. Each panel consist of a wood frame with mineral wool absorbent material *ISOVER Acustilane 70* 50 mm thick and $30 Kg/m^3$ density, with an average absorption coefficient of $\alpha_w = 0.7$) covered with

fabric. The walls of the longest dimension of the room were covered with 138×65.5 cm acoustic panels (Figure 4.1). Two folding screens (with panels of the same materials of 200×65.5 cm) were also built for the shortest dimension of the room, and placed between the loudspeaker array and the walls, reducing the effective size of the room in this dimension. The ceiling had already an acoustic treatment with hanging acoustic panels, and the floor has not been treated.

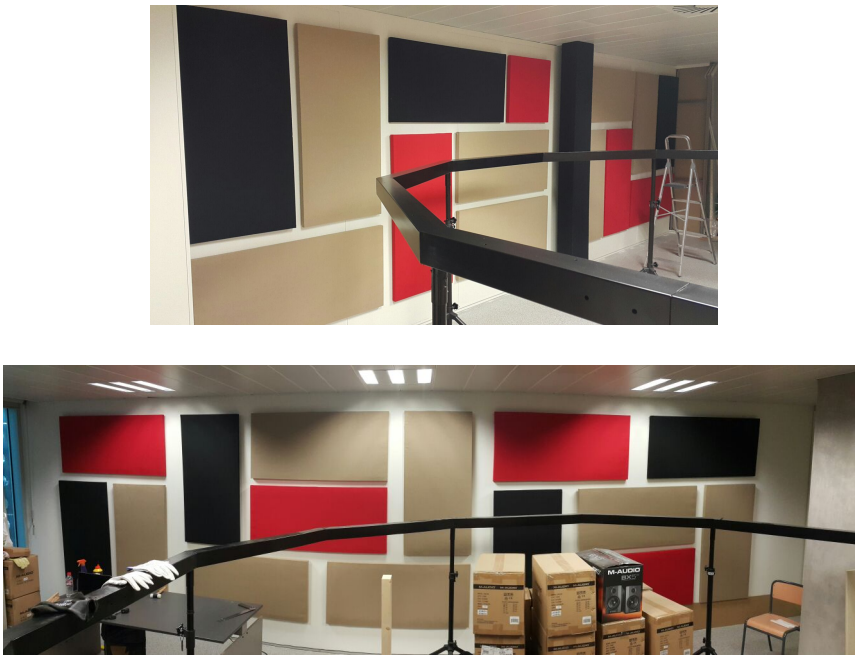


Figure 4.1. Absorbent acoustic panels for the conditioning of the HRTF measurement room

With the panels and the final refurbishment, the reverberation time is reduced to $RT_{60} = 169$ ms. Therefore, the direct measure obtained in the acoustically treated room was still a BRIR. The rest of the reverberation was removed by post-processing of the measurement, as described in Section 4.3.

Using a non-anechoic room to measure HRTF, overcoming the inherent difficulties of capturing reflections, is an innovative option to make individual HRTF measurements more affordable and therefore binaural spatial sound more accessible.

4.2.2 The speaker set-up and hardware

The loudspeakers used are self-powered *M-Audio BX5 D2* monitors consisting of a 5" woofer and a 1" tweeter, providing a ±1 dB flat frequency response between 53 Hz and 22 kHz. Their good performance makes it almost unnecessary to correct their response in measurements.

Using a two-way speaker system introduces an angle error in elevation because a point source is not used. However, since the radius of the array is 2 meters, the error at this distance is negligible. In addition, a two-way configuration offers other advantages, such as an extended frequency band covering almost the entire audible spectrum.

Figure 4.2. Loudspeaker set-up for HRTF measurements in the acoustically conditioned room

Human hearing is more sensitive to localization in the horizontal plane [1]. Because of this, a circular array of 72 loudspeakers separated by an angle of 5° has been employed in the set-up to cover the horizontal plane of listening. The radius of the circle is exactly 2 meters, and the loudspeakers are exactly touching each other in the front, avoiding any gap between them, thus reducing possible diffraction artifacts. A specifically designed

support system has been built to hold the complete horizontal array. Two additional 8 loudspeakers circular arrays of 1 meter radius were deployed on the ceiling and floor with an angular resolution of 45°, and aiming at the listener's position at +45° and −45°, respectively. Figure 4.2 shows the set-up of all the 88 loudspeakers inside the acoustically treated room.

All these loudspeakers are connected to four *MOTU 24I/O* soundcards, synchronized and controlled by the same single PCI Firewire card (Figure 4.3).

Figure 4.3. Soundcards rack used for HRTF measurements

To record the responses of the subjects, two miniature omnidirectional microphones *Knowles FG 23329-P07* are placed into the ear canals, inserted into holed soft foam earplugs. The earplugs are shortened to allow a full insertion into the ear canal. The fact that they are disposable earplugs allows for both hygienic and economically affordable use to record the HRTFs of many subjects. The phantom power adapter was also manually made to supply the 1.2 V to polarize the capsules of both microphones (two versions were built, one powered by AA batteries and another by standard 48 V phantom power from conventional soundcards). A dedicated *Roland OCTA-CAPTURE* soundcard with high quality microphone preamplifiers

is used to receive the signals from the microphones. The Figure 4.4 shows the microphones dimensions, the modified earplugs and the phantom power adapters.

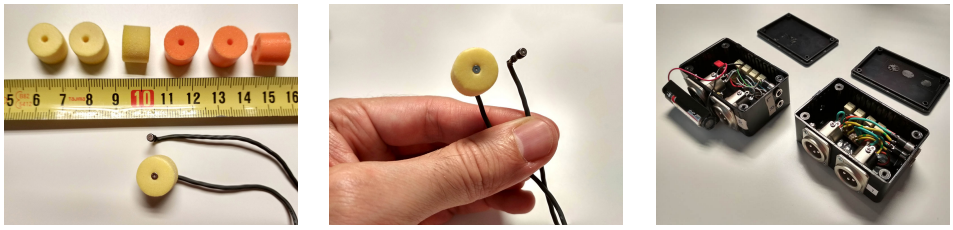


Figure 4.4. Miniature microphones, modified earplugs and phantom power adapters used for HRTF measurements

The task of correctly positioning the subject to be measured is quite time consuming if precision is sought, with the consequent annoyance for the person to be measured and the prone to errors in the measurements. To assure a fast and accurate oriented positioning four laser pointers were attached to the horizontal array at the angles 0°, 90°, 180° and 270°. Using these lasers as a reference, the people to be measured are properly placed in just a few seconds and excellent accuracy can be achieved in their orientation. A detail of one laser pointer and the process of adjusting the lasers are illustrated in Figure 4.5.

Figure 4.5. A laser pointer in place (right) and process of adjusting them (left)

4.2.3 Measurement software. Exponential sweep and Multiple exponential sweep method

There are different techniques to measure the impulse response IR of a linear system that have been applied to the specific case of acoustic measurement. Some particular issues have to be considered in this case of application. The signal-to-noise ratio (SNR) of an acoustic measurement is affected by the background noise and also the equipment, especially the power amplifier and the speakers. In addition, nonlinear distortions and time variability can also affect to the measurements, mostly due to possible saturation of the loudspeaker membrane and the nonlinearity of the power amplifier, and because of small movements of the measurement equipment during the measure. Because of these distortions, the measurement chain amplifier-loudspeaker-room-microphone-amplifier has to be described as a weakly nonlinear system. Then, a proper measurement technique for the case of HRTF should at least partially cover these particular issues.

The technique called periodic impulse excitation (PIE) uses a pulse train as the excitation signal but has poor signal-to-noise ratio (SNR) [166]. The dual-channel FFT uses Gaussian white noise as the excitation signal, but its high crest factor makes necessary many repetitions of the same measurement to obtain a good SNR [167]. Among the family of the pseudorandom sequences, the most popular is the maximum-length sequence (MLS), but the disadvantage of these techniques is that they are sensitive to nonlinear distortions [168].

The measurements with Exponential Sweep (ES) signals [130, 167] can overcome some of the previous problems. ES measurement allows to separate the linear and nonlinear parts of a weakly nonlinear system, provides with high SNR measurements, and has low sensitivity to transient noises that can disturb the measures. Besides, the processing of the measure is fast using FFT and it allows to define the measured frequency range.

The ES is a sweep in frequency from ω_1 to ω_2 with length T in seconds, as described by Equation 4.1

$$x(t) = \sin\left[\frac{\omega_1 T}{\ln\frac{\omega_2}{\omega_1}}\left(e^{\frac{t}{T}\ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right] \qquad t \in [0, T] \tag{4.1}$$

A weakly nonlinear system excited by an ES produces a response signal $y(t)$ which include higher order harmonics. Applying this response $y(t)$ in

Equation 4.2, the complete system identification is obtained in $h(t)$

$$h(t) = y(t) * x'(t) \tag{4.2}$$

where $x'(t)$ is the inverse of $x(t)$ with respect to the convolution, $x(t) * x'(t) = \delta(t)$. The inverse sweep compensates for the group delay and the magnitude spectrum of the excitation signal $x(t)$, and can be derived from Equation 4.3

$$X'(\omega) = \frac{X(-\omega)}{|X(\omega)^2|} \tag{4.3}$$

where $X(\omega)$ is the complex spectrum of the excitation signal $x(t)$ and $X'(\omega)$ is the complex spectrum of the inverse sweep.

The result $h(t)$ is a series of harmonic impulse responses of which the main one is the linear part of the measured system and the rest are the successive nonlinear harmonic distortions. These harmonic distortions can be cut out of the $h(t)$ signal obtained to maintain only the linear part of the system, proving that ES is a very robust acoustic measurement method against nonlinear distortion, small movements of the measurement equipment and occasional polluting noises during measurement. Besides, SNR can be improved by just increasing the duration T of the excitation signal $x(t)$, which is also necessary for measuring inside non-anechoic environments.

For our HRTF measurement purposes, it is therefore desirable that long ES be used to obtain clean, high SNR measurements. But long T sweeps of several measurement points around the person will make the measuring process quite long in time. Then the measurement becomes a time-consuming process, a tedious experience for the measured person and prone to errors due to the probable movements during the measuring. This is a typical problem described in most of the HRTF databases listed in section 4.1. That is the reason why Multiple Exponential Sweep Method (MESM) [154] was also implemented to be used in our measurement system.

MESM is based on ES and takes full advantage of the ES method but reduces the time of measurement when measuring many acoustical channels, in our case the acoustic path of many loudspeakers located in different spatial points to the ears of the measured person. In the MESM the exponential sweeps are reproduced overlapped in time, which results in a

shorter measurement duration of the multiple HRTFs. This method provides a correct identification of the linear part of weakly nonlinear systems when excited with correctly timed excitation sweeps, and an optimization of the measurement duration at a given SNR. The timing of the reproduction of the excitation signals is based on the interaction of two mechanisms, interleaving and overlapping.

The interleaving mechanism results from delaying the excitation signal of a second system in a way that its IR is placed between the IR and the second-order harmonics of the first system. The process can be generalized to interleave the linear responses of many systems. The requirement is to sufficiently separate the beginning of the first main IR from the end of the second non-linear harmonic, with sufficient space to accommodate the IRs of the rest of the systems. The separation is achieved by stretching the sweep to get (Equation 4.4)

$$\eta < \frac{\tau_2 - L_2}{L_1} + 1 \tag{4.4}$$

where η is the number of interleaved systems, L_1 and L_2 are the length of IRs of the first and second systems, and τ_2 is the exact distance from the start of the second harmonic relative to the start of the IR. In this case, the sweeps must have a minimum duration of T', as described by Equation 4.5

$$T' = [(\eta - 1)L_1 + L_2] \log_2 \left(\frac{\omega_2}{\omega_1}\right) \tag{4.5}$$

With the new sweep duration T' the beginning of the kth harmonic, relative to the beginning of the IR, is given by Equation 4.6

$$\tau_k^{'} = \frac{T'}{\ln(\omega_2/\omega_1)} \ln k \tag{4.6}$$

To apply the interleaving procedure, the excitation sweeps are played at the time $(i-1)L_1$, where i is the index of the measured HRTF, with $0 < i \leq \eta$.

An additional improvement can be achieved by introducing also the overlapping mechanism. The overlapping mechanism consist on delaying the reproduction of the excitation sweep of the second system time enough to avoid the overlap of the harmonics of the second system with the IR

of the previous one, otherwise the information about the linear part of the previous system is destroyed. Then, the reconstructed IR of a second system will be located in time between the nonlinear harmonics of the previous system and its own nonlinear harmonics. The minimum delay of the excitation sweep preventing superimposing the information is given by Equation 4.7

$$\Delta t_{ov} = \tau_k + L_1 \tag{4.7}$$

where k is the maximum number of harmonics above the noise floor of the system measured.

Both mechanism, interleaving and overlapping, can be combined to form the complete MESM. For a given η, the excitation time for the ith sweep of N systems is calculated by Equation 4.8

$$t_i = L_1(i-1) + \left\lfloor \frac{i-1}{\eta} \right\rfloor \tau'_k \tag{4.8}$$

where $0 < i \leq N$ and $\lfloor x \rfloor$ denotes the nearest lower integer of x. During the excitation of the systems the response signal $y(t)$ is recorded, and the extraction of a particular IR is done by windowing the signal $h(t)$. The shift of the window corresponds to the measured system and can be derived from Equation 4.8. The measurement duration of all the HRTFs with MESM finally depends on the parameters ω_1, ω_2, L_1, L_2 and k, which are practical values that should be derived from previous ES measurements under real conditions in the actual system.

In our measurement system, a problem was found when implementing MESM. An additional intermodulation distortion produced by the miniature microphones, and not covered by the MESM, made us unable to use the complete MESM to measure in our room. Just interleaving mechanism can be used to prevent any IR from being superimposed by those intermodulation distortions. Taking this into account, the practical conditions of our room, the loudspeakers and their interaction with the whole system, and the use of sweeps long enough to obtain a high SNR, the full measurement time is reduced from 7 minutes and 28 seconds using successive single ES (of 5 seconds), to 2 minutes and 28 seconds with interleaved MESM (88 channels measured with 10.11 seconds interleaved ES, ranging from 80Hz to 22kHz). The time reduction is important not only to make the measure more accessible. Since faster measurement on real subjects is less prone to

errors due to movement or discomfort of the measured person, the result will be a more accurate and reliable measurement.

4.2.4 Procedure protocol and measurement checking

Each person to be measured goes through the same procedure:

1 - Firstly, they are informed about all the process.

2 - Miniature microphones with earplugs are inserted into their ear canals.

3 - They are seated in the middle of the loudspeaker set-up, and with the aid of the laser pointers their position is adjusted to accurately be oriented with respect to the speaker array, as well as their height is regulated with the variable height seat.

4 - The actual measure is then performed with interleaved MESM during 2 minutes and 28 seconds.

5 - A rapid and automatic checking of the measurements is done. If any problem is detected the measurement can be immediately repeated.

6 - The measure is directly stored in a SOFA format file (see comments in the next Section 4.2.5).

The miniature microphones *Knowles FG 23329-P07* are placed into the ear canals inserted in the modified foam earplugs. Using a pair of tweezers, the earplugs with the microphones inside are completely introduced into the ear canals, with the microphone pointing outwards. The distance from the entrance of the ear canal to the microphones is intended to be identical between ears and few millimeters inside. When the foam earplug expands the microphone is fixed and the ear canal blocked (Figures 4.6). This results in blocked ear canal condition measurements [67], which include the full spatial information given to the ear. In addition, the variation between subjects of HRTFs measured at the blocked ear canal is smaller than at the same point with open ear canal, thus the choice of measure point at the blocked ear canal generate robust measurements [83].

The positioning of the subjects to be measured is simple thanks to a chair with wheels. Once the people are placed, they only have to remain still during the 2 minutes and 28 seconds that the actual measure lasts with interleaved MESM. The adjustment of height, displacement and orientation

Figure 4.6. Examples of pinnae with microphones in blocked ear canal condition

Figure 4.7. HRTF measurement process

of the person is easily done thanks to the reference lasers, with very precise results. Figure 4.7 shows the measurement process of a person.

Immediately after the measurement, a quick and automatic check routine of the measurements is carried out. This includes checking the correct position of the left-right channels, levels of recording, the correct channel order (loudspeaker or measurement point) and the position of the subject. The latter is checked by verifying the degree of symmetry of the ITD curves. If the person was correctly oriented, the ITD plot must be quite symmetrical with respect to the median plane, but if the subject was not correctly positioned or has moved during the measurement, the ITD plot will show its symmetry axis displaced from zero degrees or one lobe greater than the other. Figure 4.8 shows examples of symmetrical and asymmetrical ITD (of azimuth in milliseconds) indicating correct and incorrect position of the subject during the measurement of the HRTF.

The fast measurement and its checking process allows to repeat the measurement if necessary without much disturbance or annoyance to the measured person.

4.2.5 SOFA format

The Spatially Oriented Format for Acoustics (SOFA) is a format structure for storing spatial acoustic data defined by the standard AES69-2015 [21]. It provides a full description of the data structure, digital file and some possible predefined common types of measures referred as “conventions”. Among them, HRTF, HRIR and BRIR are considered.

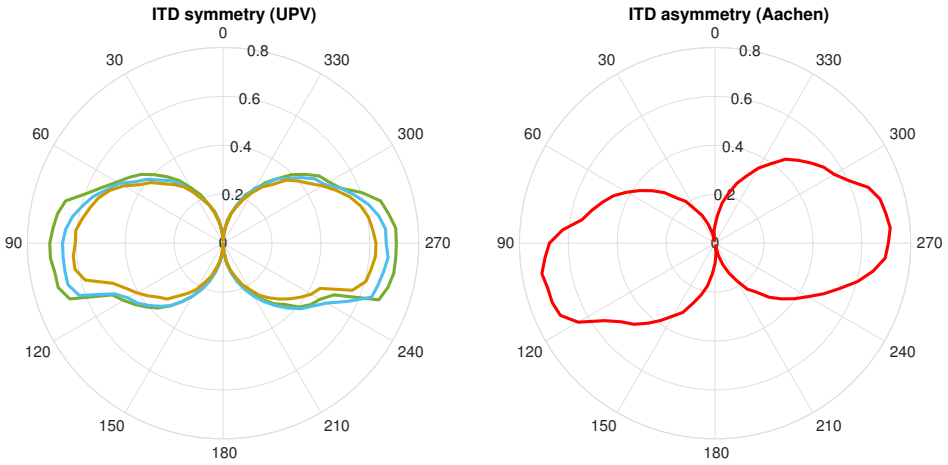


Figure 4.8. Symmetrical ITD indicates the correct position (left), asymmetrical ITD indicates the incorrect position (right) of the measured person

The measurements obtained in our system are stored in this format, based on an API for Matlab available online by the SOFA standardization project [169]. All the HRTF measurements from different spatial points of one subject are stored together as a detailed and individual collection in the same file. It also includes the geometrical coordinates of each spatial point related to each measure, as well as large amount of metadata describing measurements, equipment, post-processing, etc. Figure 4.9 shows the measured spatial points of one person, stored in a SOFA file.

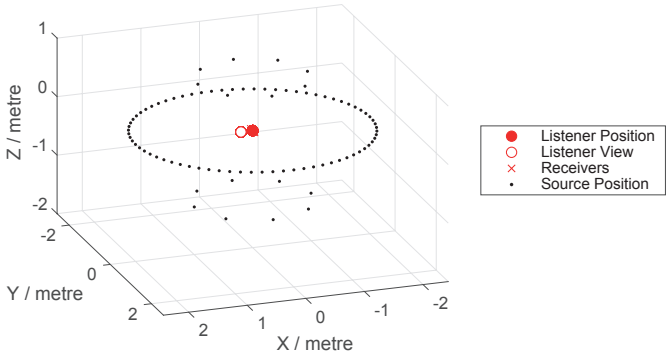


Figure 4.9. Spatial resolution of the raw measurements of one person, stored as a BRIR collection in a SOFA file

As a proposed standard, the SOFA format allows widespread sharing of HRTF information and reuse of previously developed software. Recent academic and commercial software for binaural reproduction and synthesis (many of them in the form of audio plugins) uses the SOFA format as a personalized HRTF input [107, 170, 171, 172, 173, 174]. This is indicative of the growing interest in individualization of HRTF and binaural sound.

4.3 Post-processing of the measurements

The raw measurements obtained directly with the measuring system can be improved and refined to get real HRTFs. By means of post-processing some basic corrections can be performed as well as some more complicated refinements. The raw measurements go through the next steps of processing:

1 - Level and speaker response correction using calibration measurements of the speakers.

2 - Almost complete removal of reflections by variable time-frequency windowing.

3 - Lowest-frequencies reconstruction.

The results of the post-processing are also stored in SOFA format, including a description of the post-processing stages in the metadata.

4.3.1 Level and loudspeaker response correction

Calibration measurements must be performed in order to equalize the level of all measurements and correct loudspeaker responses, which are included in the measured IRs. Figure 4.10 shows a picture of an *Earthworks M30* microphone taking reference measurements from the loudspeakers. A calibration factor that relates the level of all the loudspeakers is extracted from these measurements. The speaker response is also corrected on all speakers by an anechoic measurement of the response of that speaker model. To overcome any slight differences between the different actual loudspeakers of the same model and avoid other unwanted spectrum changes, 1/3 octave smoothed and minimum phase inverted responses are used for correction. The minimum phase filter ensures that the original phase and time rela-

tionships of the measured IR are not altered. Equation 4.9 describes the calibration and loudspeakers response correction,

$$Hc_i(\omega) = CF_i \frac{Hraw_i(\omega)}{Hspeak(\omega)} \tag{4.9}$$

where $Hraw_i(\omega)$ is the original raw binaural transfer function of the measurement obtained from each spatial point, CF_i is the calibration factor for each loudspeaker, $Hspeak(\omega)$ is the 1/3 octave smoothed and minimum phase response of the loudspeaker, $Hc_i(\omega)$ is the corrected and calibrated response, and i is the index for each loudspeaker and measured point, with $1 \leq i \leq 88$.

Figure 4.10. Calibration measurements with reference microphone

4.3.2 Removal of reflections by Frequency Dependant Windowing

The raw measurements recorded directly with the measurement system are actually Binaural Room Impulse Responses (BRIR). By post-processing them with variable time-frequency windows, the effects of the room are almost completely eliminated, obtaining *quasi* HRTFs. The process employed is a variation of the Frequency-Dependent Windowing (FDW) [175], also referred to as Frequency Dependant Truncation [176] for the specific case of removing reflections of impulse responses. Similar use of time-frequency windowing is found in previous works using two temporal windows [177, 178], one for high frequencies and other for low frequencies. In these works, reflections where partially removed and the rest softened, from HRTFs measured in semi-anechoic environments.

With truncation windows, the BRIR is cut in time keeping reflections out of the window. To eliminate all reflections the window must be very short, then the frequency resolution of such a window will not be sufficient to represent the whole spectrum, completely losing the low frequencies.

To overcome this, various time truncation windows with different time lengths can be used and take from the result only the proportional part of the spectrum where the frequency resolution is adequate. Different sizes time windows will provide different frequency resolutions according to the Equation 4.10

$$f_{res} = \frac{f_s}{N} = \frac{f_s}{f_s T_0} = \frac{1}{T_0} \tag{4.10}$$

where f_{res} is the frequency resolution (in Hz), f_s is the sampling frequency, N is the length of the temporal window in samples, and T_0 is the length of the temporal window in seconds (as can be noticed, the frequency resolution is independent of the sampling frequency).

The exact length of each temporal window will depend on the real measurement and the reflections recorded in the practical situation, because the main reflections will reach the measurement point in different instants of time. Due to our set-up and room dimensions, eight time windows are chosen with lengths 0.29, 0.58, 1.16, 2.33, 4.66, 9.33, 18.66, 37.33 ms (equivalent to 14, 28, 56, 112, 224, 448, 896, 1792 samples for 48 kHz sampling frequency), which have a frequency resolution of 3428.57, 1714.28, 857.14, 428.57, 214.28, 107.14, 53.57, 26.78 Hz respectively (Figure 4.11). A common shape for the truncation temporal windows is a modified rectangular window with ramps, e.g., halves of a Hanning window [179]. Such tempered design avoids temporal discontinuities while most of the IR waveform remains unaltered. The truncation window is also chosen to be asymmetric with respect to the maximum peak of the IR. As it extends further to positive times than negative times, the proportion of noise floor in the cropped IR is minimized. In this case, the truncation half-Hann window has always 32 samples before the maximum peak of the IR and the previously enumerated sizes after the peak. It has to be noted that the size of this previous-peak part of the window has very little influence on the frequency resolution, but avoids temporal alignment problems and guarantees that all IR content is included.

Figure 4.12 shows the useful spectral band of each temporal cropping window with an example of a real measured BRIR. Each of these spectral

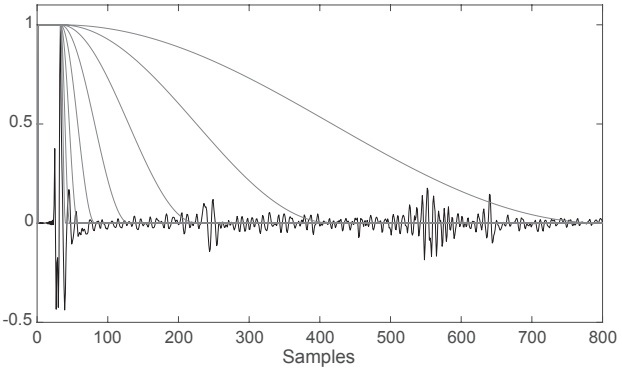


Figure 4.11. Temporal windows used in the Frequency Dependant Windowing to eliminate reflections from an example of one real measured BRIR

bands is gathered into a single spectral response that consequently does not include most of the reflections. The transition boundaries between adjacent bands are smoothed within a range of 1/6 octave to avoid any possible abrupt variation. In Figure 4.12 it can be seen graphically how up to the fifth window the removal of reflections is complete, however in low frequency from the band of about 500 Hz some reflections are still retained.

The complete result can be observed with an example in Figure 4.13, where the temporal and spectral representation of a measurement is presented before and after the Frequency Dependant Windowing processing, with and without reflections. Therefore, with this post-processing the measured BRIRs are transformed into *quasi* HRIRs.

A new method beyond Frequency Dependant Windowing is proposed in Section 4.4 to completely remove the remaining reflections in the band between 125 and 500 Hz.

4.3.3 Lowest-frequencies reconstruction

As described above, the measurement system is adjusted to use exponential sweeps starting from 80 Hz. This is done to avoid irregularities at the lower frequencies, where the loudspeaker is close to its reproduction lower limit and could produce unwanted distortions.

To compensate for the incomplete data of HRTFs at low frequencies measured with loudspeakers that lack energy in this frequency range, dif-

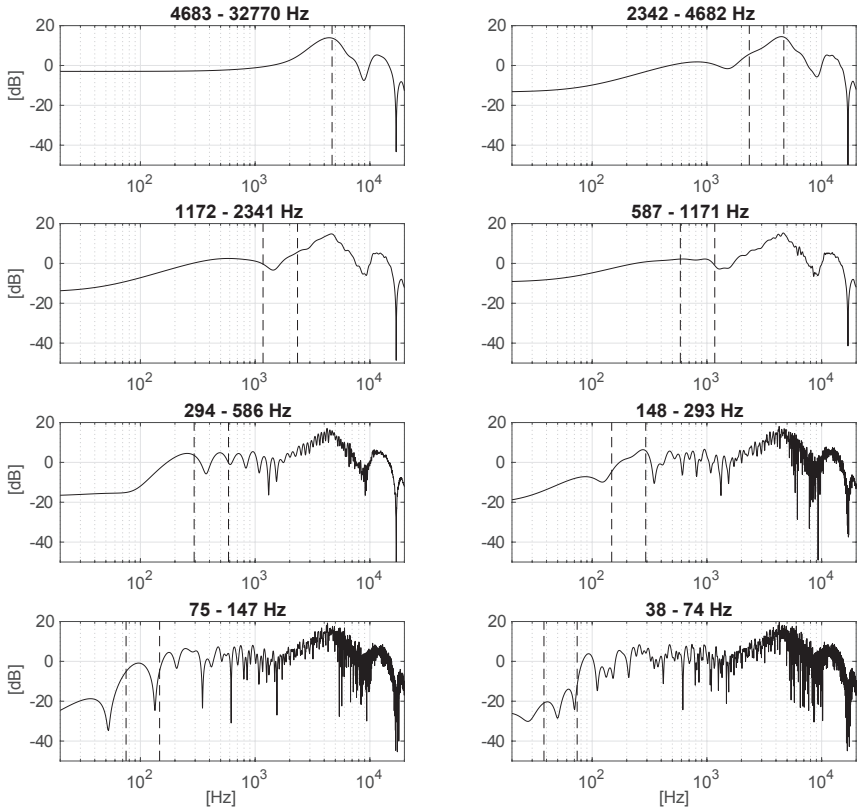


Figure 4.12. Useful spectral band corresponding to each temporal window used in the Frequency Dependant Windowing to eliminate reflections from an example of one real measured BRIR

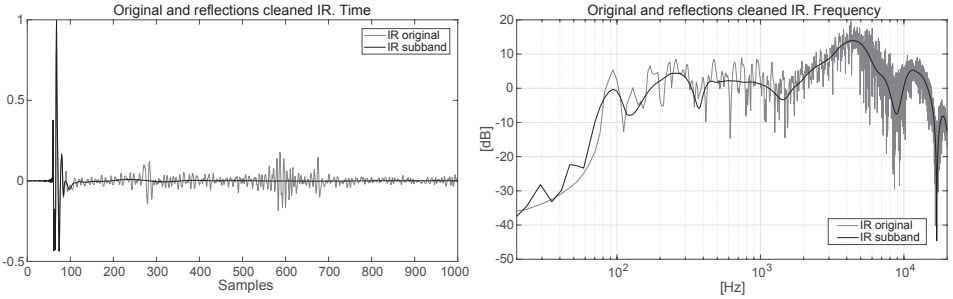


Figure 4.13. Temporal and spectral representation of one real measured BRIR and the *quasi* HRIR obtained

ferent strategies have been proposed. The missing information could be completed by means of geometric models of head and torso [180], with results from the boundary element method [181] or by cross-over filtering with an adequate low-frequency response [182]. All methods have in common that they lead to a monotonic function for the phase and an approximately linear magnitude response at low frequencies. This can be expected as the person does not present an obstacle to sound waves with large wavelengths. The approach by Xie [183], also employed in [184], takes in consideration this finding and have been chosen here. In this method the magnitude response is set to a constant value and the phase is extrapolated linearly.

The magnitude below a certain frequency is replaced by the mean value of a previous band and the phase is linearly extrapolated from the same frequency range. This method assumes that the data in this frequency range are reliable, but if the measurement is done under non-anechoic conditions, this band is likely to be affected by comb filtering due to reflections. Frequency Dependant Windows for cropping the signal without the floor or ceiling reflections (usually distances between 1.5 and 2 m in common rooms) do not have sufficient frequency resolution in this band, and longer cropping windows have these reflections present. Therefore, this method must be used with care here, due to our non-anechoic condition. According to our actual measurements, the spectrum below 100 Hz is reconstructed by the mean magnitude value and the extrapolation of the linear phase of the 100 - 300 Hz band.

The reconstruction of this lowest frequency band can be done more reliably for HRTFs measured under non-anechoic conditions if the method proposed in the next Section 4.4 is previously used to eliminate the reflections below 500 Hz. Then, the band directly above 100 Hz will be cleaned of comb filtering effect due to reflections, and therefore less conservative cut-off frequency values can be used for the reconstruction of the lowest frequencies.

The correction of the low-frequency range of the HRTFs bears the advantage that the implausibly high group delays at low frequencies are lowered, and the HRTFs can be truncated which saves storage space and makes their usage computationally more efficient. On the other hand, even though very low frequencies have practically no contribution to spatial localization there are other perceptual benefits [1, 182]. The reconstruction of these lowest frequencies makes it possible to preserve the timbre charac-

teristics of real sound sources in order to recreate natural-like and realistic virtual sounds.

4.4 Proposed method to remove low-frequency reflections by Plane Wave Decomposition

The transformation of the measured BRIRs into HRIRs without reflections does not present a simple solution. A combination of different signal processing techniques and the application of the method proposed here can be used to address the elimination of all recorded reflections together with the measured HRTF.

Previously we have seen how reflections at high frequencies can be eliminated with Frequency Dependant Windowing or simply IR cropping (see Section 4.3.2), down to approximately 500 Hz in the constructed room. It has also been commented on the flat reconstruction of the lowest frequencies while preserving the ITD, below 100 Hz in the case of our room (see Section 4.3.3). There remains therefore a gap in the spectrum to be completely resolved, the band between 100-500 Hz.

Important information is found in this spectral band, such as torso and shoulder reflections, which are used as perceptual cues for localization, especially for elevated directions [24, 88, 185]. If measurements are done in a non-anechoic room, this band is usually contaminated with room reflections, since the dimensions and proportions of a common room (distances from the ears to the walls, floor and ceiling between 0.5 and 2 meters) coincide with the generation of main reflections that affect this band (comb filtering effects between 680 and 170 Hz). It is also interesting to note that if the information in this band between 100-500 Hz is correct and reliable, the spectral reconstruction below 100Hz can be based on these frequency values immediately above and will be more accurate and real.

In this section, a method is proposed to eliminate reflections below 500 Hz of the measured BRIRs, by means of a proper compensation of the most significant reflections. It will be firstly described as a general method and secondly validated with an example and a real measurement from our constructed room. The proposed method employs measurements with a microphone array and uses Plane Wave Decomposition (PWD) to extract the reflections from measurements of the room.

4.4.1 Problem formulation and background

To set the basis for describing the generalized method, the basic notation will be defined and the description of the problem will be presented. In addition, the theoretical background of PWD not previously commented will be presented in this Section.

Signal model

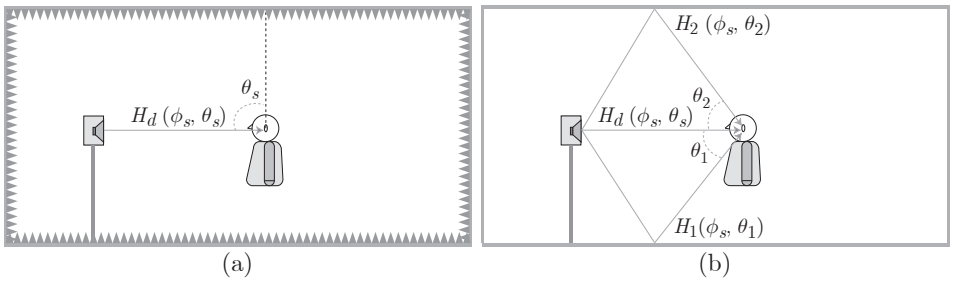


Figure 4.14. Anechoic and non-anechoic HRTF measurement (a) Anechoic measurement with no reflections (b) Non-anechoic measurement with relevant early reflections

Considering an ideal anechoic environment, as in Figure 4.14 (a), the sound pressure measured at one of the ears resulting from a sound source signal, $s(t)$, coming from the spatial direction defined by the azimuth and elevation angles (ϕ_s, θ_s), can be expressed as:

$$y_d(t) = s(t) * h_d(t; \phi_s, \theta_s) * h_{hrir}(t; \phi_s, \theta_s) \tag{4.11}$$

where $*$ denotes convolution and $h_d(t; \phi_s, \theta_s)$ is the direct-path acoustic channel. The term $h_{hrir}(t; \phi_s, \theta_s)$ represents the head-related impulse response (HRIR) corresponding to direction (ϕ_s, θ_s).

In non-anechoic conditions, as represented in Figure 4.14 (b), multiple reflections coming from the different surfaces occur inside the room and the measured acoustic pressure now contains a non-desired component as:

$$y(t) = y_d(t) + y_r(t) \tag{4.12}$$

with

$$y_r(t) = s(t) * \left(\sum_{m=1}^{M} h_m(t; \phi_m, \theta_m) * h_{hrir}(t; \phi_m, \theta_m) \right) \tag{4.13}$$

where M is the total number of significant reflections and $h_m(t;\phi_m,\theta_m)$ denotes the acoustic impulse response for the acoustic path of the m-th reflection coming from direction (ϕ_m,θ_m). As a result, the measured HRIR in a non-anechoic measurement setup will include the anechoic HRIR plus the contribution of the non-desired acoustic path components:

$$\begin{aligned}\tilde{h}_{hrir}(t;\phi_s,\theta_s) &= h_d(t;\phi_s,\theta_s) * h_{hrir}(t;\phi_s,\theta_s) \\ &+ \sum_{m=1}^{M} h_m(t;\phi_m,\theta_m) * h_{hrir}(t;\phi_m,\theta_m)\end{aligned} \tag{4.14}$$

In the frequency domain, the measured HRTF translates to the addition of the anechoic HRTF plus multiple reflections coming from different directions, filtered by their corresponding traveled acoustic channel, i.e.

$$\begin{aligned}\tilde{H}_{hrtf}(\omega;\phi_s,\theta_s) &= H_d(\omega;\phi_s,\theta_s)H_{hrtf}(\omega;\phi_s,\theta_s) \\ &+ \sum_{m=1}^{M} H_m(\omega;\phi_m,\theta_m)H_{hrtf}(\omega;\phi_m,\theta_m)\end{aligned} \tag{4.15}$$

where $H_m(\omega;\phi_m,\theta_m)$ represents the Fourier transform of the reflection acoustic path $h_m(t;\phi_m,\theta_m)$ and $H_{hrtf}(\omega;\phi,\theta)$ is the Fourier transform of $h_{hrir}(t;\phi,\theta)$.

BRIR cropping limitation

Consider a non-anechoic measurement scenario, as in Figure 4.14 (b). At the listening position, the arriving signal is a mixture of the direct sound and two typical significant room reflections. In this example they are mainly originated on the floor and the ceiling. However, the lateral and opposite walls also produce reflections, but they are not considered here for the sake of simplicity. A typical measured impulse response in such type of scenario is shown in Figure 4.15, corresponding to a medium-size room with a standard ceiling height of 2.6 meters. For a source to listener distance of 1 meter and a source to ceiling distance of 2 meters, the floor reflection arrives at 4.5 ms, while the ceiling reflection arrives at 12 ms approximately. The direct signal and the two main reflections can be identified arriving after the main signal.

Cropping the measured response before the arrival of the first echo would preserve the direct path and eliminate subsequent reflections, but

such a short window implies a significant loss in frequency resolution, which at lower bands can be considerably critical. However, for the high-frequency part of HRTF this procedure may be sufficient. As commented in Section 4.3.2, the frequency resolution of a time window is given by Equation 4.10. Consider the response of Figure 4.15 as an example. The first reflection arrives approximately 4.5 ms after the direct signal, so if the impulse response is cropped to a duration of 4 ms, the resulting frequency resolution is just 250 Hz, which can be enough for high frequencies but will ruin the information contained in the low frequency response.

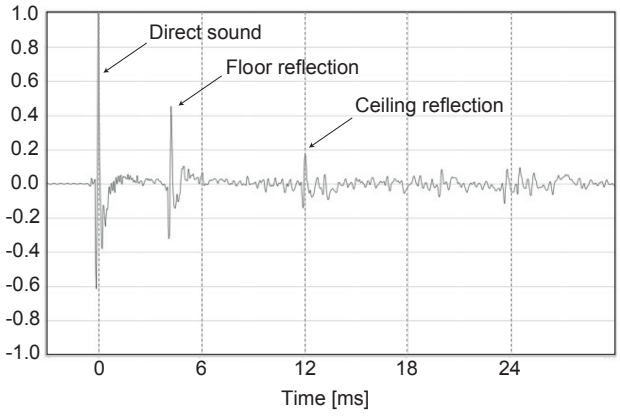


Figure 4.15. Typical impulse response in non-anechoic measurement room with large floor and ceiling reflections

Plane Wave Decomposition background

A given sound field can be decomposed into its plane wave components according to the principle of superposition. Assuming a continuous pressure distribution $P(\omega;\phi,\theta,r_0)$ on an open sphere, and its corresponding spatial Fourier coefficients, $\mathring{P}_{nm}(\omega;r_0)$, the Plane Wave Decomposition (PWD) would return the plane wave components D for a specific spatial decomposition direction and angular frequency $(\omega;\phi_\mathrm{d},\theta_\mathrm{d})$:

$$D(\omega;\phi_\mathrm{d},\theta_\mathrm{d},) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \frac{1}{i^n j_n\left(\frac{\omega}{c}r_0\right)} \mathring{P}_{nm}(\omega;r_0) Y_n^m(\phi_\mathrm{d},\theta_\mathrm{d}) \qquad (4.16)$$

The angular functions $Y_n^m(\phi, \theta)$ are referred to as spherical harmonics:

$$Y_n^m(\phi, \theta) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi} \tag{4.17}$$

where $P_n^m(\cos\theta)$ are Legendre functions of the first kind of order n and mode m. The radial part in Equation 4.16 depending on ω and r_0 is written in terms of j_n, which is the nth-order spherical Bessel function of the first kind.

In practice, the pressure distribution on the sphere is sampled at a limited amount of discrete spatial sampling nodes. As a consequence, discrete sampling schemes resolve spherical harmonics up to a maximum order N. While ideal PWD for $N \to \infty$ corresponds to a spatial Dirac pulse, order truncation results in a widened main lobe and additional side-lobes, decreasing substantially spatial resolution. In fact, practical signal processing applications may limit the maximum order N, so that $N < \frac{\omega}{c} r_0$ to reduce aliasing contributions arising from discrete spatial sampling. For M discrete microphone positions defined by a quadrature grid on the sphere, the spatial Fourier coefficients in the spherical wave domain are given by the summation:

$$\mathring{P}_{nm}(\omega; r_0) = \sum_{j=1}^{M} \beta_j P(\omega; \phi_j, \theta_j, r_0) Y_n^m(\theta_j, \phi_j)^* \tag{4.18}$$

where β_j are weighting factors that account for the selected spatial grid.

Radial filters compensate for the radial portion of the Helmholtz equation, scaling the amplification gain of spherical harmonic modes. Radial filters depend on the sphere configuration, which describes whether sensor nodes on the sphere are in free field or mounted on a rigid body. For an open measurement sphere with pressure transducers, as the radial part in Equation 4.16, the radial filters are directly given by

$$d_n\left(\frac{\omega}{c} r_0\right) = \left[4\pi i^n j_n\left(\frac{\omega}{c} r_0\right)\right]^{-1} \tag{4.19}$$

For pressure transducers mounted on a rigid sphere, radial filters take the following form:

$$d_n\left(\frac{\omega}{c} r_0\right) = \left[4\pi i^n \left(j_n\left(\frac{\omega}{c} r_0\right) - \frac{j_n'\left(\frac{\omega}{c} r_0\right)}{h_n'^{(2)}\left(\frac{\omega}{c} r_0\right)} h_n^{(2)}\left(\frac{\omega}{c} r_0\right)\right)\right]^{-1} \tag{4.20}$$

where $h_n^{(2)}$ denotes the spherical Hankel function of the second kind.

In a practical scenario, radial filters are not assumed to cover the entire frequency range. The modal amplification demanded by such filters is too high at low $\left(\frac{\omega}{c}r_0\right)$, leading to unstable array responses at lower frequencies due to noise amplification. Thus, the amplification of higher modes is limited in practice to a reasonable value, although the limiting operation results in a loss of spatial resolution at lower frequencies. Taking into account both the limited order imposed by discrete spatial sampling and the effect of radial filters, the response signal for a frequency/direction $(\omega; \phi_\mathrm{d}, \theta_\mathrm{d})$ is obtained by

$$D(\omega; \phi_\mathrm{d}, \theta_\mathrm{d}) = 4\pi \sum_{n=0}^{N} \sum_{m=-n}^{n} d_n \left(\frac{\omega}{c} r_0\right) \mathring{P}_{nm}(\omega; r_0) Y_n^m(\phi_\mathrm{d}, \theta_\mathrm{d}) \tag{4.21}$$

Despite the use of non-critical radial filters have an impact on the effective operational bandwidth of spherical arrays and therefore ideal constant directivity PWD response is distorted for very low frequencies, such limitation is not really relevant here, as can be seen with the validation of the method with real measurements (Section 4.4.3).

4.4.2 General description of the proposed method

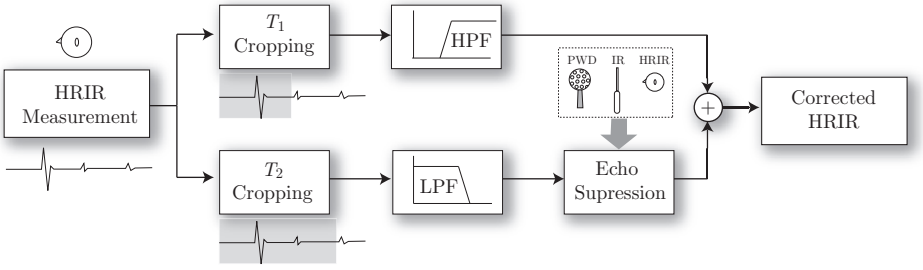


Figure 4.16. Block diagram of the proposed method

As already discussed, FDW or simple cropping can be applied to extract the information related to the high-frequency range of the desired response. However, additional processing is necessary to preserve low-frequency information while minimizing the effect of room reflections.

The general processing scheme of the proposed method is shown in

Figure 4.16. The bottom branch is aimed at processing the low-frequency part of the input BRIR signal. The effect of room reflections is cancelled based on the directional information extracted by a spherical microphone array and sound field analysis techniques. The top branch extracts the high-frequency information of the desired signal. The input BRIR is cropped before the arrival of the first reflection to keep only the direct path, leaving only the high-frequency information.

An important aspect of the proposed approach is the lowest frequency limit to be achieved, which is set at 100 Hz. For frequencies below this limit, the magnitude of the HRTF is practically flat and can be reconstructed, as described in Section 4.3.3. A frequency limit of 100 Hz implies a time interval of 10 ms after the arrival of the direct sound (or 480 samples for a sampling frequency $F_s = 48$ kHz). Thus, reflections occurring within such time window are the ones that must be properly handled. Note that the proposed technique effectively extends the frequency range of non-anechoic HRTF measurements to cover an important part of the low-frequency response, between 100 and 500 Hz or even 1 kHz depending on the reflections time of arrival.

The proposed method involves the use of different sound capture techniques and loudspeaker set-ups. For the sake of simplicity, let us consider only the removal of the floor reflection. The suppression of the ceiling reflection can be addressed by extending the same procedure in a similar manner. Nonetheless, reflections in the ceiling can be avoided more easily by means of suitable acoustic panels, as they are not stepped on.

The following steps describe the general procedure of the proposed method to remove the reflections of the low-frequency part of the BRIR.

Step 1: HRIR cropping

Let us assume the non-anechoic HRTF measurement set-up depicted in Figure 4.17 (a), corresponding to a source direction (ϕ_s, θ_s). As an example, the frontal direction $\phi_s = 0$ and $\theta_s = 0$ is considered, obtaining the measurement $\tilde{h}_{hrir}(t; 0, 0)$, which follows the model of Equation 4.15. The first and most relevant reflection affecting the measurement comes from the floor surface, with direction (ϕ_s, θ_1). There are many other contributions arriving to the listener, but since they arrive considerably later, they can be discarded by cropping. The measured response can be cropped before the arrival of this first room reflection by selecting a cropping time T_1.

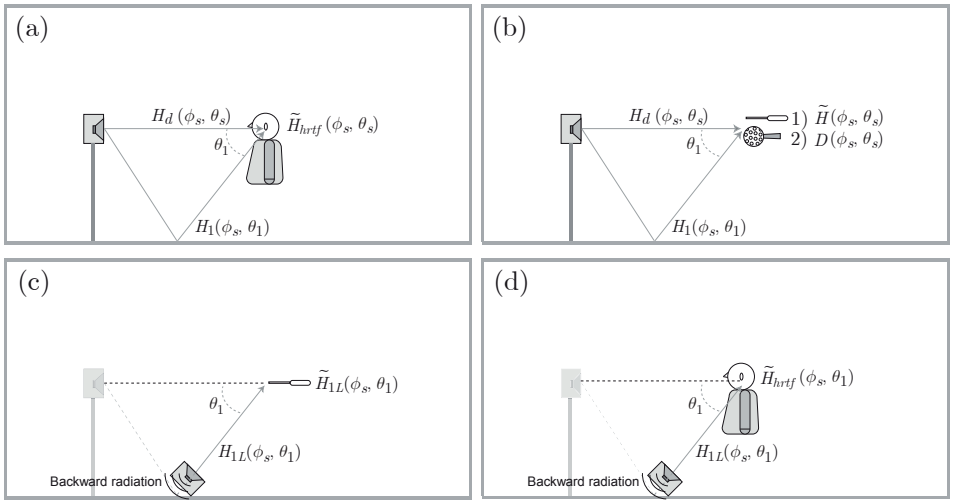


Figure 4.17. Measurements for the correction procedure. (a) Non-anechoic HRTF measurement (b) Impulse response measurement with ominidirectional microphone and acoustic channel estimation based on spherical array processing (c) Echo path impulse response measurement (d) Echo path HRTF measurement

However, depending on the type of cropping window used, such short time results in a poor frequency resolution, e. g. between 250 Hz and 500 Hz for $T_1 = 4$ ms, following the example of Figure 4.15. Thus, the HRTF could only be measured reliably above those frequencies, but no information will be contained in the range between 150 Hz and 500 Hz (depending on the chosen size of the windows).

To increase the frequency resolution a longer window with cropping time T_2 can be used. As an example based on Figure 4.15, a cropping time $T_2 = 11$ ms would increase the frequency resolution to 90 or 180 Hz, extending the frequency range to a factor close to 1.5 octaves with respect to T_1. The cropping time T_2 is selected slightly before the arrival time of the second relevant reflection (see Step 2). Thus, the cropped HRIR can be written as

$$\tilde{h}_{hrir}^{T_2}(t;\phi_s,\theta_s) = h_d(t;\phi_s,\theta_s) * h_{hrir}(t;\phi_s,\theta_s) + h_1(t;\phi_s,\theta_1) * h_{hrir}(t;\phi_s,\theta_1) \tag{4.22}$$

leading, in the frequency domain, to the HRTF

$$\tilde{H}^{T_2}_{hrtf}(\omega;\phi_s,\theta_s) = H_d(\omega;\phi_s,\theta_s)\cdot H_{hrtf}(\omega;\phi_s,\theta_s) + H_1(\omega;\phi_s,\theta_1)\cdot H_{hrtf}(\omega;\phi_s,\theta_1) \tag{4.23}$$

An even more accurate high-frequency spectrum can be obtained if FDW with several windows is applied beyond simple cropping (see Section 4.3.2), but here this process is reduced to two temporal cropping windows for the sake of simplicity.

To estimate the actual frontal HRTF with increased frequency resolution, the effect from the reflection must be suppressed. This implies estimating not only the acoustic path corresponding to such reflection, but also the HRIR for its direction. The following steps address such estimation process.

Step 2: Acoustic channel estimation

The second step involves two measurements, as depicted in Figure 4.17 (b).

Impulse response measurement: First, the listener is substituted by a flat-response omnidirectional microphone to measure the impulse response between the source and the listener's position, $\tilde{h}(t;\phi_s,\theta_s)$. By inspecting this response, the cropping times T_1 and T_2 can be determined just before the first and second relevant reflections, respectively. Cropping the impulse response at time T_2 results in the following model for the measured signal

$$\tilde{h}^{T_2}(t;\phi_s,\theta_s) = h_d(t;\phi_s,\theta_s) + h_1(t;\phi_s,\theta_1) \tag{4.24}$$

or, in the frequency domain

$$\tilde{H}^{T_2}(\omega;\phi_s,\theta_s) = H_d(\omega;\phi_s,\theta_s) + H_1(\omega;\phi_s,\theta_1) \tag{4.25}$$

Spherical array measurement and PWD: Second, to separate the contribution of the two described acoustic paths, an M-channel impulse response is recorded with an spherical microphone array positioned at the location of the previous omnidirectional microphone (Figure 4.18). Such multi-channel impulse response is denoted as $\tilde{P}^{T_2}(\omega;\phi_j,\theta_j)$, $j=1,\ldots,M$, where the T_2 superscript indicates that temporal cropping is also applied to avoid reflections. The cropped array response is analyzed by means of PWD, using Equations 4.18 and 4.21, to obtain the desired responses at

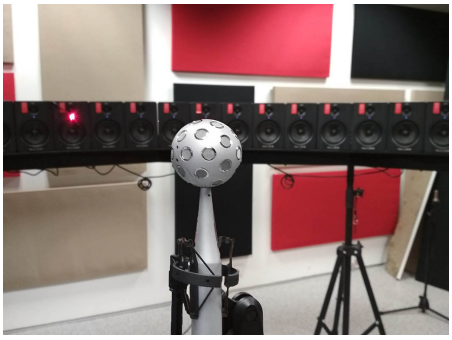


Figure 4.18. Measurement with spherical microphone

directions (ϕ_s, θ_s) and (ϕ_s, θ_1). The analysis results in the two plane-wave components $D(\omega; \phi_s, \theta_s)$ and $D(\omega; \phi_s, \theta_1)$.

It is important to note that, while the signals obtained by PWD are related to the actual acoustic channels, they may contain some amplitude differences that are modeled here by an unknown filter $Q(\omega)$ as follows:

$$D(\omega; \phi_s, \theta_s) = Q(\omega) H_d(\omega; \phi_s, \theta_s) \tag{4.26}$$

$$D(\omega; \phi_s, \theta_1) = Q(\omega) H_1(\omega; \phi_s, \theta_1) \tag{4.27}$$

The filter $Q(\omega)$ takes into account both the frequency response effect of the PWD and that of the microphones making up the spherical array. By substituting Equations 4.26 and 4.27 into Equation 4.25, and omitting the variables ω and ϕ_s for notation simplicity, the measured response can be written as

$$\tilde{H}^{T_2}(\theta_s) \approx \frac{1}{Q} \cdot (D(\theta_s) + D(\theta_1)) \tag{4.28}$$

so that the filter $Q(\omega)$ is estimated as

$$\hat{Q} = \frac{D(\theta_s) + D(\theta_1)}{\tilde{H}^{T_2}(\theta_s)} \tag{4.29}$$

The acoustic paths can therefore be estimated as

$$\hat{H}_d(\theta_s) = \frac{1}{\hat{Q}} D(\theta_s) = \frac{\tilde{H}^{T_2}(\theta_s)}{D(\theta_s) + D(\theta_1)} D(\theta_s) \tag{4.30}$$

$$\hat{H}_1(\theta_1) = \frac{1}{\hat{Q}} D(\theta_1) = \frac{\tilde{H}^{T_2}(\theta_s)}{D(\theta_s) + D(\theta_1)} D(\theta_1) \tag{4.31}$$

At this point, an estimate of the main acoustic paths have been obtained, but the effect of the HRTF for the direction of the main reflection is still completely unknown. The third step addresses such issue.

Step 3: Reflection path measurements and suppression

As illustrated in Figure 4.17 (c), the acoustic transfer function is again measured with the omnidirectional microphone, but this time using a loudspeaker placed on the floor and oriented towards the direction of the reflection θ_1. It should be mentioned that the measured acoustic channel, denoted as $\tilde{h}_{1L}(t; \phi_s, \theta_1)$, includes the bass enhancement effect resulting from the floor placement of the loudspeaker. Indeed, due to the speaker boundary interference response [186], the loudspeaker behavior in Figure 4.17 (b) for frontal radiation is different from the one observed when the loudspeaker is placed on the floor. Note that, due to the orientation of the loudspeaker, the reflections will arrive later in time than T_2. Then, as with $\tilde{h}^{T_2}(t; \phi_s, \theta_s)$, this new response is also cropped with the same window length to discard reflections, resulting in $\tilde{h}_{1L}^{T_2}(t; \phi_s, \theta_1)$.

Subsequently, the microphone is substituted with the subject, measuring and cropping the HRIR from that direction (see Figure 4.17 (d)). The resulting HRTF can be written as:

$$\tilde{H}_{hrtf}^{T_2}(\omega; \phi_s, \theta_1) = H_{hrtf}(\omega; \phi_s, \theta_1) \cdot H_{1L}(\omega; \phi_s, \theta_1) \tag{4.32}$$

where $H_{1L}(\theta_1)$ is the frequency-domain equivalent of $h_{1L}(\theta_1)$, already known from the onmidirectional microphone measurement. Thus, the HRTF for the echo direction can be estimated as:

$$\hat{H}_{hrtf}(\omega; \phi_s, \theta_1) = \frac{\tilde{H}_{hrtf}^{T_2}(\omega; \phi_s, \theta_1)}{\tilde{H}_{1L}^{T_2}(\omega; \phi_s, \theta_1)} \tag{4.33}$$

Finally, according to Equation 4.25, an estimate of the echo-free HRTF for the desired source direction can be obtained as:

$$\hat{H}_{hrtf}(\theta_s) = \frac{\tilde{H}_{hrtf}^{T_2}(\theta_s) - \hat{H}_1(\theta_1)\hat{H}_{hrtf}(\theta_1)}{\hat{H}_d(\theta_s)} \tag{4.34}$$

where the azimuth angle ϕ_s and frequency ω have been again omitted for notation simplicity.

4.4.3 Validation of the method with real measurements

Real experiments that have been carried out to validate the proposed reflection compensation method. To analyze the performance, measured and corrected HRTFs have been compared to the corresponding HRTFs obtained in an anechoic chamber using the same measuring equipment.

The actual measurements of BRIR, and the acoustic paths recorded with the spherical matrix and an omnidirectional microphone were taken in the constructed room described in Section 4.2, and the verification measurements were taken in an anechoic chamber. The Earthworks M30 [187] microphone was used to measure the required omnidirectional impulse responses at the different stages of the method. The spherical array recordings were captured by an em32 Eigenmike from mh acoustics [188], composed by 32 individual electret capsules inserted on a rigid sphere of radius 4.2 cm, following the shape of a pentakis dodecahedron with the transducers placed at its vertices. The BRIR measurements were acquired with the Brüel & Kjær Head And Torso Simulator (HATS) 4100 [189]. The direction selected to carry out the experiment was the frontal one, ($\phi_s = 0°, \theta_s = 0°$). All the recordings considered a sampling frequency of $F_s = 48$ kHz.

The PWD processing was performed by means of the open-source SOFiA MATLAB library [190]. The Eigenmike array allows a spherical harmonic decomposition up to the order $N = 4$. The reflection from the back and lateral walls are not considered now in the low-frequency processing, since they are highly absorbed by acoustic materials placed on these surfaces. The reflection on the ceiling can be handled following the same procedure as the one on the floor, but using complementary measurements from the elevated loudspeaker array.

The processing dealing with echo suppression will consider the frequency range between 100 and 1000 Hz. As already described, frequencies above 1000 Hz can be reliably measured by cropping the measured HRIRs. On the other hand, the magnitude response of HRTFs below 100 Hz is completely flat, and its phase can be easily reconstructed by ensuring that the ITD is properly preserved at low frequencies.

HRIR and impulse response cropping

As a first step, both the BRIR to be compensated and the room impulse response, are measured with the dummy-head device and the flat-response

omnidirectional microphone, respectively. Thus, the signals $\tilde{h}_{hrir}(t; 0°, 0°)$ and $\hat{h}(t; 0°, 0°)$ are acquired. Since all the measurements correspond to the same azimuth $\phi_s = 0°$, the azimuth angle will be omitted for simplicity in what follows. Given the specific geometry of the measurement set-up, the reflection from the floor arrives approximately at 4.4 ms (211 samples after the direct sound, at $F_s = 48$ kHz), leading to a low-frequency limit between $1/0.0044 = 227$ Hz to 454 Hz depending on the windowing employed. Following the approach described in the Section 4.4.2, T_1 is fixed to 200 samples in order to avoid this reflection. The acquired impulse response is shown in Figure 4.19, where such main reflection can be clearly identified, as well as the second one. Therefore, T_2 is fixed just before this second reflection at 420 samples. Note that the longer window T_2 contains the effect of the first reflection that must be suppressed with the proposed compensation method. The windows used to crop the measured temporal responses are also depicted in Figure 4.19, which have the shape of a soft decay rectangular window having a Hanning profile.

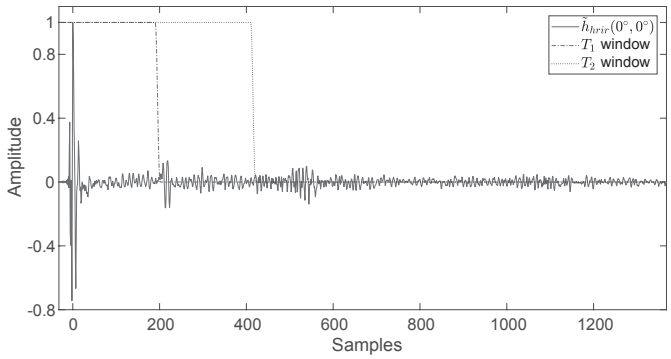


Figure 4.19. Measured impulse response from the frontal direction, $\tilde{h}(t; 0°, 0°)$, and selected cropping windows

Acoustic channel estimation

In the next step, the multichannel impulse response at the same location is measured with the Eigenmike spherical microphone array, extracting the signals $D(\omega, 0°)$ and $D(\omega, -45°)$ by means of PWD. Both are represented in Figure 4.20 for the frequency range of interest. The result after performing the equalization operation of Equations 4.30 and 4.31 are also represented as $\hat{H}_d(0°)$ and $\hat{H}_1(-45°)$. It can be seen that the sum of both corrected answers $(\hat{H}_d(0°) + \hat{H}_1(-45°))$ perfectly matches the original fre-

quency response measured with the omnidirectional microphone, $\tilde{H}^{T_2}(0°)$, which includes the combination of the two paths. The need for such equalization is also demonstrated by showing the addition of the two PWD signals ($D(0°) + D(-45°)$), which results in a magnitude difference between 10 and 12 dB due to factors such as the frequency response of the array capsules and other effects derived from the PWD processing.

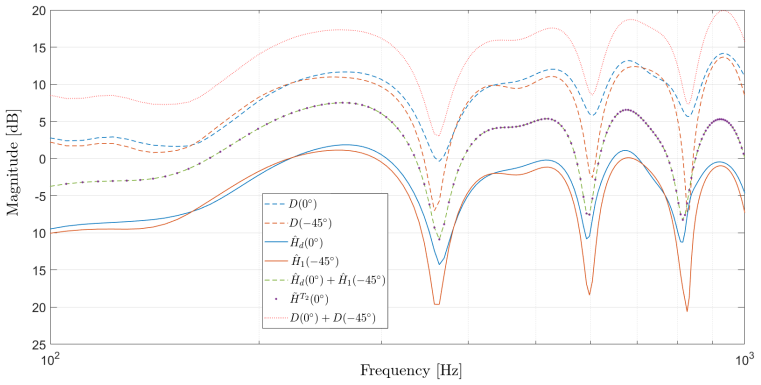


Figure 4.20. PWD Components, $D(0°)$ and $D(-45°)$, and estimated acoustic channels, $\hat{H}_d(0°)$ and $\hat{H}_1(-45°)$. The addition of the two estimated acoustic paths matches the measured omnidirectional response $\tilde{H}^{T_2}(0°)$

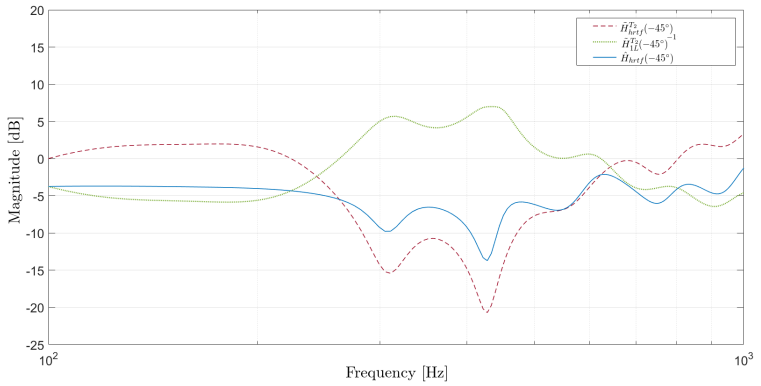


Figure 4.21. Measured ($\tilde{H}_{hrtf}(-45°)$) and compensated ($\hat{H}_{hrtf}(-45°)$) HRTFs from the echo direction, together with the inverse of the measured transfer function $\tilde{H}_{1L}(-45°)$

Reflection path measurements and compensation

For the next measurement step, the HRTF for the reflection direction must be properly estimated. To this end, a loudspeaker having the same azimuth angle, $\phi_s = 0°$, but from the lower loudspeaker array, is selected. This loudspeaker has the elevation angle of the reflection ($\theta_s = -45°$) and points towards the listener position. The response is measured both with the flat-response microphone and with the dummy-head device, resulting in the signals $\tilde{h}_{hrir}(t, -45°)$ and $\tilde{h}_{1L}(t, -45°)$, respectively. Due to the back radiation of the loudspeaker on the floor, the original room transfer function presents a relevant boost in the range 100−250 Hz. The compensation is performed by applying Equation 4.33. To this end, a regularized inverse filter [191] with $\beta = 0.05$ is considered. Figure 4.21 shows the measured $\tilde{H}_{hrtf}(-45°)$, the inverse filter $\tilde{H}_{1L}(-45°)^{-1}$ and the compensated version $\hat{H}_{hrtf}(-45°)$, in the frequency domain for the frequency range of interest. As observed, the result is the cancellation of the low-frequency boost at the frequencies of interest, with a reduction in magnitude close to 5 dB.

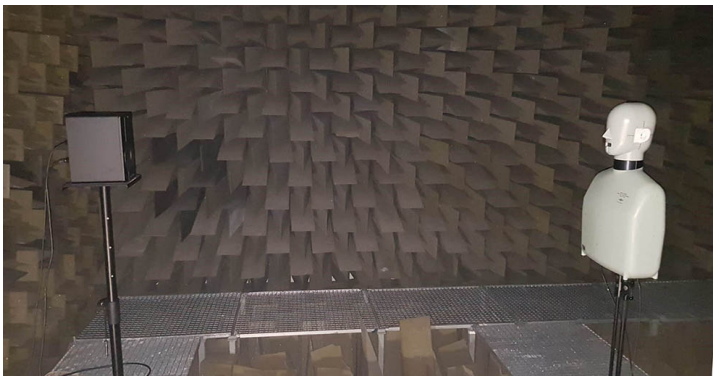


Figure 4.22. Measurements in the anechoic chamber

Final HRTF estimation and discussion

The last step corresponds to the final estimation of the reflection-free HRTF for the frontal direction, as given by Equation 4.34. Anechoic chamber measurements were acquired under the replica of the measurement conditions of the non-anechoic configuration, as shown in Figure 4.22. Thus, the same distance between the source and the listener and the same measurement equipment was considered to avoid changes due to different loudspeaker

responses. The compensated low-frequency response obtained from the T_2 window is crossed-over at 1 kHz with the high-frequency response obtained by the T_1 cropping. Following the experimental set-up described above, the result for the HRTF obtained after processing the high and low frequencies separately and adding them back together is shown in Figure 4.23. The corresponding anechoic measurement is shown together to evaluate the temporal error in the equivalent HRIR.

In order to properly interpret the results, different frequency responses have also been represented in Figure 4.24. Firstly, the HRTF measured in the anechoic chamber, $H_{hrtf}(0°, 0°)$, is shown as reference and target. On the other hand, the figure also shows the non-compensated HRTFs obtained in non-anechoic conditions using the cropping windows T_1 and T_2, i.e. $\tilde{H}^{T_1}_{hrtf}(0°, 0°)$ and $\tilde{H}^{T_2}_{hrtf}(0°, 0°)$. Finally, the HRTF estimated with the proposed compensation method, $\hat{H}_{hrtf}(0°, 0°)$, is represented.

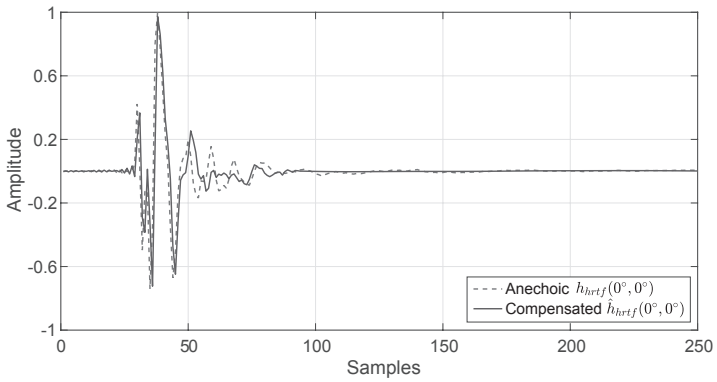


Figure 4.23. Comparison between the compensated HRIR and the corresponding measurement in anechoic chamber

The responses corresponding to all the non-anechoic curves overlap in the high-frequency range (above 1 kHz), as expected from the the sub-band processing scheme. In the low-frequency range, the two non-compensated responses show undesired effects. The response obtained from the longer window T_2 presents a remarkable comb filtering effect resulting from the main reflection on the floor. This effect is highly mitigated in the response corresponding to the the shorter T_1 windowing, although other inaccuracies appear derived from the loss in frequency resolution and, to a less extent, from other effects related to the neighboring loudspeakers of the array. In

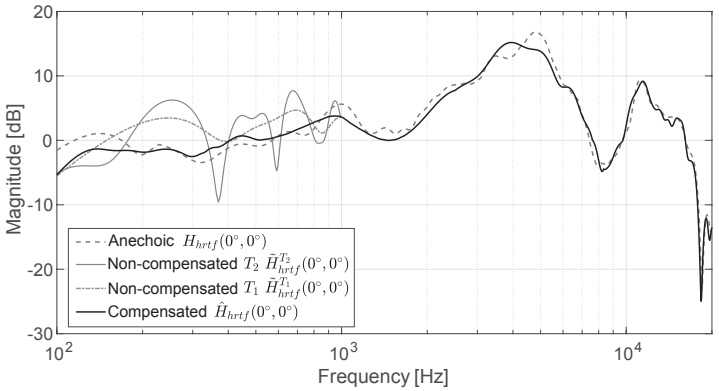


Figure 4.24. Comparison between the anechoic, non-compensated and compensated HRTFs

contrast, the HRTF obtained with the proposed compensation method is free from the low-frequency ripples produced by the room reflections and shows a response that is closer to the one measured in the anechoic chamber.

The most notable differences between the compensated HRTF and the anechoic one appear in the 100-150 Hz band, which are probably due to the effects caused by the T_2 cropping window. However, above this frequency, the results are very well aligned with the objective of the proposed method, which aimed at the full-band suppression of room reflections in non-anechoic HRTFs. This can be more clearly observed in Table 4.1, which shows a more detailed evaluation of the error. It contains the average error with respect to the anechoic response for the estimated and non-compensated signals in half octave bands (according to ISO standard frequencies). The error has been computed by averaging the absolute value of the error in dB for each frequency bin within each band, i.e.

$$\text{Error}_i\ [\text{dB}] = \frac{1}{N_i} \sum_{\omega_k \in B_i} \left| \bar{H}_{hrtf}(\omega_k)[\text{dB}] - H_{hrtf}(\omega_k)[\text{dB}] \right| \tag{4.35}$$

where N_i denotes the number of frequency bins in band B_i and $\bar{H}_{hrtf}(\omega_k)$ the response under evaluation at the k-th frequency bin. For the compensated version, the error is always below or around 1 dB, except for the first band due to the reason previously commented. However, for the non-compensated version T_1, the error is higher than 5 dB for one band and is around 3 or 4 dB for other two bands. An important result is the obtaining

of the 500-700 Hz band, which contains the effect of shoulder reflections. The numerical analysis of the measurements confirms both the validity of the proposed method and its advantages.

Finally, it is worth commenting on an interesting aspect that, although not related to the low-frequency correction of the response, is important to take into account. By looking at the HRTF results, some differences are observed around 4-5 kHz. There is a level mismatch between the anechoic signal and the estimated HRTF. Such undesired effect is thought to be related to the acoustic diffraction caused by loudspeaker edges. In fact, when measuring inside the non-anechoic constructed room, speakers followed a circular array arrangement, working as an infinite screen. As a result, diffraction is minimized and it might occur at a much lower amplitude. On the other hand, the loudspeaker in the anechoic chamber was set alone, producing wave diffraction on the edges. As a result, the frontal radiation due to this effect appears as peaks in the temporal response after the arrival of the direct sound. Note that the array-based measurement set-up improves this issue, obtaining an even clearer HRTF.

ISO 1/2 octave band	**Non-compensated T_1**	**Compensated**
125 - 175 Hz	1.36 dB	2.21 dB
175 - 200 Hz	3.84 dB	0.47 dB
200 - 250 Hz	4.44 dB	0.44 dB
250 - 350 Hz	5.16 dB	0.77 dB
350 - 500 Hz	1.81 dB	1.18 dB
500 - 700 Hz	3.10 dB	0.42 dB
700 - 1000 Hz	2.28 dB	1.09 dB

Table 4.1. Average error per 1/2 octave band for the non-compensated measurement and the proposed method with respect to the anechoic measurement

4.5 Supplementary headphones measurements and compensation filters

As described in Chapter 3 *Effects of Headphones in perception*, the response of the headphones employed to reproduce binaural sound has an important influence on perceptual characteristics such as perceived quality and spacialization. The localization of sound sources and the naturalness of the timbre are affected by the response of the headphones [67, 115], which can ruin the realism and accuracy of the binaural listening experience.

For an authentic simulation, the auralization system should be transparent, so the synthesized signal should be indistinguishable from the natural sound field. Therefore, one main concern is to optimize the compensation of the electroacoustic transducers involved in the binaural simulation system. The loudspeaker and the microphones used for BRIR recording and the headphones used for reproduction need to be considered in this compensation.

Headphones have a non-flat frequency response, therefore, it is necessary to use an equalization filter to compensate the effect of the Headphone Transfer Function (HpTF). Essentially, the transfer function of the headphone should be cancelled out with the equalization filter, so that when playing the equalized signal through the headphones, both transfer functions should counteract each other and the listener receives an unaltered version of the rendered binaural audio [95].

When evaluating headphone compensation, it should be kept in mind that the transfer function of the headphone comprises of the transfer function of the transducer itself and of the transfer function from the transducer to the individual's ear canal [192]. Appropriate design goals for headphone frequency responses have been proposed [73, 116, 193, 194], and depending on the application, free-field or diffuse-field responses calibrated headphones are accepted. However, according to measurements in [115], differences in frequency responses of different headphones can be as large as variations in HRTFs. Besides, depending on the type of headphone (extraaural, circumaural, supraaural, suspended on concha etc.), leakage due to small air gaps was shown to contribute to variability in the low frequency response [195]. Furthermore, the variability of measurements of Headphone Transfer Functions (HpTF) for different subjects is significant

and perceptually noticeable [95, 116]. For a single headphone, both studies revealed interindividual differences up to ±10 dB. Therefore, individual headphone compensation was recommended for binaural reproduction by different authors [77, 95, 116, 196].

The HRTF and HpTF are highly dependent on the morphology of the ear [194]; therefore, the highest degree of authenticity is only achieved when an individualized pair of HRTF and headphone equalization filter (HpEQ) are used [192]. Even equalization of headphones with generic non-individual HpEQ filters still yield significant perceptual benefits compared to unequalized reproduction [197], including a reduction in coloration, an increase in overall quality, and an improved perception of externalization and distance [77, 197, 198]. Thus, headphone equalization is always recommended for binaural reproduction.

Apart from interindividual variations of HRTFs and the variations of different headphones transfer functions, the variability only due to repeated placement of headphones on the same subject has also a certain impact [98, 192]. For a supraaural headphone, differences of ±4dB below and up to ±10dB above 10 kHz were reported in [98], and in [192], largest differences were reported below 500 Hz and above 10 kHz. In [98] it was shown that the result of an equalization based on a single transducer measurement without repositioning can become worse than having no compensation at all. Inverse filters should therefore be derived from an average of multiple measurements carried out while successively repositioning the headphones.

In order to compensate for the HpTF, headphone equalization filters (HpEQ) can be calculated using frequency-dependent regularization [199], which has been shown to perform better than other methods in perceptual tests [200]. In general, the goal of regularization is to avoid the excessive boost of certain frequencies that happens if the direct inverse of the transfer function is used as a compensation filter, thus preventing distortion and sensitivity to measurement errors [199]. In the particular case of headphone compensation, regularization prevents the inversion of narrow notches at high frequencies, which could lead to ringing artifacts when the headphones are repositioned after the measurement [200]. For this reason, a frequency-dependent regularization parameter must be set to a higher value at frequencies where those notches are present. While this parameter has traditionally been manually adjusted by expert listeners [97, 200], Gomez Bolaños et al. [97] have proposed an automatic procedure by which to

calculate it, demonstrating positive objective and perceptual results. The latter approach is the adopted method to compensate the response of our measured headphones. According to this method, the regularized inverse (here called sigma regularized inverse) $H_{SI}^{-1}(\omega)$ of a headphone response $H(\omega)$ is calculated with Equation 4.36,

$$H_{SI}^{-1}(\omega) = \frac{H^*(\omega)}{|H(\omega)|^2 + [\alpha(\omega) + \sigma^2(\omega)]} D(\omega) \tag{4.36}$$

where $D(\omega)$ is a modeling delay to ensure that the filter is causal, $\alpha(\omega)$ is the parameter which defines the bandwidth and maximum amplification of the filter, and $\sigma^2(\omega)$ is an estimator of the amount of regularization needed within the inversion bandwidth. Equations 4.37 and 4.38 define parameters $\alpha(\omega)$ and $\sigma(\omega)$, respectively.

$$\alpha(\omega) = \frac{1}{|W(\omega)|^2} - 1 \tag{4.37}$$

$$\sigma(\omega) = \begin{cases} |H(\omega)| - |\hat{H}(\omega)| & \text{if} \quad |\hat{H}(\omega)| \geq |H(\omega)| \\ 0 & \text{if} \quad |\hat{H}(\omega)| < |H(\omega)| \end{cases} \tag{4.38}$$

$\alpha(\omega)$ is calculated from a unity-gain passband filter $W(\omega)$, which delimits the bandwidth within which the headphones are equalized. $\sigma(\omega)$ is defined as the negative deviation of the headphone response $H(\omega)$ from a smoothed version $\hat{H}(\omega)$, which will be larger in zones with narrow notches [97].

The actual HpTF measurements carried out were made with exponential sweeps [130] (also described in previous Section 4.2.3) of 5 seconds of duration in the 20-22000 Hz band, with 5 different repositions over each subject. Then the Gomez Bolaños' method [97] is employed to obtain the individual HpEQ filters (or HpTF compensation filters) for each person and headphone model. A 1/6 octave smoothed version of the average of the five repositions ($H(\omega)$) is applied to the sigma regularized inversion method, with a 1/2 octave smoothed version ($\hat{H}(\omega)$) as the regulator for the $\sigma(\omega)$ parameter. The band where the inversion is applied is 20-18000 Hz. Figure 4.25 shows the average of the responses for one headphone model $H(\omega)$, the direct inverse $H^{-1}(\omega)$ and the sigma regularized inverse $H_{SI}^{-1}(\omega)$.

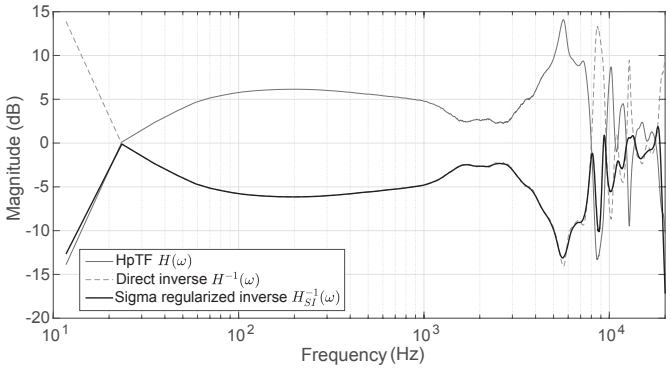


Figure 4.25. Example of an averaged HpTF $H(\omega)$, its direct inverse $H^{-1}(\omega)$ and the sigma regularized inverse $H_{SI}^{-1}(\omega)$

As seen in Figure 4.25, the HpEQ filter (sigma regularized inverse $H_{SI}^{-1}(\omega)$) is similar to the direct inverse of the HpTF, except that amplification is reduced outside the defined headphone bandwidth (20-18000 Hz in this case) and in zones with narrow notches; this is particularly noticeable around 8 and 13 kHz.

The obtained HpEQ filters are transformed to their minimum phase version in order to remove any possible delay and to have a short length. In this way, these filters can be used directly to reproduce binaural content with individualization (headphone model and personal coupling effect) with low computational cost.

The headphones measurements performed include the response of the microphones employed for the BRIR recording. Thus, the headphones compensation filter HpEQ also corrects for the microphone responses. Moreover, the possible influence of the position of the microphones inside the ear canal is also compensated with this filter, since both the BRIR and the HpIR of different headphones are recorded in the same session with the same individual microphone position.

A set of individual HpTF from various headphones models has been measured for each different individual. With this collection, individualized HpEQ compensation filters have been created for each subject and headphone model following the described procedure. The collection of measurements and compensation filters of the different headphone models includes the responses measured over the generic dummy heads Brüel 4100 [189]

and Neumann KU100 [201].

4.6 Conclusions and future work

To conclude this chapter, a summary of achievements can be presented to assess the work and results obtained in relation to the HRTF measurements.

A complete measurement system has been constructed from scratch, both hardware and software. It includes the conditioning of the room acoustics to reduce the reverberation time and the main reflections from the walls, the construction of a set-up with 88 loudspeakers (an array of 72 for the horizontal plane, 8 high and 8 low positions) driven by four sound cards, the adaptation of miniature microphones to be inserted into the ear canal of people, and an optical laser system to place the subjects to be measured with precision. Besides, the measurement software that allows the use of Exponential Sweeps (ES) [130, 167] and the Multiple Exponential Sweep Method (MESM) [154] was implemented. Additional functions were also developed to quickly check the measurements and their accuracy, as well as to store the measurements in SOFA format [21].

Specific and adapted post-processing has been implemented and developed in order to perform corrections and refinements in the measurements. It includes level and speaker response corrections, removal of reflections of mid and high frequencies with a variant of Frequency Dependant Windowing [175, 176] and the reconstruction of the lowest frequency band [183]. This post-processing of the measured BRIRs leads to the obtaining of *quasi* HRTFs clean of reflections.

Furthermore, a new method has been proposed and validated to completely remove the reflections that affect the low frequency band (from 100 up to 1000 Hz depending on the measurement conditions and configuration). It is based on measurements with spherical microphone array and Plane Wave Decomposition (PWD), and is able to eliminate the main reflections of the room that are responsible for comb filtering effects in this frequency band.

The measurement system also contemplates the individualized measurements of different headphone models on each person to be measured. The obtained HpTF is used to generate individualized HpEQ filters with

an automatic regularized inversion method [97], which ensures the absence of timbre coloration of the headphones and significantly improves the spatial localization, the perception of externalization and the naturalness of virtual sound sources.

The accomplishment of the initial objectives proposed in the introduction of this chapter leads to the following main conclusions:
1 - High accuracy measurements can be made with the constructed system, with fast acquisition (2 minutes and 28 seconds) in a non-anechoic environment. A combination of different techniques and innovative solutions have been implemented and developed to meet the requirements of accessibility to individualized HRTF measurements, and to be performed in an affordable installation that does not depend on anechoic chambers.
2 - The personal measurements that are made with the constructed system and the implemented methods enable to do research on the individualization of the HRTF. The precision and reliability of the measurements and therefore their usefulness is demonstrated by various experiments involving perceptual tests, presented and described in Chapter 5.

Finally, it should also be noted that a HRTF collection of more than 30 people has been measured, as well as their individual HpTF with different headphone models. The gathered data is of great value to investigate the individualization of HRTF and is already being used to synthesize binaural audio content.

Future work

Several improvements have been planned for the constructed measurement system. The most immediate may be to increase the number of elevated loudspeakers to acquire a denser spatial grid of real measurements. It has also been considered the possibility of introducing a turntable to rotate the person to be measured, thus creating a faster measurement system that can combine static measurement points and dynamic orientation of the subject.

Interpolation techniques to increase the spatial resolution of the HRTF collection has also been pretested over the measurements obtained. It will be desirable to optimize an interpolation technique to our specific case.

With the aim of performing faster measurements, it would be interesting to try other miniature microphone models that do not produce intermodulation distortions with MESM measurements. Avoiding this problem

with different microphone models, the measurement duration could be drastically reduced with a full MESM (interleaving and overlapping) probably to less than 2 minutes.

Finally, a very interesting future work will be to refine the FDW post-processing software in combination with the proposed method based on PWD to eliminate reflections. A better adjustment of both methods specifically for our particular room and system and considering all the directions measured, will be very useful to obtain a complete cleaning of the reflections of the HRTFs in an automatic and fast way.

5 HRTF individualization tools

The individualization of HRTF is a major milestone to be achieved in the development of commercial binaural sound reproduction systems, as discussed in the Chapter 2 *Background*. A personal HRTF introduces such an improvement in the perception of realistic sound, not only considering spatial perception, that has made the individualization of HRTF a focus of research.

HRTF individualization is useful for correcting classic perceptual problems such as front-to-back confusion, erroneous perception of elevation, inaccuracy in the general localization of sound sources, lack of sound externalization or inside-the-head effect [24, 202, 203, 204, 205]. All of these perceptual problems can be eliminated by using the individual HRTF of the listener or they can also be improved with a HRTF adapted to the person's own [68, 91, 203, 204, 206]. The basic problem of individualization is that HRTF depends on the morphology of the individual and that our sense of hearing is very sensitive to differences in HRTF. Thus, in order to be accurate, HRTF must be obtained from real, direct measurements of the individual himself. Therefore, individual HRTF can be obtained either by directly measuring the response of the individual or by synthesizing it from a precise 3D model of the subject's morphology [207, 208].

As already mentioned, an additional possibility is to adapt an HRTF so that it becomes individualized, getting as close as possible to the individual's own perception. In this way, the individualized binaural sound with an HRTF that has gone through an individualization process can solve one or several of the problems listed above, obtaining a significantly more correct result that can be crucial depending on the case of use or application. There is therefore a need for tools to induce individualization of HRTF. There are different strategies to try to individualize HRTFs. An introduction to different techniques can be found in Section 2.4.3 *HRTF individualization techniques*.

In this work the HRTFs are studied from a perceptual point of view in order to propose some of these useful tools from a practical perspective, with the general aim of the individualization of the HRTF. The study of HRTF has been divided into two different parts, on the one hand the analysis of the spectral form (magnitude of the HRTF) and on the other hand the investigation on the temporal interaural relations (Interaural Time Difference, ITD). This criterion is given by the studies which have demonstrated that a minimum phase HRTF together with the reconstruction of the ITD, results in HRTF perceptually indistinguishable from the originals, given that our sense of hearing is little sensitive to phase variations without a previous reference [82, 209, 210, 211, 212]. Thus, it has been sought the development of individualization tools for the magnitude of the HRTF (which can be corrected to be minimum phase) and for the ITD. In this way, the results of both studies are perceptually complementary.

In Section 5.1 *HRTF magnitude parametric modeling*, a method for parametrizing the HRTF magnitude has been analyzed and tested. The tested algorithm manages to extract the values of the relevant peaks and notches in the HRTF and synthesize them with a set of filters. This parametrization gives the possibility to study perceptually the personal variations of the HRTF, giving freedom to manipulate and simplify them if desired. In this case, a method for direct individualization of the magnitude has not been developed, but a tool for its manipulation that can be used for fast and efficient modeling that allows its adaptation.

In Section 5.2 *ITD scaling*, a perceptual study in relation with different variations of ITD is presented. The confrontation analysis of the perceptual results together with antropometric data has led to the generation of a simple tool (from a practical point of view) for individualization of

ITD. Adaptation is achieved through the scaling of the ITD towards the individual's own values.

5.1 HRTF magnitude parametric modeling

5.1.1 Introduction and motivation

Generic binaural cues can be classified into interaural differences (ITD and ILD) and timbre variations produced by spectral modeling of sound as it reaches the eardrum [2]. When there are no interaural differences or these are very small, for example in the middle sagittal plane (or median plane) of the listener, the spectral modeling (i.e. the shape of the HRTF magnitude of each ear) is used perceptually for the location of sound sources. Spectral information (HRTF) is therefore of great relevance for sound localization, especially for critical locations such as positions at different elevations [78] or in the so-called cone of confusion, which have the same ITD cues, or the torus of confusion which have the same ILD cues [213].

Determining the fundamental frequencies of the peaks and notches of the HRTF is therefore of great interest for the general study of spatial sound perception, as well as for the synthesis of HRTFs [214]. Specifically, spectral distortions caused by pinnae in the high-frequency range approximately above 5 kHz act as cues for median plane localization [78]. That is, the shape of the spectrum in the high frequency range is a key cue responsible for the perception of the elevation of a sound source. Mehrgardt and Mellert [209] have shown that the spectrum changes systematically in the frequency range above 5 kHz as the elevation of a sound source changes. Shaw and Teranishi [193] reported that a spectral notch changes from 6 to 10 kHz when the elevation of a sound source changes from -45° to +45°. Iida et al. [215] concluded that spectral cues in median plane localization exist in the high-frequency components above 5 kHz of the transfer function of concha. Hebrank and Wright [216] carried out experiments with filtered noise and reported that spectral cues of median plane localization exist between 4 and 16 kHz; front cues are a 1-octave notch having a lower cutoff frequency between 4 and 8 kHz and increased energy above 13 kHz; an above cue is a 1/4-octave peak between 7 and 9 kHz; a behind cue is a small peak between 10 and 12 kHz with a decrease in energy above and below the peak. Moore et al. [217] measured the thresholds of various spectral peaks and

notches. They showed that the spectral peaks and notches that Hebrank and Wright regarded as the cues of median plane localization are detectable for listeners, and thresholds for detecting changes in the position of sound sources in the frontal part of the median plane can be accounted for in terms of thresholds for the detection of differences in the center frequency of spectral notches. Asano et al. [218] carried out median plane localization tests with the HRTFs smoothed. The results indicate that major cues for judgment of elevation angle exist in the high-frequency region above 5 kHz, and that the information in macroscopic patterns is utilized instead of that in small peaks and dips.

The results of these previous studies imply that spectral peaks and notches due to the transfer function of the pinnae in the frequency range above 5 kHz prominently contribute to the perception of sound source elevation. Furthermore, there might be a potential of HRTF modeling based on the knowledge on spectral cues.

Then, the median plane is perfect for testing any HRTF modeling algorithm. The median plane, by minimizing interaural differences, allows to concentrate on HRTF modeling without taking into account the temporal relationships between the two ears, i.e. the ITD. Thus, if a HRTF modeled for the median plane produces a correct localization of the sound source, we can confirm that the HRTF modeling is correct without the influence of other interaural variables. The ITD can then be synthesized in parallel by other methods to obtain a complete HRIR.

There are previous methods for modeling the magnitude of HRTF. Parametric methods describe HRTFs with a limited set of parameters, reducing the amount of information needed to encode HRTFs [219]. Blommer and Wakefield [220] present a parametric model based on logaritmic error criterion, Haneda et al. [221] and Liu [222] propose common-acoustical-pole and zero models, Durant and Wakefield [223] suggest the use of their implemented method of modeling based on genetic algorithms, and Kulkarni and Colburn [224] provide with two IIR model arquitectures. All these methods are focused on looking for a model that simplifies HRTF to perform more efficient processing, but all of them (except for [224]) do not provide perceptual validation. Iida et al. [225] present a parametric model to extract and recompose peaks and notches of HRTFs; they perform perceptual tests to investigate the role of the extracted parameters and conclude that the parametric HRTF recomposed using the first and second notches and the

first peak provides almost the same localization accuracy as the measured HRTFs. This modeling concentrates on correctly determining each peak and notch to obtain an accurate parameterization of the HRTF.

The present work about HRTF modeling has two purposes:
1 - The first is to evaluate an already existing method to parametrically model spectrums [226] to see if it is capable of model HRTF that simulates the vertical localization of sound. The parametric HRTF is recomposed only of the spectral peaks and notches extracted from a measured HRTF, and the spectral peaks and notches are expressed parametrically as a filter with central frequency, gain, and quality factor. Vertical sound source localization in the median plane is evaluated to avoid interaural differences.
2 - The second is to explore the number of parametric filters that determine the spectral resolution (number of peaks and notches of the HRTF) needed to preserve the vertical localization. A reduced number of filters can provide with a simpler HRTF model that can have a significant computational advantage. Besides, a perceptually appropriate low-order representation of the HRTF may provide insight into sound localization mechanisms. This information can be very useful, as gaining this insight could result in computational methods for generating HRTFs that would not rely upon making empirical measurements from individuals.

In pursuit of these goals, a perceptual experiment was conducted. In this experiment, different subjects compared their own individually measured HRTFs with simplified versions of them. The simplified versions of these HRTFs include only some of the original peaks and notches. To model these peaks and notches, a parametric modeling method previously developed in the doctoral candidate's research group [226] was used. This parametric modeling method was originally intended to generate efficient filters for equalizing loudspeakers, and provides an optimized chain of second order section (SOS) where each section is an IIR PEAK filter, defined by the parameters gain, central frequency and quality factor. The perceptual experiment aims to test the already existing modeling method with the different purpose of modeling HRTFs, and wants to find out the number of peaks and notches that need to be modeled to achieve localization results comparable to those achieved with individual HRTFs measured from real subjects.

The methods used to carry out the perceptual experiment are described below, detailing the HRIR measurement and the pre-processing used in the

experiment. This is followed by a presentation of the parametric model procedure and a description of the perceptual tests performed. Finally, the results are analysed and some conclusions are extracted from the data obtained.

5.1.2 Individual HRIR measure and preprocessing

HRIR measurement

To measure the individual HRIRs in different vertical positions, a specific support was built in the shape of an arc of circumference to hold the loudspeakers in five angles of the median plane (65°, 40°, 15°, 0° and −20°) equidistant from the subject's ears 1.5 m (Figure 5.1). As these angles will be employed in the perceptual test, an acoustically transparent curtain was placed in front of the loudspeaker setup (Figure 5.2) to avoid visual cues or to reveal the location of test positions to the measured subjects who will subsequently perform the perceptual test.

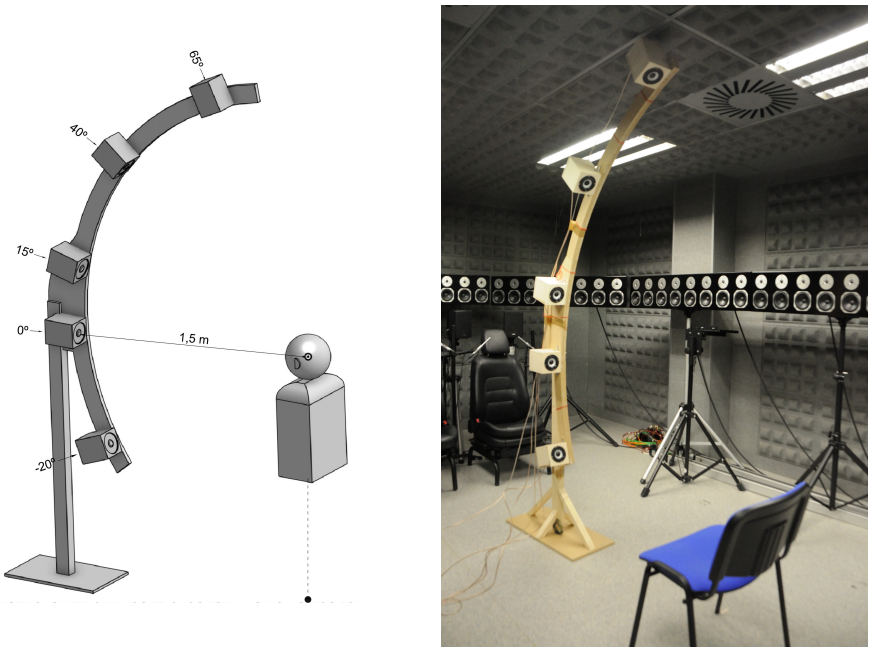


Figure 5.1. Loudspeakers for the measure of the median plane individual HRTFs

The HRIRs of the different angles were measured for each subject following a procedure similar to that explained in Section 4.2.4 (but with the different speaker setup). Miniature omnidirectional microphones Knowles FG23329 (described in Section 4.2.2) were employed to measure with blocking ear canal i.e. blocking meatus condition [83]. The microphones were placed inside the ear canal using a soft earplug with a hole in the middle (see Figure 4.6 for examples), which allows reliable measures and is easy to implement [67, 178]. Then, exponential sweep signals were used for measuring the impulse responses [130], as is fully described in Section 4.2.3.

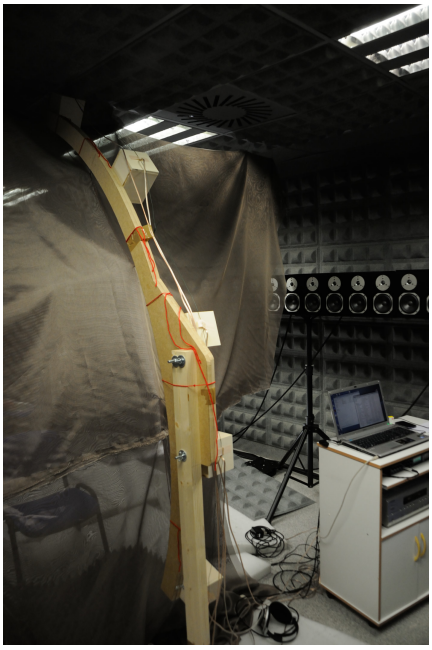


Figure 5.2. View of the set up for the measure of the HRTFs from behind the acoustically transparent curtain

Loudspeakers response correction

The loudspeaker responses were also measured individually with the same procedure, using the same microphones employed to measure the HRIR of the subjects. The magnitudes of these loudspeaker measurements were used to correct each of the measured HRTF of the subjects, with a procedure similar to that described in Section 4.3.1. Smoothed 1/3-octave versions of

the loudspeaker inverted responses were applied to the individual measured HRTFs (each vertical position was corrected with the corresponding loudspeaker response), and the final responses were reduced to 2048 samples. This makes it possible to eliminate the effect of the loudspeakers and the error due to possible disparities between them.

Removal of reflections

The measurements acquired were actually Binaural Room Impulse Responses (BRIRs), because they also include the reflections of the room. Then, a variant of the Frequency Dependant Windowing (FDW) method described in Section 4.3.2 was employed to remove those reflections and obtain *quasi* HRIRs. The acoustic treatment of the experimental room produces rather damped reflections, so the FDW achieves very good results in these measurements. With this process the main room reflections recorded in the BRIR were mostly erased, but preserving the information of the personal HRIR. Five variable time windows are used in this case. The time windows employed were basically rectangular with a fade out at the end (half Hanning) to allow for a soft transition. The first reflection due to the room geometry is found 24 samples away from the peak of the impulse response (sampling frequency 48 kHz). Hence, five different lengths were used to separate information of the impulse response, 24, 48, 96, 192 and 384 samples. As the HRIRs were recorded with a sampling rate of 48000Hz, these windows provide a spectral resolution of 2000, 1000, 500, 250 and 125 Hz respectively. A Fast Fourier Transform (FFT) were applied to these five pieces of temporal information. From the resulted spectra, the corresponding subband was taken according to the different resolutions of the temporal windows. Then, the five spectral subbands were collected to construct the magnitude response, 125 to 250 Hz, 250 to 500 Hz, 500 to 1000 Hz, 1000 to 2000 Hz and 2000 to 20000 Hz. The transitions between the subbands were summed with a 1/6 octave of resolution to smooth the response and avoid unwanted distortions in the transitions. Figures 4.11 and 4.12 shows examples of time windows and the corresponding extracted frequency bands (for a different number of samples and frequency bands). This methodology ensures that the most important frequency band for elevation (5 to 20 kHz) is free of reflections, even though the measurements were not performed in anechoic conditions.

As a result of the processing of the measurements, the HRTFs of each subject were obtained in the band from 100 to 20000 Hz. In Figure 5.3, an

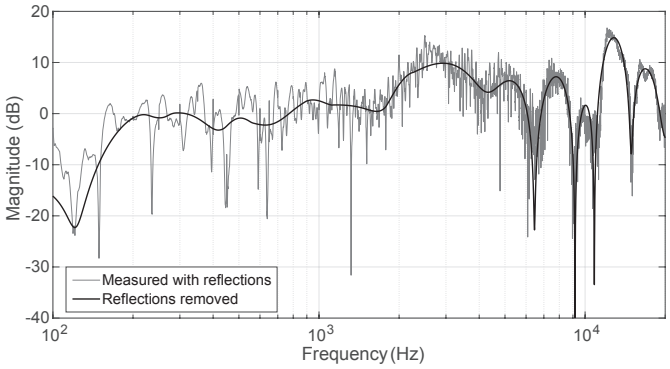


Figure 5.3. Example of a measurement with reflections and the obtained HRTF

example of an original BRTF measurement and the corresponding HRTF obtained is shown.

Headphone compensation

Besides, the Headphone Transfer Function (HpTF) of the reference headphones to be used during the perceptual test was measured over the subject's ears (five measurements with reposition for each subject).

The reference headphone Sennheiser HD800 headphone was used for the reproduction of all the different stimuli of the test. To compensate the spectral response of the headphone, the measured HRTFs as well as the individualized modeled HRTFs (as will be seen in the next section) used in the test were corrected using the automatized regularized inverse method [97] described in Section 4.5. The compensation is applied as described by the Equation 5.1

$$H_{compensated}(\omega) = \underset{\text{or } model}{H_{measure}}(\omega) \cdot HpTF_{SI\ HD800}^{-1}(\omega) \tag{5.1}$$

where $H_{measure\ or\ model}(\omega)$ is the measured or modeled HRTF, $HpTF_{SI}^{-1}(\omega)$ is the compensation filter to equalize the headphone, and $H_{compensated}(\omega)$ is the compensated response to apply to each stimulus of the perceptual test.

5.1.3 Parametric model description

The parametric modeling method employed for this experiment is based on a previous algorithm initially developed to equalize loudspeakers [226] in the same research group of the doctoral candidate, and later employed to simulate elevation in Wave-Field Synthesis systems [227]. In this work, the algorithm has been used to model the HRTF of different subjects. The method described in the two previously cited papers provides a way to compute a filter chain of Second Order Sections (SOS) PEAK IIR filters that models the peaks and notches of an HRTF.

A peak filter is a second-order parametric filter, which enhances or attenuates with again $G(dB)$ at a bandwidth centered at a frequency f_c, and with a quality factor Q. Equation 5.2 shows the normalized analog prototype peak filter where A is the square root of the filter gain in volts.

$$H_{PEAK}(s) = \frac{s^2 + s\frac{A}{Q} + 1}{s^2 + s\frac{1}{AQ} + 1} \tag{5.2}$$

This minimum phase filter will be transformed to the discrete domain $H(z)$ by bilinear transformation, and the computation of the digital filter coefficients a_1, a_2, b_0, b_1 and b_2 from the filter parameters f_c, Q, G can be found in [228].

In this experiment, the target frequency response to model is a 1/3 octave smoothed version of the measured HRTFs in the median plane. The number of peaks and notches to model is a controlled factor N_{peaks}. In order to find the chain of PEAK filters that models the peaks and notches of the HRTFs, the algorithm is applied in two stages: First, it sets the initial values of the parameters for the PEAK filter $H_{PEAK\ i}(\omega)$. To do this it searches the biggest area determined by a peak or notch in the HRTF to model ($H_{measured}(\omega)$), and assigns the parameters of the filter. The central frequency f_i is calculated as the geometric mean between the zero-crossing points of the area, the log-gain G_i as the value of the magnitude at f_i and the quality factor Q_i by looking for the -3 dB points if they exists or is selected directly as 2.5. The second stage optimize the parameters of the filter by performing 300 iterative random variations over the initial values computed previously. Then, the two stages of the process are repeated for the second PEAK starting from the new filtered response $H_{measured}(\omega) \cdot H_{PEAK\ 1}(\omega)$.

When the N_{peaks} stages of the filter are designed a postoptimization procedure is done to improve the final response. This is performed to take into account the interaction effects between each PEAK. Variations of the values of the parameters in a range of 2% are applied in a process of 500 iterative random variations.

According to Equation 5.3, the resulting N_{peaks} PEAK filters model the measured HRTF $H_{measured}(\omega)$:

$$\prod_{i=1}^{N_{peaks}} H_{PEAK\ i}(\omega) \approx H_{measured}(\omega) \tag{5.3}$$

Finally, the algorithm provides with the parameters f_i, G_i and Q_i, as well as the a_i and b_i coefficients of each PEAK filter.

A detail to take into account is that the parametric modeling process was applied separately to each of the channels (left and right) of the HRTFs measured, thus preserving the influence of the morphological differences between ears in the same subject.

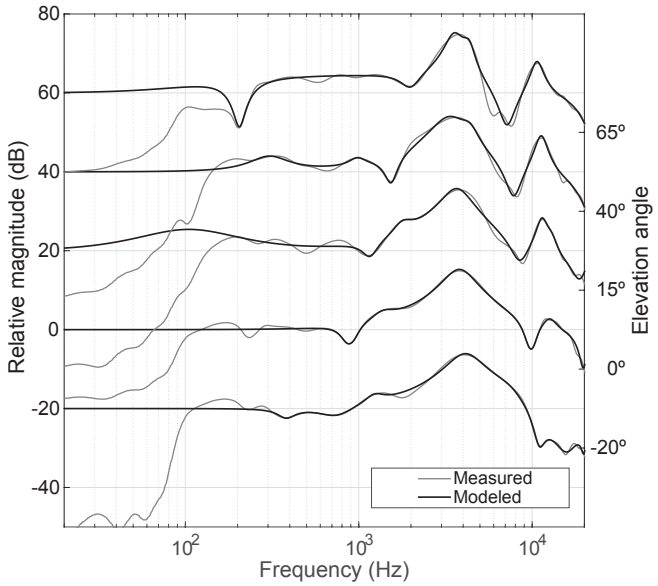


Figure 5.4. Comparison of the HRTF measured and modeled ($N_{peaks} = 8$) for different elevation angles of one subject. (offset 20 dB)

Figure 5.4 shows the comparison between the HRTF measured for one person and the corresponding modeled versions employing 8 PEAK filters ($N_{peaks} = 8$), for the elevation angles of the median plane $-20°$, $0°$, $15°$, $40°$, $65°$. The magnitudes are represented with a 20 dB offset to enable their visualization, and the measured versions are represented after the removal-of-reflections pre-processing described in the previous section and the 1/3 octave smoothing commented before.

5.1.4 Test description

The complete methods described previously of measuring, preprocessing and modeling the individual HRTFs, as well as the headphone compensation, were completely automatized. This made it possible to measure and compute the modeled responses of a subject with different number of peaks in just some minutes, and then have the subject ready to perform the perceptual test.

The objective of the test is to evaluate the capability of the parametrically modeled HRTFs to simulate the localization perceived with individually measured HRTFs, and also to check how many parametric filters are needed for that. As commented before, the median plane was chosen to test the parametric model, because it is considered that the spectral information is the main cue for median plane localization, not taking into consideration the interaural differences.

Some preliminary tests performed with the parametric model made it possible to realize the difficulty of the task of perceiving and locating the height of a sound source in the median plane. This experience led to take some design decisions:

- It was necessary to include a reference stimulus to compare the stimulus under evaluation. The reference was chosen to be at $0°$ of elevation.
- The type of sounds to be reproduced in the test need to have an important content in the two higher octaves, where is most of the information used to locate sounds in the median plane.
- Due to the difficult of the task of locate sounds in height, expert listeners were preferred.
- To evaluate the performing of the different number of PEAK filters to model the HRTF, the individual measured HRTF was included in

the test as a hidden reference.

– As the HRTF under evaluation were modeled from individual measured HRTF, the measured HRTF of another person was included in the test as an anchor to examine the effect of individualization.

The task of the listeners was to identify the angle of the sound source direction of the different stimuli. All the stimuli were presented to the subjects in a double-blind manner and in a random order. The participants could freely listen the different stimuli as many times as they wanted and then they indicated the value in degrees of the position perceived. They use as a reference a visual scale placed in a curtain in front of them, which marks the values each 10° from −30° to 70°, as shown in Figure 5.5.

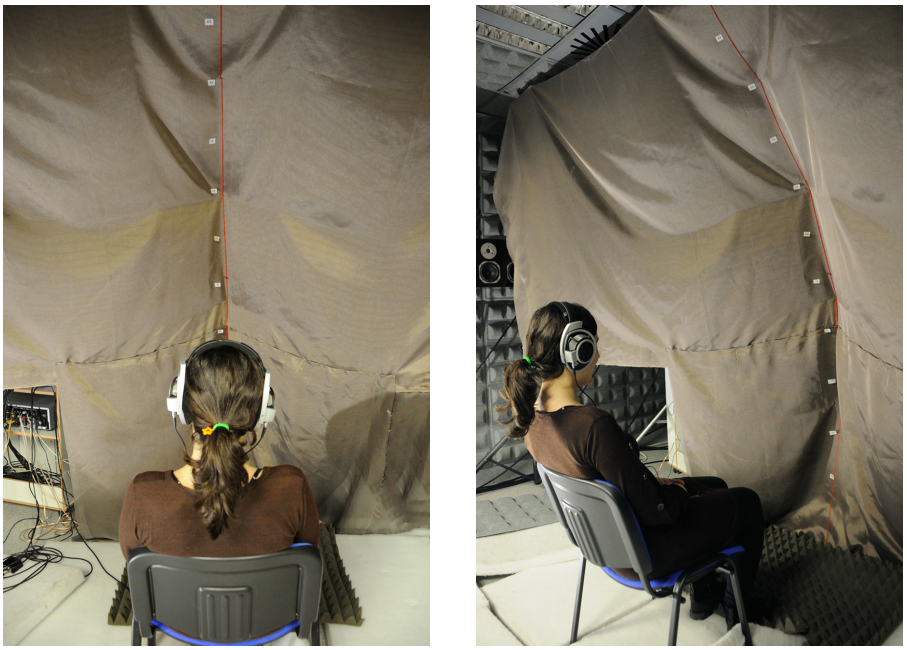


Figure 5.5. Subject performing the test in front of the acoustically transparent curtain and the visual scale (10° of separation)

The individual HRTFs of each participating subject were measured in the median plane at the elevated angles of 65°, 40°, 15°, 0° and −20°,

consequently the sound sources directions to test with the modeled HRTFs were the same.

Just before the actual perceptual test was performed, a brief training phase was conducted. In this phase, participants listened twice to some of the real loudspeakers and were first asked about the perceived position and then given the actual location. It was found that in the second round of listening the participants located the elevated sound sources quite well. In addition, this small training also helped to introduce the participants to the task of the test. Both during the training phase and during the actual test, participants were instructed to adopt an initial reference position with the head facing straight towards the 0° mark of the visual reference, before listening to each sound.

The perceptual test consists of two different parts, each of them modeled a different number of peaks and notches from different spectral bands. As previously said, spectral distortions caused by pinnae in high frequencies approximately above 5 kHz are described to be main cues for median plane localization [215, 218]. For this reason the modeling of the HRTFs was focused mainly in high frequencies.

The first part of the test evaluated the band from 750 to 20000 Hz, modeled with three different number of filters (N_{peaks}): 2, 4 and 6 PEAK filters. A burst of 2 seconds of pink noise high pass filtered from 1000 Hz was the sound employed. Then, the number of stimuli was: (3 modeled HRTF + 1 measured HRTF hidden reference + 1 other person HRTF anchor) x 5 angles = 25 stimuli. The sounds used in the test were high pass filtered to avoid the possible roll-off effect of the first filter.

Lower frequencies were included in the localization task of the second part of the test. The reason was to consider the possible influence of lower frequency peaks and notches that could help in the externalization of the sound sources. The cleaning preprocess is not perfect and some remains of the room effect can influence the externalization of the sound sources. It can also help to generate naturalness in the real sounds, so it was decided to include the sound of a guitar in this part. The impulsive sound of the chords of the guitar can be useful to localize the sound sources. A bigger number of PEAK filters was chosen to model the extended lower frequency band of the second part, making it comparable to the modeling of the band of the first part.

The second part of the test evaluated the band from 200 to 20000 Hz, modeled with two number of filters (N_{peaks}): 6 and 8 PEAK filters. A burst of 2 seconds of pink noise and a guitar sound of 5 seconds, high pass filtered (HPF) from 300 Hz, were the sounds employed for the second part of the test. The number of stimuli on this part was: (2 modeled HRTF + 1 measured HRTF hidden reference + 1 other person HRTF anchor) x 5 angles x 2 sounds = 40 stimuli.

Six expert listeners (2 women and 4 men) performed the test (24 to 36 years, average age of 29). The average run time was 7 and 11 minutes respectively for each part.

5.1.5 Analysis of the results

Figure 5.6 shows the results of the first part of the test. It represents the mean (with 95% confidence intervals) of the perceived angles for each type of HRTF, that were chosen for all the subjects who performed the test. The diagonal lines in the graphs indicate the ideal correct answers. A slight

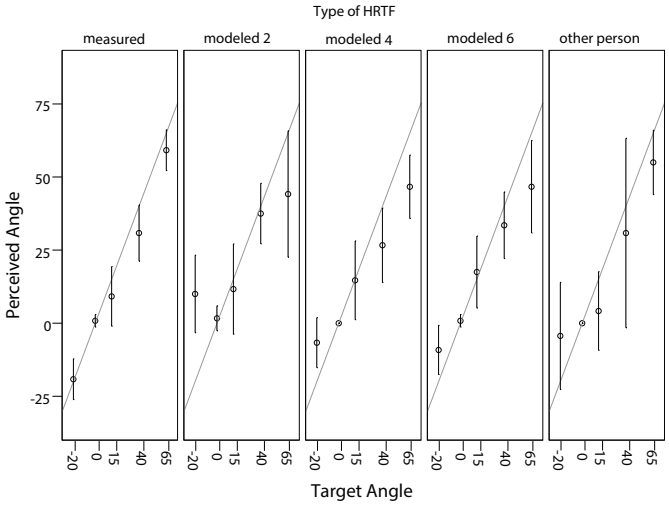


Figure 5.6. Part 1: Mean of the perceived angles for the measured, 750-20000 Hz modeled and other person HRTFs. For pink noise HPF 1000 Hz. (Confidence Intervals 95%)

general tendency to choose positions getting closer towards zero degrees can be seen. In the first block, "measured" HRTF indicates the results for the

individual HRTF measured from the subject itself, then there can be found the results for the three HRTF "modeled" with 2, 4 and 6 PEAK filters, and finally the results obtained with the non-individual HRTF of "other person" that acts as anchor stimuli.

As expected, Figure 5.6 shows that the "measured" HRTF produce the best adjustment of the answers to the real localization angles. The "modeled" HRTF with 2 filters presents significant localization errors for the elevation angles $-20°$ and $65°$, besides showing wider confidence intervals. However, starting from 4 PEAK filters, the "modeled" HRTFs resemble the localization performance obtained with the measured HRTFs. The "other person" HRTF have consistent results too, but the wider confidence intervals of the data reveal the worst accuracy of all the answers.

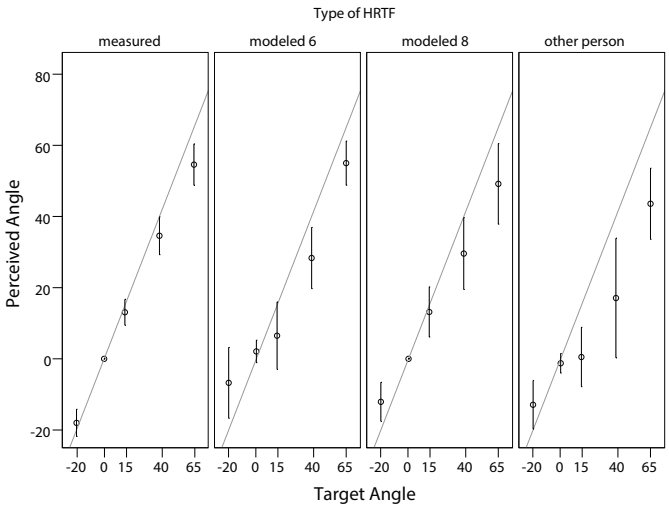


Figure 5.7. Part 2: Mean of the perceived angles for the measured, 200-20000 Hz modeled and other person HRTFs. For all sounds HPF 300 Hz. (Confidence Intervals 95%)

In Figure 5.7 the general results for the second part of the test are shown. It represents the mean (with 95% confidence intervals) of the perceived angles for each type of HRTF, for all the subjects who performed the test. Similarly to the results of the first part of the test, the same tendency to respond with angles towards zero degrees is found. In addition, it is also observed that the results obtained with "measured" HRTF are the most accurate and that the results of "other person" HRTF are those with the

worst accuracy, exhibiting higher confidence intervals. Besides that, the "modeled" HRTFs with 6 and 8 PEAK filters give similar results in the 200-20000 Hz band as the obtained in the first part of the test with the "modeled" with 4 and 6 filters in the 750-20000 Hz band. This is logical, since the increase in the number of filters for the second part was motivated by the extension of the frequency band with respect to the first part, to allow a similar resolution of the modeling in both cases.

These "modeled" HRTF versions of the second part of the test that include lower frequencies show a slight improvement in the localization of the lowest position $-20°$. This is probably due to some reflections from the floor that boost the mid-low frequencies and are better simulated with the low-frequency extension.

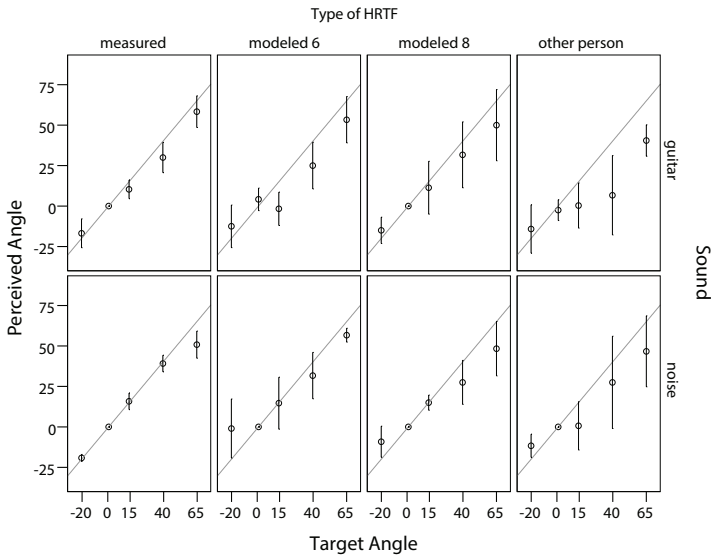


Figure 5.8. Part 2: Mean of the perceived angles for the measured, 200-20000 Hz modeled and other person HRTFs. For pink noise and guitar HPF 300 Hz. (Confidence Intervals 95%)

The type of sound have a strong influence in the localization results in the second part of the test. Figure 5.8 shows the results of the second part of the test, with each type of sound displayed separately. The pink noise shows more accurate results than the guitar sound. It is noticeable that the guitar sound produces worse localization results in the case of

"other person" HRTF. This seems to suggest that the impulsive temporal characteristic of the sound of the guitar strings may be more influenced by the individualized HRTF than in the case of stationary broadband noise.

5.1.6 Conclusions and future work

An experiment has been carried out to evaluate the performance of a parametric model for the individualization of the HRTFs in the median plane. Individual HRTFs from real people have been measured and then have been modeled with various numbers of PEAK filters. A perceptual test has been performed with the same measured people and their parametrically modeled HRTFs.

The parametric model is based on an existing algorithm [226] originally intended to equalize the response of loudspeakers. With the conducted experiment it has been proved that the same algorithm can be used to model HRTFs in the median plane with perceptually similar results that the obtained with individual measured HRTFs.

In addition, it has been shown that a low-order representation of the median plane HRTFs also provide with enough perceptual cues to properly localize the position of an elevated sound source. More specifically, it has been demonstrated that starting from 4 PEAK filters, it is possible to model an HRTF in the 750 to 20000 Hz band, which is capable of achieving localization results that resemble those obtained with the individualized HRTF measured. Similar results have been also obtained for the band of 200 to 20000Hz with a modeled HRTF starting from 6 PEAK filters.

Besides, the modeled HRTFs which include lower frequencies have been found to slightly improve the localization of the lowest $-20°$ position tested in the median plane.

Additionally, the non individualized HRTFs included in the test as anchor obtained worse results than the versions modeled from the individual HRTFs, which confirms the strong influence that the personal HRTF have in the perception of elevated sound sources. In particular, non-individualized HRTFs have been found to produce worse localization results, especially with guitar string sounds, suggesting that perception may be more sensitive to individualization with impulsive sounds than with broadband noise.

According to the data obtained with the perceptual experiment, the tested parametric model is revealed as a powerful tool to investigate the perceptual mechanisms of the HRTF. Psychoacoustic studies on the role of the peaks and notches in HRTF could easily be performed with this parametric model. In addition, the data reduction that parametrization implies can make it possible to generate computationally efficient binaural sound.

Future work

The experience acquired in this test inspires some improvements to continue this work in the future, both to upgrade the algorithm that performs the parametric modeling of HRTFs and for the design of perceptual tests to evaluate it.

It would be interesting to develop other ways of evaluating the perception of height, focusing on the comparison of positions. It would also be desirable to compare more impulsive and natural sounds with individualized and non-individualized HRTFs against synthetic broadband sounds. From the results of these tests, it is clear that the modeling algorithm can improve its performance by limiting the detection of certain peaks and notches to specific bands of the spectrum.

An interesting idea to explore is to use the parameters obtained as a result of the modeling to feed a machine learning algorithm, with which to evaluate the perceptual mechanisms of the auditory localization. Then, the parameters as well as the real HRTF would be the features that could help to identify the most relevant peaks and notches (or the relationships between them) to localize sound sources, and therefore employ them in an individualization HRTF synthesis algorithm.

5.2 ITD scaling

5.2.1 Introduction and motivation

The Interaural Time Difference (ITD) was described in the duplex theory by Lord Rayleigh [23] as a relevant cue for localization and an important factor of individualization specially for low frequency and azimuth localization [229, 230].

In the previous Chapter 5.1 a parametric modeling was described to replicate and model the magnitude of HRTFs. The obtained modeled magnitude does not record on the temporal information of the HRTF or the temporal relation between the left and right ears. We can then base on the assumption of a HRTF reconstruction with minimum phase magnitude and interaural phase differences approximated by constant time delay values between both ears, which will act as ITD [82, 210, 211, 212]. According to these studies, this reconstruction is perceptually indistinguishable from an HRTF measured in free field. Consistent with the above, this chapter studying the ITD is complementary to the previous magnitude modeling described, and required to achieve a complete HRTF reconstruction.

Different studies has been carried out in the past to model or estimate the ITD. Many of them present a model to calculate the ITD based on anthropometric dimensions. Woodworth [231] introduced one of the earliest models, based on the sound pressure transmission on the surface of a rigid sphere, given by Equation 5.4

$$ITD_{Woodworth} = \frac{a}{c_0}(sin\ \phi + \phi) \quad \text{for} \quad 0 \leq \phi \leq \frac{\pi}{2} \tag{5.4}$$

where a is the radius of the head, c_0 is the speed of sound and ϕ is the azimuth angle of incidence of the sound. Later, Kuhn [210] presented a more accurate model specific for lower frequencies, with the Equation 5.5

$$ITD_{Kuhn} = \frac{3a}{c_0} sin\ \phi \quad \text{for} \quad (ka)^2 \ll 1 \tag{5.5}$$

where k is the wave number. Other authors introduced elevation dependence to refine the previous models, as Larcher and Jot [232] model Equation 5.6

$$ITD_{Larcher} = \frac{a}{c_0}(sin^{-1}(sin\ \phi\ cos\ \theta) + sin\ \phi\ cos\ \theta) \tag{5.6}$$

and the simpler extension of Equation 5.4 by Savioja et al. [233]

$$ITD_{Savioja} = \frac{a}{c_0}(sin\ \phi + \phi)\ cos\ \theta \tag{5.7}$$

where θ is the elevation angle of incidence of the sound. Algazi et al. [234] proposed an average head radius formula to improve the use of the previous models, based on the head width w, depth d and height h

$$a_{Algazi} = 0.51w + 0.18d + 0.019h + 32 \quad (mm) \tag{5.8}$$

There are other models that take into account more anthropometric details or that approaches to the problem from different perspectives and techniques. Busson [235] consider the elevation and ear position dependency, Algazi et al. [180] include in their model the shoulder reflection, whereas Duda [236] and Bomhardt [237] propose geometrical ellipsoid models. Different analysis tools are employed in Aussal [238] which applies the principle component analysis (PCA) including head and torso dimensions, or in Zhong [239] that based their study on spherical harmonic basis functions. Similarly, Zhong [240] applies spatial Fourier analysis and multiple regression and Katz [208] derived the ITD by boundary element method calculations of the HRTF.

In this work, following the initial premise, a perceptual approach is considered for the study of the ITD. There is an interesting and related previous work by Lindau [241] that seeks the adaptation of ITD through a real time perceptual test. Algazi [24] also uses a perceptual criterion to explore elevation localization considering ITD among other cues. As mentioned before, here the idea is to approximate the ITD individual estimation problem from a perceptual perspective. To this end, an exploratory test was proposed to investigate the perceptual relationships between objective and anthropometric measurements with the ITD. A group of listeners were asked to locate binaural sound sources with different scaled ITD of their own HRTF and two dummy heads HRTF, and the results were contrasted with individual objective data.

5.2.2 BRIR measurements and extraction of objective variables

To investigate the individualization and personal characteristics of the ITD, individual HRTF measurements were required. Because of that, each participating subject who performed the perceptual test was previously measured and individual objective data was extracted from these measurements. In the same manner, some individual stimuli were also prepared for each subject. As the greatest variations of ITD occur at azimuth angles, the measurements and the perceptual test were focused on the horizontal plane.

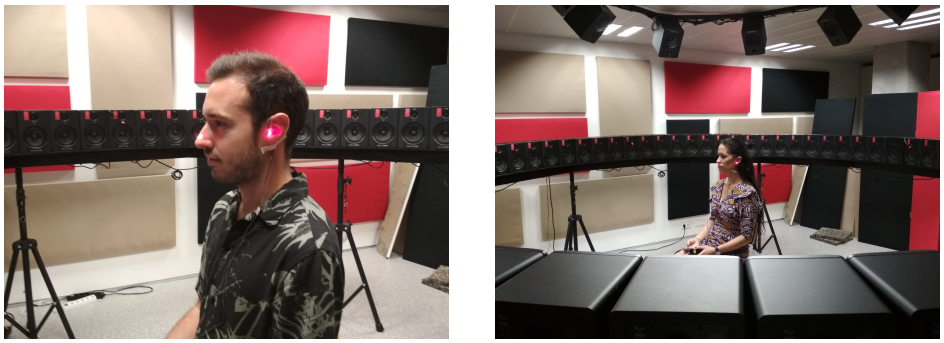


Figure 5.9. Examples of BRIR measurement of two people

Binaural Room Impulse Responses (BRIR) measurements were in fact performed in a measurement room and a system developed during the work of this thesis. A full description of this measurement system set-up and its characteristics can be found in Chapter 4. The resulting measurements include the BRIRs of the horizontal plane of each subject with 5° of resolution (72 different angles of incidence of the sound), with a high degree of position precision. Besides, a processing of the BRIR measurements was done to partially remove the reflections of the room and reduce the room effects, obtaining almost clean HRTFs. The method for the processing of the measurements is described in Section 4.3. The *quasi* HRTFs obtained were used to determine some objective individual data. Figure 5.9 illustrates the measurement process of two of the subjects.

Besides, two dummy heads were also measured in the same conditions and their HRTFs employed in the experiment, the widely used and known Brüel 4100 [189, 242] and Neumann KU100 [201, 243].

Individual HpTF responses of the reference headphone model Sennheiser HD800 were also measured for each subject participant in the perceptual test. The mean of five repositioned measurements was employed to generate individual inverse filters to compensate the response of the reference headphones used in the test. These filters were obtained with an automatic regularized method [97], which produces perceptually better equalization than the regularized inverse method with a fixed factor. A more detailed explanation of the method can be found in Section 4.5.

The following objective variables were measured or extracted from the BRIR measurements:

- Calculation of the ITD:

There are different methods for calculating or estimating the ITD, which are usually classified into three families:
1-Onset threshold (based on the detection of the time difference between the first onsets of the signals in the two ears. Onsets are selected relative to predetermined level thresholds as the first point in a signal that exceeds that level value [234]).
2-Cross-correlation (based on the relation for which the time delay maximizes the coherence between the signal of the two ears [91, 244]).
3-Group delay (based on phase analysis and estimation of group delay, by applying either a linear fit on the excess-phase component of the signals or a cross-correlation between the minimum and excess-phase component of the HRIR [245, 246]).
An extensive review and comparison between these different methods and some variants can be found in [247].

As this experiment is planned from a perceptual point of view, the selected method to estimate the ITD should be perceptually oriented. The first onset threshold detection method with a threshold of -3dB, applied on 300-3000Hz band pass filtered HRIRs, was employed here to estimate the ITD. The band-pass filter allows to suppress the high frequency contributions of the pinnae in the HRIR, and to avoid possible unwanted low frequency fluctuations due to reflections not completely eliminated in the measured BRIR. A threshold of -3dB provides high accuracy in the detection of local HRIR maximums, while ensuring homogeneous estimates between different subject measurements. According to previous studies [248], the chosen method closely resemblance the perceptually most relevant method to calculate the ITD. The HRIRs were previously upsampled

to 96kHz in order to have a higher resolution in the calculation of the ITD and its subsequent manipulations. Figure 5.10 shows the calculated ITD of the 21 subjects who took part in the perceptual test.

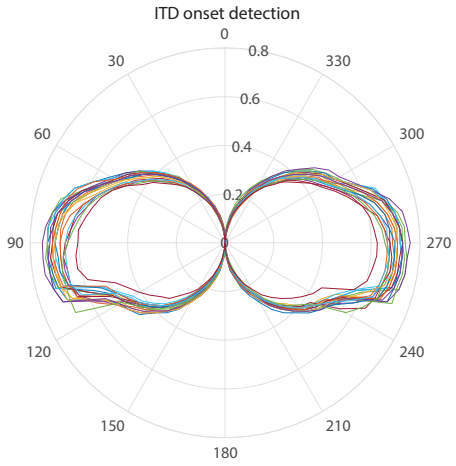


Figure 5.10. Polar plot of the ITD of the 21 participants measured for the perceptual test. Radial units are in ms

- Calculation of the ILD:

Although this perceptual study explores the ITD, the Interaural Level Difference (ILD) was also calculated for all subjects. The ILD objective values were examined in relation with the perceptual results. Equation 5.9 defines the ILD

$$ILD(f,\phi) = 20\ log\frac{|H_L(f,\phi)|}{|H_R(f,\phi)|} \tag{5.9}$$

where f is the frequency, ϕ is the source direction, and $|H_R(f,\phi)|$ and $|H_L(f,\phi)|$ respectively denote the magnitudes of the right and left HRTFs [249].

As this cue is perceptually dependent on frequency, ILDs were calculated for three different octave frequency bands centered on 500Hz, 2000Hz and 5000Hz. Figure 5.11 shows the calculated ILD for the three frequency bands of the 21 subjects who participated in the perceptual test.

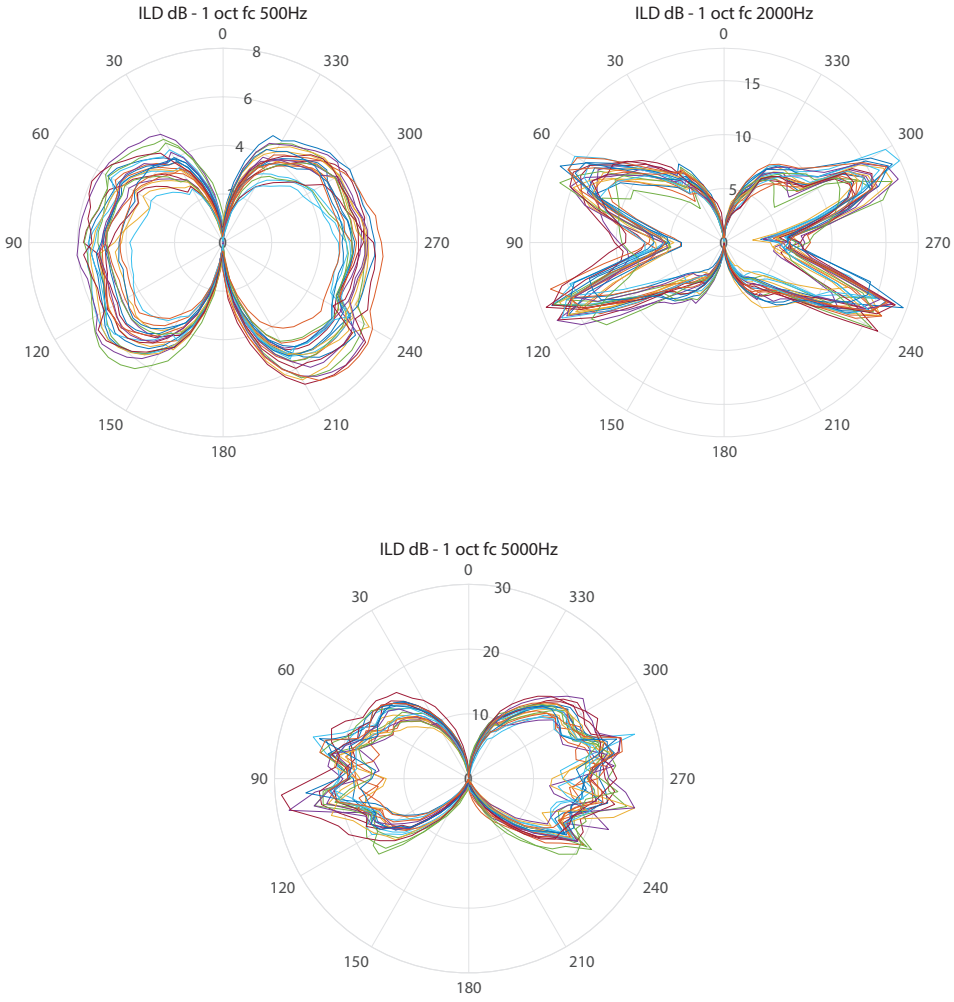


Figure 5.11. Polar plots of the ILD for one octave frequency bands centered in 500Hz, 2000Hz and 5000Hz, of the 21 participants measured for the perceptual test. Radial units are in dB

- Calculation of the Spectral Distortion:

The Spectral Distortion (SD) gives an objective score of the difference between two spectra. The SD between the subject own individual HRTF and the HRTF of the two dummy heads measured (Brüel and Neumann) was calculated for each measured angle [250], and then averaged for all source directions [249]. Equation 5.10 described the calculation

$$SD = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{W}\sum_{w=1}^{W}\left(20\ log\frac{|H_{indiv}(f_w,\phi_n)|}{|H_{dummy}(f_w,\phi_n)|}\right)^2} \tag{5.10}$$

where $|H_{indiv}(f_w,\phi_n)|$ and $|H_{dummy}(f_w,\phi_n)|$ denote the magnitude responses of the individual and dummy head HRTFs, W is the sample length of the HRTFs, N is the number of measured azimuth directions, f_w is the frequency, and ϕ_n is the source direction.

- Anthropometric measurements:

Two morphological dimensions were measured for each subject, the intertragus distance and the perimeter of the head. Measurements on real people with different head shapes make it necessary to establish some kind of reference points to achieve comparable and repeatable measures.

The intertragus distance describes the separation between the entrances of the ear canals. This measure was extracted from scaled pictures of each subject [206]. Figure 5.12 shows some pictures examples.

The perimeter of the head appears in many studies as a relevant anthropometric dimension for the ITD, but its definition is not always clear or practical. Some studies give loose head perimeter definitions [249, 251], while other have a too specific measure but not a practical approximation [237]. So, three different head perimeter measurements were done to explore a practical, repeatable and specific measure. They were labeled as *perim_head1* (through the highest point of the forehead and just above the ears), *perim_head2* (over the eyebrows and just above the ears) and *perim_head3* (over the eyes and the ears). See Figure 5.13.

Figure 5.12. Photographs of some subjects with scale reference for the extraction of the intertragus distance

Figure 5.13. Example of the three different head-perimeter measurements: *perim_head1*, *perim_head2*, *perim_head3*

5.2.3 ITD manipulation

Differences in the ITD are assumed to affect to azimuth localization. To investigate how these differences influence to each individual, a series of scaled ITD versions of the measured HRTFs were presented to them.

The original ITD of the two dummy heads and each subject were calculated from the processed and upsampled *quasi* HRIRs, with an onset detection method and a perceptual criterion, as said above. Then, scaled versions of these original ITD were calculated proportionally to -15%, -12%, -9%, -6%, -3%, 0%, 3%, 6%, 9%, 12%, 15% (which is the same as 0.85, 0.88, 0.91, 0.94, 0.97, 1, 1.03, 1.06, 1.09, 1.12, 1.15 scale factors) for the two dummy heads, and -6%, -3%, 0%, 3%, 6% (which is the same as 0.94, 0.97, 1, 1.03, 1.06 scale factors) for each individual HRTF. Figure 5.14 shows the scaled ITD variations of the dummy heads and an individual example.

The scaled ITD variations were applied as simple delay differences to the measured 96kHz upsampled BRIRs for each azimuth angle. The resulting BRIRs were then convolved with the minimum-phase HpTF compensation filter and downsampled to 48kHz. These BRIRs with modified ITDs were the ones used to generate the stimuli of the perceptual test. This procedure of modifying the ITD in BRIR has been tested in other studies [241] without any subject being able to reliably discriminate the reconstructed responses from the originals, in addition to showing robustness and results subjectively free of artifacts.

5.2.4 Test description

The objective of the perceptual test was to evaluate the individual subjective localization in relation to ITD variations. Scaled ITD variations of two dummy heads and own individual BRIRs were presented by headphones to each subject participant in the test. They were asked to locate virtual sound sources in the horizontal plane, with the aid of a real visual reference.

The different stimuli presented to the subjects depend on the following characteristics:

- Type of dummy: three BRIR sets were employed with each subject, from the measurements on the two dummy heads (Bruel 4100 and Neumann KU100) and from the own individual.
- ITD variations: the ITD of the BRIRs were modified as previously de-

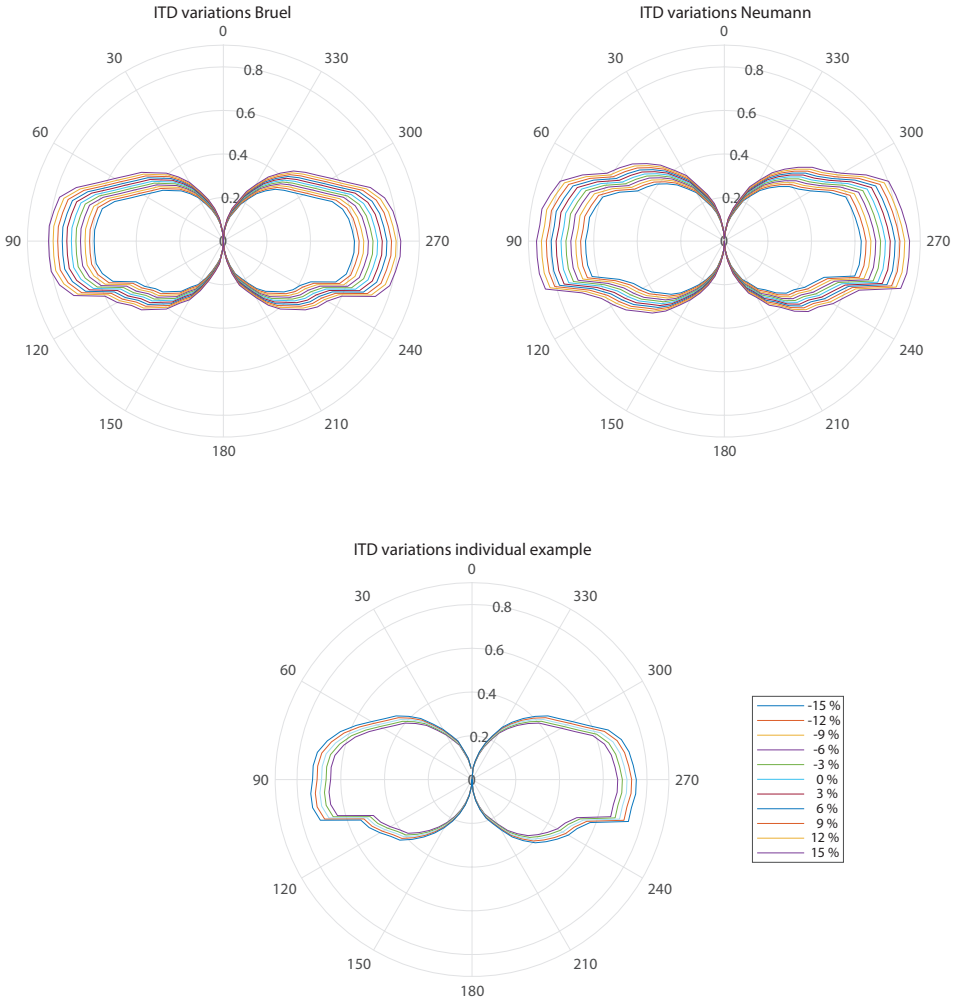


Figure 5.14. Scaled ITD of the two dummy heads and one individual example, presented in the perceptual test. Radial units are in ms

scribed, resulting in eleven scaled versions for the dummy heads and five scaled versions for the own individual BRIRs (see Figure 5.14). The ITD of the own individual was also inserted into the BRIRs of the two dummy heads, as another special case.
- Angles: To avoid a large number of angles that would extend the duration of the test, the subjects were assumed to have symmetrical perception on their left and right sides. Then, symmetrical angles to the medium plane were used and treated afterwards as a single position. In previous informal tests, this procedure was found to facilitate the natural use of the space by the subjects and to improve the position of the subjects with respect to the visual reference scale. Six different angles on the horizontal plane were chosen for the study: 0°, 20°, 40°, 60°, 75° and 90°. Considering the symmetrical criterion, these angles of study could randomly become 0°, 340°, 320°, 300°, 285° and 270° in reproduction.
- Type of sound: Three different excerpts of sound were used. Guitar, female voice and pink noise. The guitar sound (5 seconds) was chosen because it was composed by various impulsive sounds, while retaining enough bass content. On the other hand, the female voice (15 seconds) consisted partly of long sung vocalizations. These two excerpts simulate acoustic sounds that each subject had previously experienced. By contrast, the pink noise (4 seconds) was a broadband artificial sound.

Taking into account all the previous characteristics, the total number of stimuli presented to each subject was:
$(12\ ITD_{variations} \times 2\ BRIR_{dummy} + 5\ ITD_{variations} \times 1\ BRIR_{individual}) \times 6\ angles \times 3\ sounds = 522\ stimuli$

These stimuli with all characteristics combined were presented randomly to each subject.

A head-tracker was implemented to be used during the test. It was based on Arduino platform and the BNO055 sensor, according to [252]. A real time software reproduction responsive to the head-tracker, was also implemented for the test. They employ USB serial communication by UDP (User Datagram Protocol). The signals for all azimuth positions (each 5 degrees, 72 positions) were pre-rendered for each stimuli sound, then they were reproduced in the specific angle directions chosen for the test. The custom real time software reproduction crossfaded the renderings of each five degree angle position, showing a smooth and undetectable spatial interpolation. The performing of the test was done in the same room and

position of the measurement process. This made it possible to use as a visual reference the same set of loudspeakers used for the measurements. The head-tracking along with the coherence between the virtual sounds and the acoustic of the room, allowed excellent externalization, as has been previously demonstrated [253]. The perceptual effect was so good that all participants came to believe that at least some of the sounds were reproduced by the loudspeakers, many of them even believed that all the sounds were reproduced by the loudspeakers, although this detail was not the focus of study.

A simple GUI was also made for subjects to annotate the answered angles and control the perceptual test in a double-blind manner, Figure 5.15. Each stimulus was looped with a one-second pause until a response was chosen, and the play and stop controls were also available to the participant. To ensure that each subject understood the task and was familiar with the environment, the graphical user interface and the procedure, a brief training introduction was conducted.

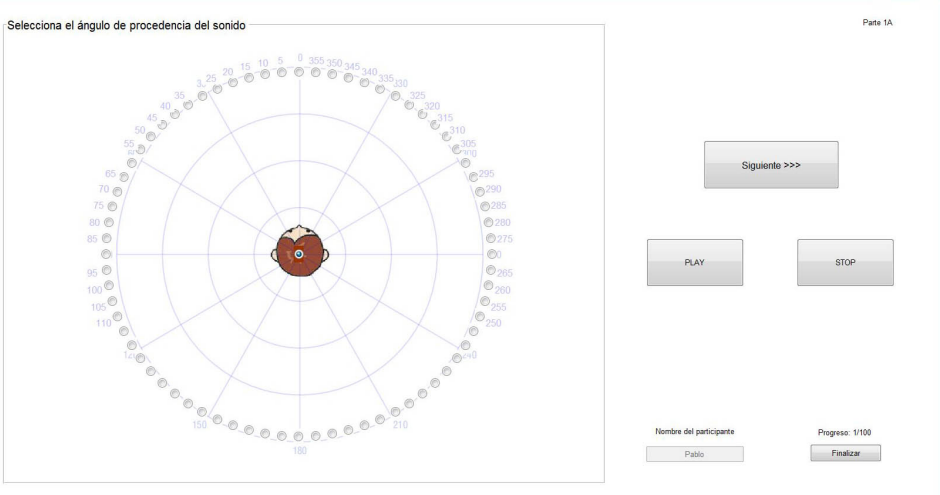


Figure 5.15. GUI for the perceptual test of ITD variations

The visual reference was the same loudspeaker array used for the measurement. Each speaker was labeled with the number of degrees as seen from the measurement and test position. All stimuli were judged looking to 0°. The same lasers used during the measurements were employed to as-

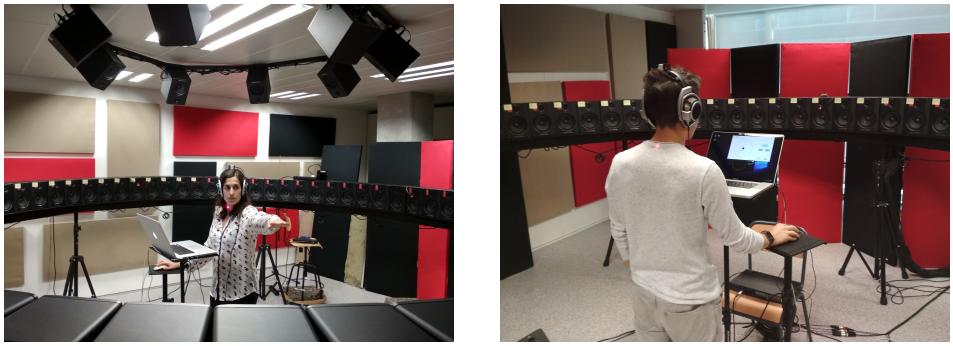


Figure 5.16. Subjects performing the ITD variations perceptual test

sure the correct reference position during the test. The head-tracker used for real-time reproduction was also employed to check the subjects face pointing direction, so the stimuli could be judged always looking to 0°.To avoid problems with angular positions outside the subject's field of vision, each participant was instructed to act as follows: If the stimulus appears to sound within the field of vision while looking at 0°, simply note the angle of the chosen (apparent) sound source. If the stimulus seems to sound out of the field of vision while looking at 0°, raise one arm pointing to the sound source and then turn your head to check the visual reference angle you are pointing at. Then, to avoid the possible mismatch between listening at the 0° angle and listening oriented towards the tested angle, the real time playback was limited to the ±10° arc around 0°, so that when a subject rotated the head beyond this limitation the playback was muted, returning when the subject faced an angle around 0° reference angle. This procedure made it possible to take advantage of the head-tracking reproduction and preserve the integrity of the perceptual task. Pictures of the performing can be seen in Figure 5.16.

Twenty-one people participated in the experiment, six women and fifteen men, ages 20 to 41 years (mean 29.23, median 27). The entire procedure included a pre-session to measure each individual and the actual perception test was conducted over other days. To ensure that results are not affected by hearing fatigue, the test was divided into four sessions of about twenty minutes, and performed in two different days. Each day consisted of two sessions separated by a fifteen-minute break.

5.2.5 Outlier responses treatment

Due to the difficult of the task, the long duration of the test and possible distractions of the subjects, an outlier treatment was performed over the answers of each subject to avoid extreme responses values that can spoil the statistical analysis. This intra-subject outlier treatment was applied to the mean value of the standard deviation of the answers, for each ITD variation presented for each dummy head. In this way, it is possible to detect an outlier response compared to others in the same conditions (dummy head and ITD variation).

The *generalized extreme Studentized deviate test* (generalized ESD) [254] was employed to detect the outliers. Outlier responses were detected in nine subjects, two of them responded with two extreme answers and the remaining eight with only one (Figure 5.17). The detected outlier responses were replaced with the mean of the responses that had the same characteristics (dummy head and ITD variation). This was done with the aim of minimally disturbing the statistics and further analysis.

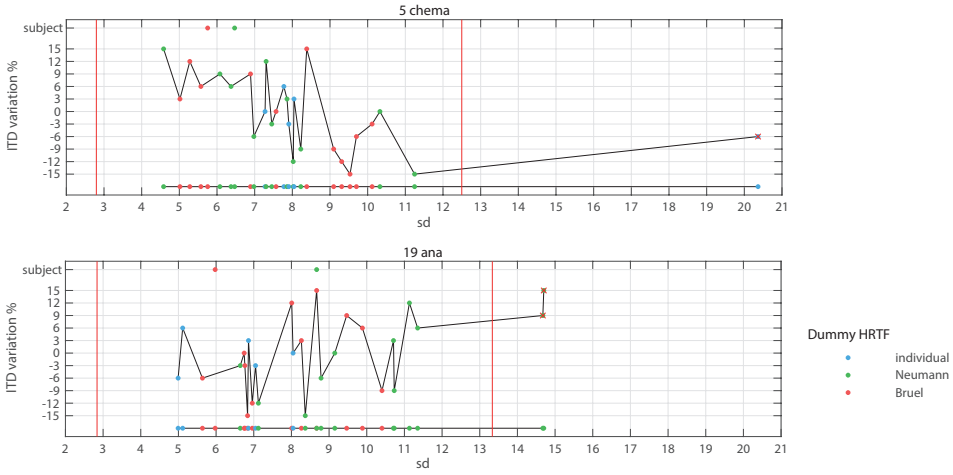


Figure 5.17. Examples of outliers detection in two subjects

5.2.6 Analysis of the results

Due to the individual characteristics of the HRTF and therefore of the ITD, simple analysis with subjects' aggregated results are not useful here. What a divergence error in the answered angles may mean for one subject, for another could have a different meaning. Differences between subjects should be maintained and taken into account during the analysis.

Besides, difficulties in the perceptual evaluation of HRTFs and the analysis of the data had been previously reported [255, 256]. These reveal that the interdependence of the perceptual cues and the effect of the learning process of the subjects during the perceptual test, make the object of study, that is, the perception of the subject, a characteristic not static but dynamic. The dynamic behaviour can affect to the degree of repeatability of the subject's answers [257] and includes the possible effect of super-normal cues [258].

Because of the individual and dynamic characteristics of the subjects, instead of discarding subjects with less accuracy in their responses, it is interesting to study the behaviour of the subjects according to their precision. Each subject have a different HRTF and each of them will perceive and locate the stimuli in different positions. The individual perception is always correct, regardless of the error in their answers or the difficulty of the proposed task. So, no inter-subject outlier treatment was done based on the error (accuracy) of their answers, but a classification of subjects based on the standard deviation (precision) of their response errors. This classification can explain different subjective characteristics: the reliability of the subject performing the test, their adaptation ability, and can also be influenced by the degree of difference of their own HRTFs with respect to the tested ones. The clustering also allows to compare the subjective responses of all the subjects together, despite the different individual perceptions. In the Figure 5.18, the clustering classification of subjects can be seen, as a function of the standard deviation of the error of their answers. Four different groups of people arise using a *k-means* clustering with Calinski-Harabasz evaluation criterion [259]. Group 1, with a lower mean standard deviation, have a very robust behaviour, while group 4, with higher mean standard deviation, shows a less reliable performance. Higher standard deviation values are expected to occur in subjects with greater differences between the tested HRTF and their own, which can also be understood as greater morphological differences. The possible effect of the learning pro-

cess was not studied, and an attempt was made to disperse its incidence as statistical noise by randomizing the angles presented to each subject. The clustering classification was included in the multivariable analysis as another variable, because it reflects a subjective behaviour that enables the different subjects to be directly related to each other.

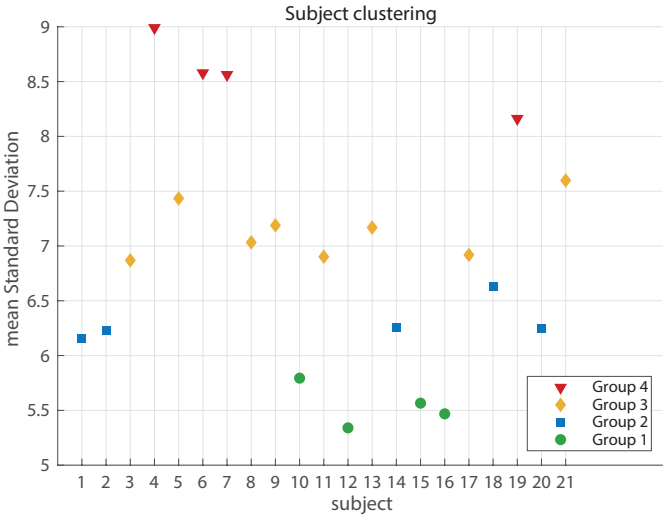


Figure 5.18. Clustering of the subjects based on the standard deviation of the error of their answers

Pearson correlation coefficients were calculated to study the linear relations of different variables and characteristics. The following variables were taken into account:
-*cluster*: cluster group classification of each subject participant in the test (Figure 5.18).
-*Std*: standard deviation of the error of the answers.
-*type_Dummy*: type of HRTF set, from dummy heads Brüel 4100, Neumann KU100 or own individual measurements.
-*ITD_variation*: percent scaled variation of the ITD (-15%, -12%, -9%, -6%, -3%, 0%, 3%, 6%, 9%, 12%, 15%).
-*MSE_varitd*: Mean Square Error of each scaled ITD variation (Figure 5.14) with each individual actual and measured ITD (Figure 5.10).
-*MSE_answers_varitd*: Mean Square Error between the answered and the target angles tested.
-*MSE_ild_500Hz*, *MSE_ild_2000Hz* and *MSE_ild_5000Hz*: Mean Square Er-

ror between the dummy heads (Brüel and Neumann) and each subject measured individual ILDs (Figure 5.11), for the bands of one octave centered in 500Hz, 2000Hz and 5000Hz.
-*SpectDist*: Spectral Distortion between the dummy heads and each individual HRTFs (Equation 5.10).
-*perim_head1*, *perim_head2* and *perim_head3*: perimeter of each individual's head, three different measures (Figure 5.13).
-*intertragus_distance*: intertragus distance of each individual's head.
-*age*: Age of each participant.

Table 5.1 gathers Pearson's correlation coefficient of the *cluster* variable with all the others. *Std* and *MSE_answers_varitd* have the expected strongest correlations as the *cluster* classification depends on those variables. It is interesting to note that *type_Dummy* and *SpectDist* variables show a weak relation with *cluster*, pointing that the effect of the HRTF set is lower than other characteristics. *ITD_variation* also gives a weak correlation with *cluster* while *MSE_varitd* have a stronger correlation, suggesting that the perceptual effect of the variation of the ITD is not linearly based (*ITD_variation* contains the linear percentage values used to scale the ITD, while *MSE_varitd* reflects the difference between the actual scaled ITD and the original individual one). Attending to the ILD variables, we can see that the higher frequency bands have more correlation with *cluster* than the lower one, especially the band of mid frequencies, *MSE_ild_2000Hz*. This can be explained as ILD being one of the dominant perceptual cues of HRTF sets, especially for higher frequencies. The predominance of the ITD in the lower frequencies and of the ILD in the higher frequencies is a perceptual mechanism that has been previously described. [229]. *Age* also shows a strong relation with the performance of the subjects, maybe for possible age-related hearing variations or more probably due to a connection of age with the comfort and ease of the subjects when performing the test. One of the remarkable relationships that arise in this analysis is the high correlation with the anthropometric measurements, perimeter of the head and intertragus distance. Of the three head perimeter measurements, *perim_head2* gives a higher correlation value, proving to be the most robust and stable measurement.

The Pearson correlation coefficients of the *MSE_answers_varitd* variable can be seen in Table 5.2. In general, high correlation values are shown

Pearson correlations for ***cluster*** variable		
Std	0.540	strong
type_Dummy	0.000	weak
ITD_variation	0.000	weak
MSE_varitd	0.119	strong
MSE_answers_varitd	0.430	strong
MSE_ild_500Hz	0.059	weak
MSE_ild_2000Hz	0.345	strong
MSE_ild_5000Hz	0.215	strong
SpectDist	-0.008	weak
perim_head1	-0.341	strong
perim_head2	-0.442	strong
perim_head3	-0.249	strong
intertragus_distance	-0.301	strong
age	0.307	strong

Table 5.1. Pearson correlation coefficients between *cluster* and the rest of variables

in this table, which is a noticeable difference from Table 5.1. These high values show a more sparse relationship of the *MSE_answers_varitd* with most variables. This observation should be taken with caution, because as previously said, the error or divergence between the answers of the different subjects does not have to mean the same thing. Therefore it is not entirely optimal to directly relate all subjects together with this variable. However, this analysis helps to understand already detected mechanisms (difficulty of the analysis due to the influence of dynamic subjects and individual characteristics) while confirming some of the detected dependencies with the other correlation Table 5.1. Then, extreme values are the most interesting to examine: variable *age* shows here a very weak correlation with *MSE_answers_varitd* (as opposed to the *cluster* correlation), suggesting that the accuracy of the subjects' responses is not related to their age, but perhaps to their precision. *ITD_variation* is again a weak relation, reinforcing the idea of non-linearity in the perception of ITD variations, and *MSE_ild_2000Hz* and *perim_head2* are again observed as the most influencing variables.

Summarizing the essential, in the evaluation of the perception of scaled variations of ITD an additional dependence is observed with the ILD (especially with the one octave band around 2000Hz) and a strong relation appears with the anthropometric parameters perimeter of the head (*perim_head2*) and intertragus distance.

Pearson correlations for ***MSE_answers_varitd*** variable		
Std	0.860	strong
cluster	0.430	strong
type_Dummy	-0.146	strong
ITD_variation	-0.052	weak
MSE_varitd	0.217	strong
MSE_ild_500Hz	0.168	strong
MSE_ild_2000Hz	0.340	strong
MSE_ild_5000Hz	0.205	strong
SpectDist	0.153	strong
perim_head1	-0.175	strong
perim_head2	-0.226	strong
perim_head3	-0.169	strong
intertragus_distance	-0.166	strong
age	0.048	weak

Table 5.2. Pearson correlation coefficients between *MSE_answers_varitd* and the rest of variables

The important relationship of the subjects' perceptual answers with the anthropometric parameters *intertragus_distance* and *perim_head2* is quite significant as it directly relates subjective data to objective and easily observable measures. Furthermore, a decision tree classification was done for the *cluster* variable (to avoid overfitting, a pruning limitation was introduced to one third of the subjects elements-responses). The result depicted in Figure 5.19 shows that the clustering of subjects can be explained by using just the variables *perim_head2* and *intertragus_distance* together as predictors, confirming also the previous correlation values. It is interesting to point out that these findings are related to other studies [206, 241, 260, 261] in which similar anthropometric measures have also been identified as possible influential parameters of the perception of the ITD and also of the general HRTF.

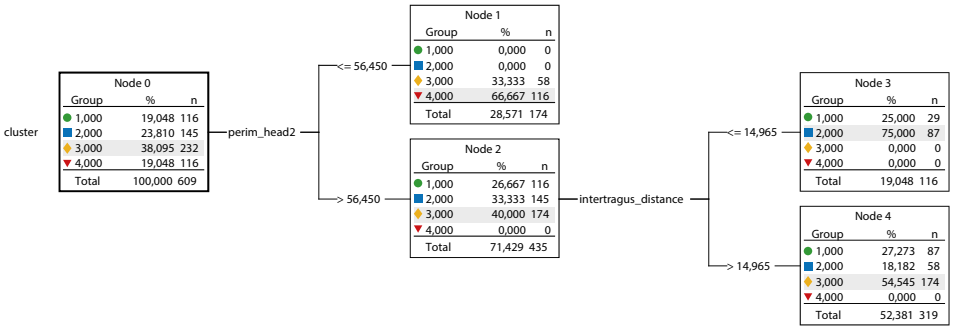


Figure 5.19. Decision tree that classifies the subjective perception of the subjects with the objective measurements *intertragus_distance* and *perim_head2*

5.2.7 Prediction of individual ITD scaling factor by polynomial equations

A key result of the experiment is the apparent randomness with which participants rated their own measurements. This behaviour have been previously observed in other studies. In [257] the repeatability and hence reliability (or lack thereof) of HRTF ratings is discussed, and in [256] they also found that individual measurements may not necessarily be the optimal when considering the more general requirements of good spatial audio reproduction beyond localization. These problems in the study of HRTF perception may be influenced by at least two reasons: super-normal cues may be acting for some people as a reinforcement for the location of some spatial positions [258], and the timbre variation of a different set of HRTF may be more pleasant for some people than the timbre of their own HRTF [256], or even enhance their listening ability as if it were a hearing-aid device. Besides, the effect of the adaptation and learning ability to listen with other HRTF [141], mixed with the previous perceptual phenomena, may produce more statistical noise and bias in the results of HRTF perception experiments.

Then, as in other previous experiments, we found that individualization of ITD is possible and desirable, but letting the subject self-adjust to its perceptual optimum is a difficult and time consuming task. Instead, it would be more convenient to provide with a generic prediction that could

fit or adjust to the individual's own HRTF. Following the approach of Lindau [241], we can try to predict individual scaling values for the ITD, to adapt the ITD of a particular HRTF set to the closest scaled values of each individual's ITD. Taking advantage of the individual measured and tested data of this experiment, we know that two anthropometric measurements (*intertragus_distance* and *perim_head2*) can be used as a predictors of the scale factor of the ITD. These can be employed to calculate a regression formula that will produce a practical individual scaling factor based on the perimeter of the head and the intertragus distance of the subject.

In the experiment, discrete scale factors were applied and evaluated. To improve the resolution of the individual scale factor, minimum error scaling factors have been calculated for each subject, based on a quadratic regression of the discrete scale factors and data of each individual. This procedure can be applied to the Root Mean Square Error of the answers of the subjects (subjective criterion - minimum localization degree errors) and also to the Root Mean Square Error between the scaled ITD values and the individual ITD (objective criterion - minimum ITD difference). It should be noted that the scaling of the ITD (and any other characteristic), will be different for each HRTF set to be adapted. Figure 5.20 shows the curves and minimums obtained with this calculation, both for the subjective and the objective criteria, and for the two dummy heads (different HRTF sets) employed in this experiment, the Brüel 4100 and Neumann KU100.

As can be seen in the Figure 5.20, the calculated minimums for some of the subjects were on the bounds of the tested range and a couple of regression curves are inverted. These subjects minimums were discarded as the data are not reliable (For dummy Brüel with subjective criterion 4 subjects, and with objective criterion 9 subjects. For dummy Neumann with subjective criterion 9 subjects, and with objective criterion 2 subjects).

Polynomial modeling of the scaling factors, in relation with the variables *intertragus_distance* and *perim_head2*, were calculated for the two different dummy heads employed in the experiment, Brüel 4100 and Neumann KU100. These regression formulas can be obtained both from the subjective responses or objective measurements minimum errors calculated before. Three-dimensional polynomial regressions formulas (scaling factor, *intertragus_distance* and *perim_head2*) of second order were obtained, in the form of the Equation 5.11

$$S_{ITD}(x,y) = p_{00} + p_{10}x + p_{01}y + p_{20}x^2 + p_{11}xy + p_{02}y^2 \qquad (5.11)$$

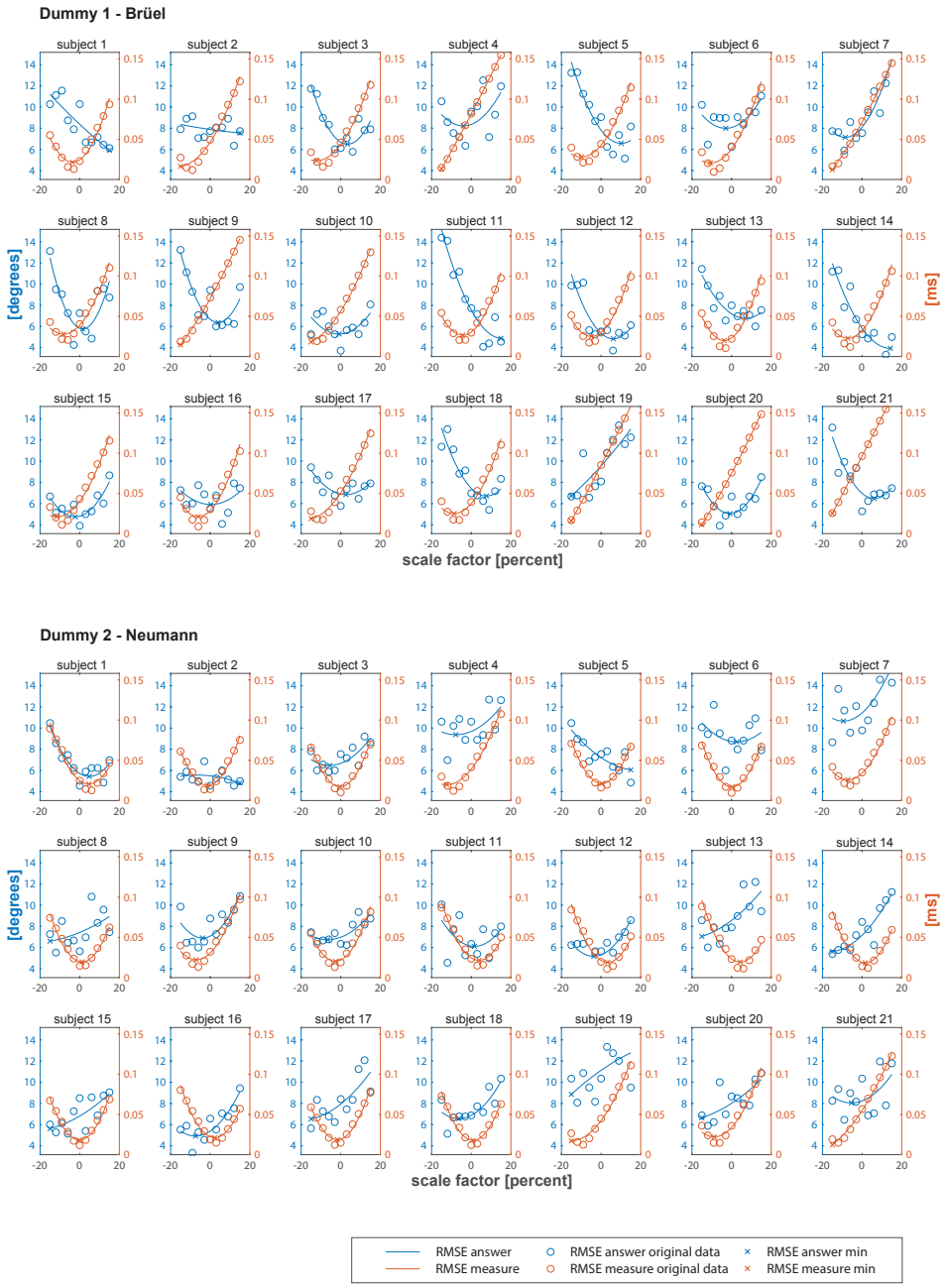


Figure 5.20. Minimum error scaling factors for the two dummy heads (1-Brüel, 2-Neumann) for all the subjects

where $S_{ITD}(x, y)$ is the scaling factor to apply to the ITD, p_{ij} are the computed coefficients for each dummy head, x is the intertragus distance and y is the perimeter of the head of the subject (measured over the eyebrows and just above of the ears), both in centimeters.

To improve the fitting of the polynomial modeling equations of the subjective criterion, a weighting factor was applied in function of the precision of the subject's responses, that is, with the inverse of the variance of the responses (Equation 5.12)

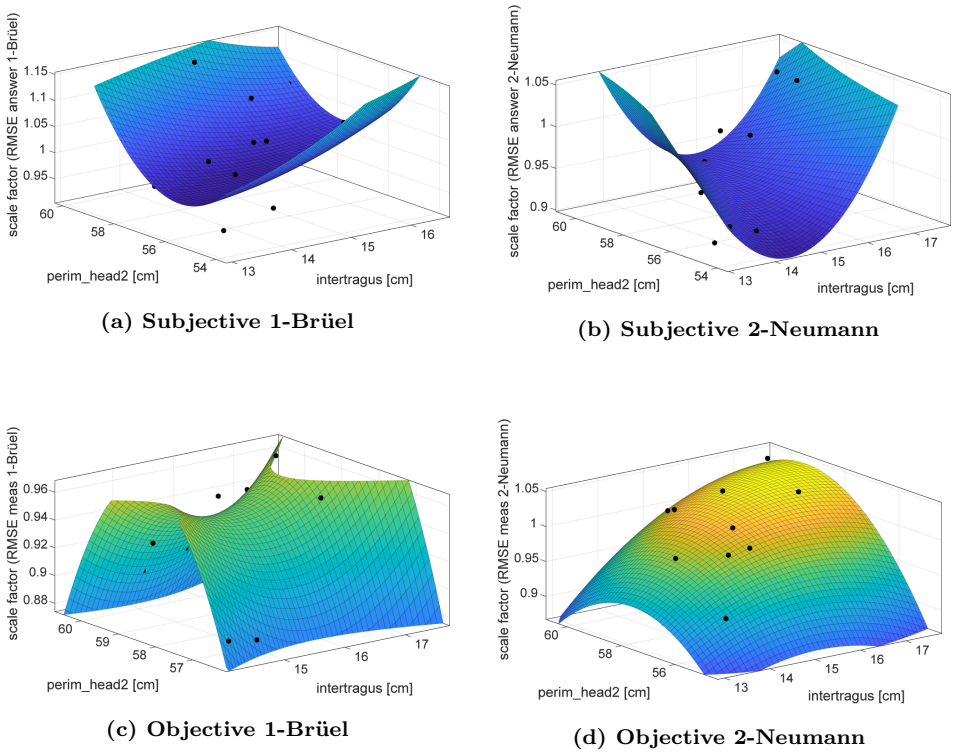
$$weightfactor_{subjective}(subject,dummy) = \frac{1}{(std_{mean}(subject,dummy))^2} \quad (5.12)$$

where std_{mean} is the average standard deviation of the answers of all the ITD variations tested, for each subject and dummy head. The normalized coefficients corresponding to the calculation are in Table 5.3, as well as the surfaces defined by these polynomials can be seen in Figure 5.21.

Based on	Answer (subjective)		Measures (objective)	
Dummy head	Bruel 4100	Neumann KU100	Bruel 4100	Neumann KU100
p_{00}	0.9943	0.9355	0.9507	1.013
p_{10}	-0.004915	0.007211	0.004937	0.02481
p_{01}	0.03921	0.006943	0.009847	0.02013
p_{20}	0.01184	0.05807	0.01509	-0.01046
p_{11}	-0.03599	-0.03844	0.01084	0.01212
p_{02}	0.04845	0.0002679	-0.04824	-0.02177
mean x	14.85	15.24	15.7	15.1
std x	1.061	1.43	0.952	1.184
mean y	57.36	57.25	58.4	57.73
std y	1.793	2.073	1.431	1.613
R-squared	0.6534	0.299	0.8036	0.8337

Table 5.3. Normalized coefficients of the polynomials, for the scaling of the ITD of two dummy heads

Looking at the values of the normalized coefficients we can see that the relative weight of the variables *intertragus_distance* (coefficients p_{10}, p_{20}) and *perim_head2* (coefficients p_{01}, p_{02}) is comparable in almost all cases. It can also be noticed that the objective criterion polynomials have a higher fitting with the data (R-squared above 0.8).



(a) Subjective 1-Brüel

(b) Subjective 2-Neumann

(c) Objective 1-Brüel

(d) Objective 2-Neumann

Figure 5.21. Surfaces defined by the polynomials that relate the ITD scaling factor with the intertragus distance and the perimeter of the head:

a) Subjective criterion for dummy 1-Brüel,
b) Subjective criterion for dummy 2-Neumann,
c) Objective criterion for dummy 1-Brüel,
d) Objective criterion for dummy 2-Neumann.

For the direct and practical calculation of the ITD scaling factor with the polynomial modeling Equation 5.11, the Table 5.4 is presented, which lists the direct coefficients without normalizing. With these coefficients the Equation 5.11 can be applied directly with the measured values of the intertragus and the perimeter of the head of any person, to obtain a scale factor for the HRTF set of the dummy heads Brüel 4100 and Neumann KU100, and thus adapt the ITD to the individual.

Based on	Answer (subjective)		Measures (objective)	
Dummy head	Bruel 4100	Neumann KU100	Bruel 4100	Neumann KU100
p_{00}	35.59	-3.848	-68.42	-24.07
p_{10}	0.7682	-0.1183	-0.9823	-0.1199
p_{01}	-1.425	0.1939	2.632	0.8826
p_{20}	0.01052	0.02841	0.01666	-0.007465
p_{11}	-0.01892	-0.01297	0.007952	0.006347
p_{02}	0.01506	6.237e-5	-0.02354	-0.008367
R-squared	0.6534	0.299	0.8036	0.8337

Table 5.4. Direct application (not normalized) coefficients of the polynomials, for the scaling of the ITD of two dummy heads

5.2.8 Conclusions and future work

An experiment has been developed and carried out to evaluate the perceptual effect of proportional scaled ITD inserted in different HRTF sets. Real HRTFs have been measured and anthropometric measurements performed to obtain objective data to be included in the analysis along with the subjective results. Two dummy heads' HRTFs have been tested with scaled variations of ITD as well as individual measured HRTFs.

The analysis of the data collected with the perceptual test has turned out to be complex, as other perceptual tests with HRTF describe in the recent literature. Most of the difficulties may be due to the fact that test subjects have a dynamic behaviour that depends on several factors: possible super-normal cues for the location of some spatial positions, influence of timbre preferences different from their own, non-linear interrelation of different perceptual variables, and the combined effect with their ability to

learn and adapt to other HRTFs.

The dispersion of the responses (the standard deviation of the error of the answers, which indicates the precision of each subject), has been found to have a significant relation with the anthropometric measurements of intertragus distance and perimeter of the head. In addition, this perimeter of the head has been defined in a specific and practical way, out of three different manners of measuring the head perimeter.

By relating these two anthropometric dimensions to the ITD scale factor that produces a minimum error for each subject, an individual ITD scale factor can be predicted for other subjects by polynomial regression, and only with their intertragus distance and head perimeter. This polynomial is specific to each set of HRTF to be adapted, and can be calculated from objective measurements or subjective responses of a group of subjects.

The polynomial equations for the individual ITD scale of two widely used dummy heads (the Brüel 4100 and the Neuman KU100) have been estimated and their coefficients are provided for practical use.

Future work

The results and findings of this experiment demand continuation with future work. A new perceptual test must be addressed to evaluate the estimated ITD scale factors with subjective and objective criteria, and to compare the results of both methods. If the objective criterion gives good results, a generalized method could be derived to adapt the ITD of any HRTF set by scaling it to any other person. Besides, more anthropometric measurements could be explored to increase the number of dimensions of the polynomials and use them to extend the scaling to three-dimensional ITD values.

Part II

Loudspeaker based systems

Distance perception comparison between WFS and VBAP

6

6.1 Introduction and motivation

Every sound field produces a spatial depth sensation. This feeling is responsible for the perspective perception of an acoustic scene. The same pattern is valid for an artificial acoustic scene. Depth is considered to be an essential attribute for spatial sound perception and the sensation of depth is highly related to the perception of distance to the sound source [20]. The successful creation of depth in a sound scene is a challenge for a spatial audio reproduction system.

For conditions where both listener and sound source are stationary, at least four possible acoustic distance cues have already been suggested to play a relevant role [1, 262]:

- Intensity: An acoustic point source in free field obeys an inverse-square law, losing 6 dB on doubling distance to the listener. It is the most important cue, but implies a previous knowledge of the source power.
- Direct-to-reverberant energy ratio: In environments with sound reflecting surfaces, the ratio of energy reaching a listener directly (with-

out contact with reflecting surfaces) to energy reaching the listener after reflecting surface contact (reverberant energy), decreases systematically with increases in source distance.

- Spectrum: For distances greater than 15 m the sound absorbing properties of air significantly modify the sound source spectrum, mainly high frequencies. Also, when the sound source is close to the listeners head, some low frequency increase may occur due to the curvature of the sound field.
- Binaural differences: Although the HRTF in far field the does not strictly change with the distance [1], slight and natural movements of the listener can modify the interaural differences allowing distance perception in the form of acoustic parallax.

In the field of music production and cinema, direct-to-reverberant energy ratio as well as distance attenuation have been employed extensively as the main distance cues with acceptable results. These techniques have been commonly applied to stereo and multichannel surround (5.1) mixes. However, with the introduction of other advanced spatial audio systems such as Wave Field Synthesis, new possibilities have emerged.

With Wave Field Synthesis (WFS), it is possible to synthesize within the whole listening area the correct curvature of the wavefront arriving to the listener from a source located at a certain distance. Although distance perception has been said to be difficult to perceive with WFS [263], other studies [264] suggest the opposite. This chapter is intended to provide deeper insight into this point by means of a perceptual listening test aimed at comparing the performance of WFS and amplitude panning systems such as VBAP with respect to their capabilities to recreate realistic distance perception cues.

An introduction and brief description of the sound systems compared in this experiment can be found in Chapter 2 *Background*, for WFS in Section 2.5 *Wave-Field Synthesis principles*, and for Vector Base Amplitude Panning (VBAP) in Section 2.3.3 *Vector Base Amplitude Panning VBAP*.

Despite the fact that the curvature of a wavefront has a more relevant effect in the HRTF at short distances to the sound source, slight movements of the listener might provide some parallax information that they may use to extrapolate distance even when the source is in the far field. In this context, it is considered that the absolute spatial resolution in azimuth is

about 5°, but a relative resolution is about just 1° or less, which is enough to notice a difference between two source directions [1].

The objective of this work is to find out if WFS provides better distance perception than other general approaches based on the phantom effect through amplitude panning. VBAP [63] was selected because it provides analytic equations for the multiple loudspeaker case. In the experiments, the listeners were seated, but they were able to move slightly their heads in order to be sensitive to some hypothetical parallax effects. The influence of different factors in the perception was studied: type of sound, listening angle and reverberation at different distances.

6.2 Test description

In order to evaluate the perception of sound distance by means of WFS and VBAP, a direct scaling by interval test [18] was selected. The test was based on a comparison between the sounds synthesized by WFS and VBAP for a virtual source at different distances from the listener, having a real sound source placed at a fixed distance.

Figure 6.1 shows a scheme of the set-up employed in the experiments. An octagonal array of 64 loudspeaker units separated 18 cm apart was deployed around the listener to reproduce WFS and VBAP. Additionally, a reference loudspeaker was placed at a distance of 4.9 m. Figure 6.2 shows a panoramic of the room with the WFS arrays used in the experiments. For testing VBAP, only some loudspeakers of the array were used, marked in grey in Figure 6.3. To avoid possible influences by visual cues, an acoustically transparent curtain was placed in front of the listener. The test was performed in a dedicated and acoustically treated listening room [265] with a T60 at $1kHz < 0.25s$ with a volume of 96 m^3, (Figure 6.1).

The standards and recommendations [135, 265, 266] related to subjective evaluation of sound where fulfilled in the experiments.

Different effects and factors were taken into account in during the test:

– Seven distances (synthesized source positions) were used to compare with the reference position: three ahead and three behind the reference distance at 1.74 - 2.46 - 3.47 - 4.9 - 6.92 - 9.77 - 13.81 meters, with 4.9 meters as reference position, (see Figure 6.1).

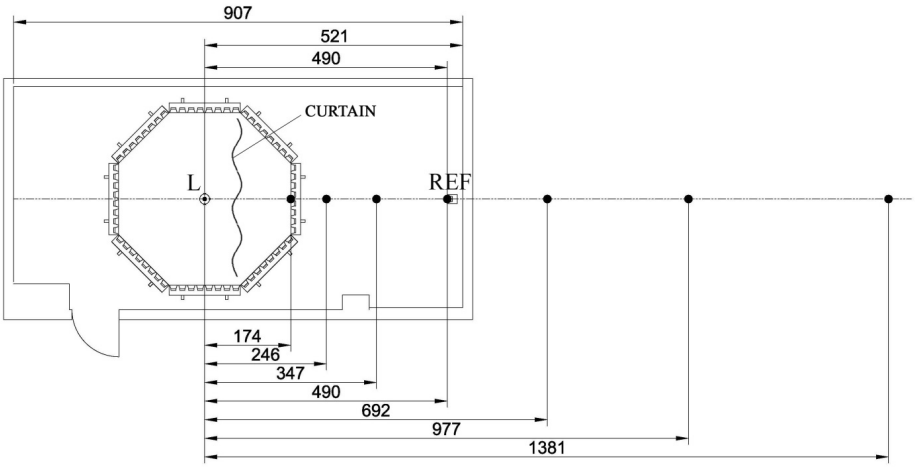


Figure 6.1. Speaker arrays and reference speaker (REF) set-up with synthesized distances (cm)

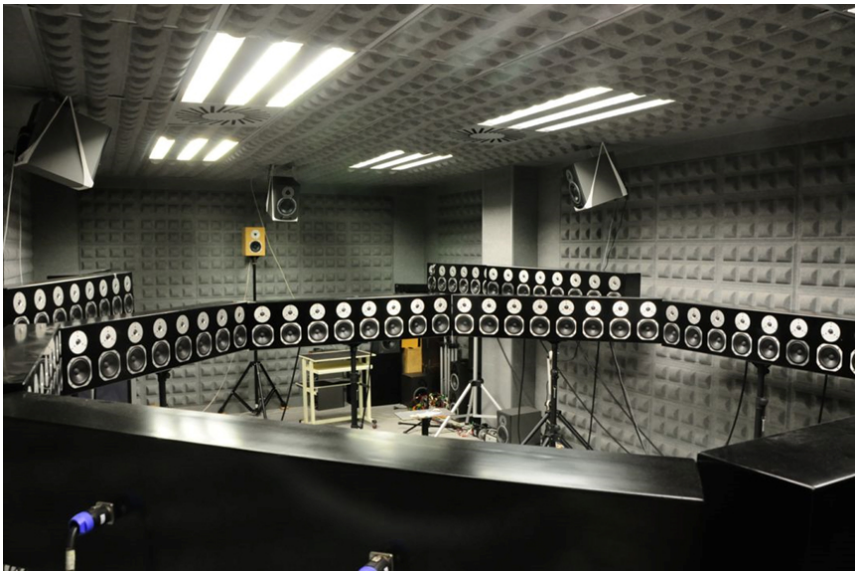


Figure 6.2. Wave-Field Synthesis set-up, where the octagon with the 64 loudspeakers can be appreciated and also the reference loudspeaker at the back in brown color

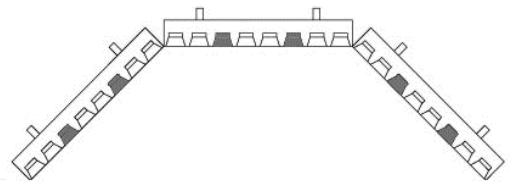


Figure 6.3. Array detail. The dark speakers are the ones employed for VBAP

– Four different types of sounds were considered: pink noise, speech, guitar and door closing. These sounds were interesting for localization according to their different spectral and temporal features.
– Additionally, the effect of synthetic early echoes was also analyzed. The same stimuli were presented to the listeners with and without echoes. The added reverberation time was approximately 20 ms, generated by means of four additional first order reflection sources from four virtual walls, calculated by the image-source method [267].
– Finally, two listening angles were studied, 0° and 90° azimuth. Note that 90° azimuth was chosen because it produces maximum interaural time differences.

Combining all the factors to be studied, a total of 224 stimuli per participant were needed. Besides these, 8 hidden references for each listening angle were added, resulting in a total of 240. All these stimuli were randomly presented.

The set-up was calibrated and equalized in order to match the frequency response of the reference speaker with the WFS and VBAP reproduction. The sound pressure level for pink noise at the reference position was 69 dBA.

A total of 25 people participated in the test, 15 male and 10 female all of them with normal audition with ages from 24 to 35. To perform the test, a graphic interface was developed, presented in a computer screen. This interface was easily controlled by the participant by means of a videogame joystick, Figure 6.4. The participants were seated in a high stool to have their ears at the same elevation than the loudspeakers. The acoustic curtain hides the reference loudspeaker and the frontal array loudspeakers to avoid visual cues or influences, Figure 6.5.

Figure 6.4. Graphic user interface of the perceptual test

Figure 6.5. Participant seated in the listening position ready to start the test. The laptop computer with the GUI, the joystick and the acoustic curtain can be appreciated

Before performing the test, some previous information was given to each participant. Moreover, the participants took part in a training phase before the test. They were able to listen to the different types of sound (pink noise, voice, guitar, and door) and also to the distance limits provided by the farthest and closest distances. Since the scale presented to the subjects was completely subjective, this experience allowed them to get an idea of the total range in which the distances would fall.

The test was performed in two phases. First, the subjects evaluated the stimuli for the 0° angle. Then, in the second part they proceeded with the 90° stimuli. The average execution time of each part was 25 minutes with a 15 minutes break between the two parts. The test procedure was quite simple; the two sounds to assess were presented one after another, followed by the reference sound. For each test signal, the subjects evaluated the perceived distance over the subjective scale with respect to the reference sound, being able to listen to both several times.

6.3 Analysis of the results

Reliability

A Cronbach's alpha was calculated with all the participants' answers. The alpha reliability of the distance answers is $\alpha = 0.992$ (for $N = 25$ participants) which indicates a very high reliability. Besides, the analysis of the responses about the included hidden references confirmed a high degree of consistency in the responses.

Aggregated results

The mean of all the perceived distances (answers) for all the stimuli and participants and classified for each system is showed in Figure 6.6. As illustrated by this figure, both systems have coherent results. The graph also shows that both systems tend to perceive closer the synthesized sources. WFS seems to approximate better the ideal behavior (diagonal).

Figure 6.7 (a) illustrates the divergence (perceived error) of WFS and VBAP for the reference position (4.9 m) and also for the hidden references. It is observed that the hidden reference has a very small deviation, indicating a nearly ideal behavior. In contrast, WFS and VBAP show a greater

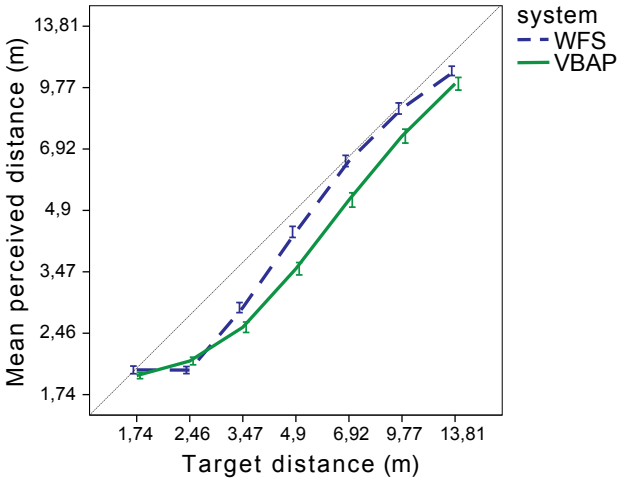


Figure 6.6. Mean ($N = 25$) of the perceived distances (m) for each system (WFS and VBAP). (95% CI)

divergence. However WFS presents a better error average than VBAP, having a statistically significant difference, as shown by the 95% confidence intervals (95% CI).

An overall comparison for all positions is showed in Figure 6.7 (b), where the total divergences of the aggregated perceived distances for WFS and VBAP are represented. In these graphs, the general behavior of each system can be compared, indicating that WFS provides lower divergence.

To study the influence of the system in the test, a paired samples t-test was performed. It showed that the system (WFS vs VBAP) has a highly significant influence in the perceived distances ($t = 16$, $df = 2799$, $p < 0.001$).

Finally, the divergence for each distance is showed in Figure 6.8, where a similar distribution is observed. Note that WFS provides better performance than VBAP, especially for distances located around the reference and behind it.

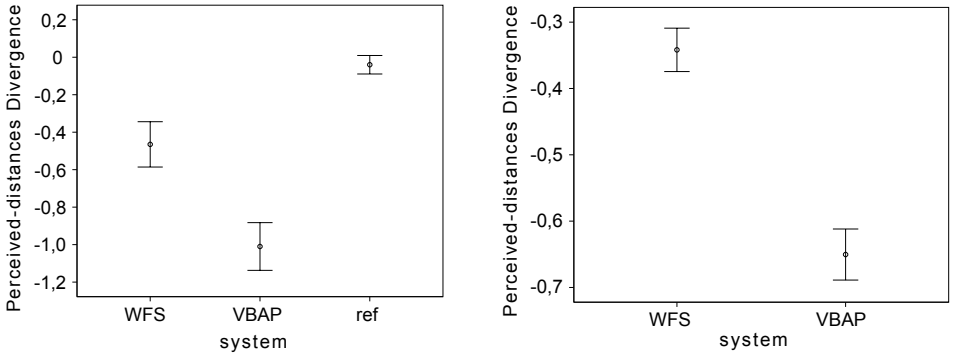


Figure 6.7. Divergence of the perceived distances (m) for the type of system: **(a)** at reference position (4.9 m) for WFS, VBAP and hidden references, **(b)** of all the distances for WFS and VBAP. (95% CI)

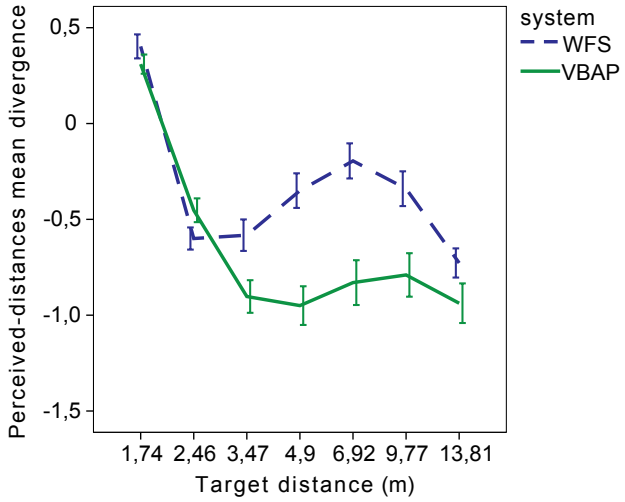


Figure 6.8. Mean divergence of each perceived distance (m) *versus* each target distance (m), for WFS and VBAP. (95% CI)

Influence of the early echoes

The added four early reflections from 4 virtual walls introduce no noticeable effect, as shown in Figure 6.9. To measure this effect an analysis of variance (ANOVA) was performed to yield that the influence of the added reverberation was not significant ($p = 0.388$).

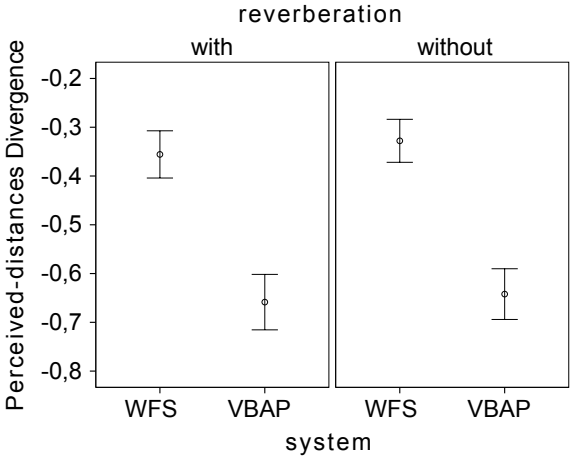


Figure 6.9. Divergence for each of the perceived distances (m) for WFS and VBAP considering reverberation. (95% CI)

Influence of the type of sound

Figure 6.10 shows that the guitar sound provides the best results, followed by the closing door sound. The voice sound is the next obtaining more correct assessments, followed closely by pink noise. It is worth to note the differences in distance divergence for pink noise and voice according to the system, where WFS outperforms VBAP. A one-way ANOVA was performed, founding that the type of sound has a significant influence ($F = 131.62$, $df = 3$, $p < 0.001$). There is also a cross relation between the type of sound and the system, with a very high significance ($F = 8.6$, $df = 3$, $p < 0.001$), corresponding to better results of WFS in general, and especially for pink noise and voice.

Influence of the listening angle

The test was performed at 0° and 90° by rotating the listener, with the same number of stimuli at each listening angle. Figure 6.11 shows that better

results are obtained for frontal listening. A one-way ANOVA shows that the listening angle has a very high significance ($F = 56.08$, $df = 1$, $p < 0.001$), but its cross relation with the system is not significant ($F = 2.55$, $df = 1$, $p = 0.110$).

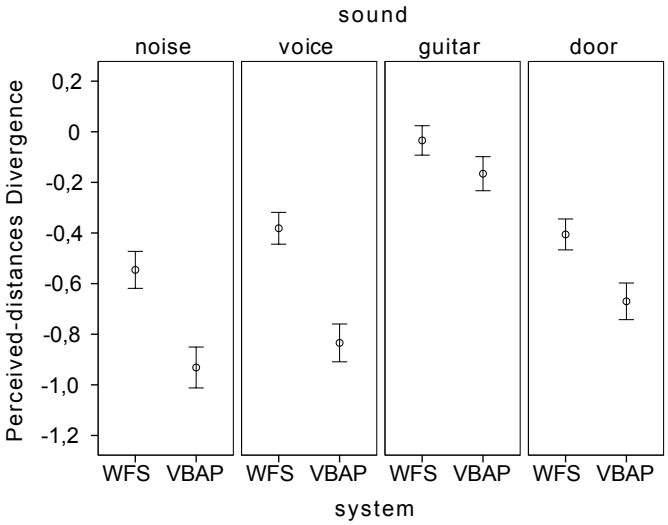


Figure 6.10. Divergence of the perceived distances (m) for WFS and VBAP, according to the type of sound. (95% CI)

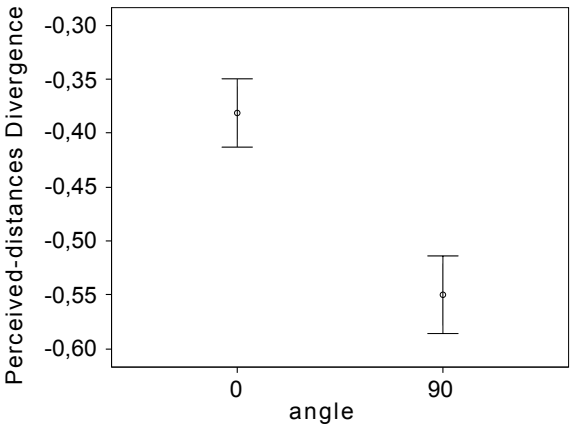


Figure 6.11. Divergence of the perceived distances (m) for the listening angles at 0° and 90°. (95% CI)

Influence of the participant

A very high significant cross-relation between the participant in the test (the listener) and the system can be seen with a one-way ANOVA ($F = 2.33$, $df = 24$, $p < 0.001$). In a graphic representation of the divergence of the perceived distances for each participant, considering separately WFS and VBAP, the difference between both systems stands out. Some of the participants (2, 9, 19) have good results with WFS and poor results with VBAP, as observed in Figure 6.12.

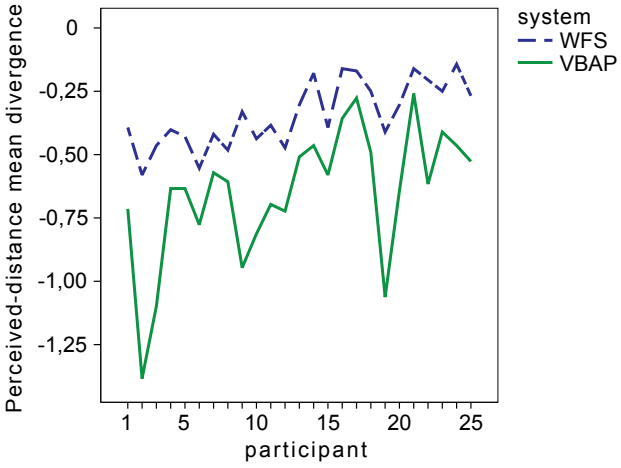


Figure 6.12. Mean divergence of the perceived distances (m) for each participant, considering the system, WFS or VBAP

6.4 Conclusions

According to the subjective perceptual test carried out, some conclusions can be listed:

- Both spatial sound systems (WFS and VBAP) have been shown capable of simulating a certain sound sense of distance based on the attenuation caused by distance.
- The type of sound is a determining factor in the perception of distance, getting better results for impulsive sounds. Moreover, WFS is capable of reproducing better sound distance than VBAP for other

sounds that are not impulsive.

- The listening angle has a high influence in sound distance perception, but its relation with the system (WFS or VBAP) is not determinant.
- The first order reflections did not provide a substantial improvement in distance perception. More experiments are needed to evaluate to what extent the introduction of multiple order reflections enhance this feeling.
- From the test results, it can be concluded that WFS has been shown to have a better overall capability to reproduce sound distance than VBAP, at least for sources placed in front of the listener.

Despite the pressure level is the main factor responsible for distance perception; this test has concluded that, at least for this set-up, WFS is better at producing distance perception cues than VBAP, as confirmed by the statistical analysis of the results.

Perceptual spatial acuity of spectrally divided sound sources

7

7.1 Introduction and motivation

Spatial sound systems with loudspeakers have evolved with a natural tendency of increase the number of speakers. Two main speaker-based approaches are used to reproduce immersive sound: amplitude panning and sound field synthesis systems. The first group includes multichannel surround systems (5.1, 6.1, 7.1 10.2...) [56], where the increasing number of speakers generally results in better quality and localization. The second group tries to synthesize a realistic sound field based on the physical equations of sound propagation, with Higher Order Ambisonics (HOA) [43] and Wave-Field Synthesis (WFS) [46] techniques as the main exponents. As in the first group, the more loudspeakers used, the better the synthesis obtained. In the case of WFS, this is because the spatial aliasing is reduced if loudspeaker density is increased, and in the case of HOA, it is because a higher order raises the minimum amount of loudspeakers needed. An introduction to these two families of methods can be found in Chapter 2 *Background*, 2.2 *Spatial sound systems classification.*

In both cases, the premise "the more, the better" (loudspeakers) is applied, but this approach obviously increases the cost and complexity of

the systems. It would, then, be interesting to have more available studies to determine what number of loudspeakers should be adequate to achieve a quality experience in each system, as well as to define the minimum number of loudspeakers to preserve this. In response to this problem, this chapter explores the feasibility of an alternative approach to reducing the number of loudspeakers in dense loudspeaker set-ups. The high frequencies are reproduced from a single loudspeaker, with a different direction from that of its panned counterpart low frequency. This approach may reduce coloration artifacts, but might, in turn, lead to misplacement or degraded quality of the sound source. Therefore, listening tests are conducted to investigate the location and quality of the source in the case where high frequencies are reproduced from a different direction than low frequencies.

Before addressing the research aspects, some previous considerations are presented. The inherent coloration effect of multi-speaker reproduction is explained, along with a review of a previous study related to the perception of simultaneous sources from different directions. The motivation and hypothesis of this work is then presented. In the Section 7.2 the approach of the experiment is explained, followed by the description of the three perceptual tests performed (Sections 7.3, 7.4 and 7.5) with the results obtained. Finally, the Section 7.6 summarizes the conclusions of the experiment and outlines some possible applications and future work.

7.1.1 Coloration effects in loudspeaker reproduction

Some descriptions and details of amplitude panning and WFS techhniques can be found in Chapter 2 *Background*, in Sections 2.3 *Panning with loudspeakers* and 2.5 *Wave-Field Synthesis principles*.

One of the negative aspects of panning systems and other systems that use multiple speakers to recreate the sound field is the alteration of the spectrum of sources (coloration), especially at high frequencies. This effect is due to the sum of signals out of phase with different speakers. For example, although the listener of a stereo system aims to be centered on the axis between the two speakers symmetrically, the distance from each speaker to each ear is slightly different. This difference of acoustic paths causes the sum of the signals from the two speakers to be out of phase. Therefore, a pronounced effect of comb filtering occurs beyond a certain frequency. These frequencies are above 1.5 kHz where the interaural distance begins to be comparable to its wavelength.

The coloration can be detected by the human ear. However, these effects have not been subjectively studied in depth, despite the existence of some works on the subject [268]. Depending on the reproduction system and the musical material employed, the effects may be more or less perceived by the listener. In any case, and regardless of the severity of these effects, it would be desirable to minimize or even remove them from the reproduction system.

The origin of the problem is the emission of high frequencies from two or more different points, therefore a direct solution would be not use panning techniques with conflicting frequencies and emit them solely from a single speaker. Obviously, this would result in a restriction of putting the virtual sources only in the places where we had a speaker, limiting the spatialization sound at first. This solution does not apply in the case of stereo playback set-ups (two speakers) or 5-channel surround systems, where the separation between speakers is very large [52]. However, in systems that use many distributed speakers, such as WFS [46], HOA [269] or VBAP systems with many channels, a solution of this kind might be approached. There would still, of course, be an error equivalent to the separation in the positioning of the speakers, but it would be less than in systems with a small number of speakers.

Given that the coloration effect is very small or negligible at lower frequencies, we might consider systems which would split the reproduction of low and high frequencies. Thus, low frequencies would be reproduced using panning or soundfield synthesis techniques (VBAP, HOA or WFS) and high frequencies, where the effects of coloration appear, would be reproduced from a single speaker. In this manner, some discrepancy could occur in the perception of the direction of arrival of the frequencies of the source, with the low and high frequencies coming from different positions. However, the hypothesis of this work is based on the premise that if the difference is not very large, the sound source can be perceived as coming from a single point instead of two different ones.

7.1.2 Concurrent Minimal Audible Angle (CMAA)

There are few studies related to the resolution of the human ear to perceive two different but simultaneous sources over time from different directions. One of the most interesting studies on this "spatial acuity" was conducted by Perrott [270]. There was a previous concept of *minimum audible angle*

(MAA) [271] used as a basis for his study, that identifies the ability of human hearing to detect an angular difference of the same source from two different directions when played sequentially. In his study, Perrott uses simultaneous acoustic events with the hypothesis that the new *minimal audible angle* for *concurrent* events (CMAA) is likely to be greater than the MAA because of the added difficulty to the human auditory system of having to separate sounds.

In the experiment, two speakers playing different signals were separated at different angular distances. The signals were two tones of different frequencies and the subject had to guess in each case whether the higher frequency tone was on the right or left. The tones were reproduced randomly so that 50% of correct answers for a certain angle meant that the listener could not distinguish where each tone was. The smallest angle at which this phenomenon began to occur was called CMAA. Another parameter studied was the frequency separation between the tones. The closer the frequencies were, the more difficult it was to distinguish them. The minimum angle is dependent on the frequency difference between the tested tones. For frontal listening this angle is around 10 degrees and can decrease to 5 if the frequency difference is small. The position of the sound sources with respect to the median plane was also found to affect the perception of the CMAA. As expected, the angle was bigger in lateral positions than in frontal ones. At a lateral azimuth position of 40 degrees the minimum angle raises to 16 degrees and up to 40 degrees with a lateral source position of 67 degrees.

7.1.3 Work motivation

The previously described work suggests the possibility that sounds emitted from different positions with little separation in space, can be perceived by the human auditory system as coming from one single point. The signals employed in that study were two tones, but what would happen if more complex signals were used? for example, two people speaking from different points.

On the other hand, it might be interesting to study the effects of separating low and high frequencies at a certain angular distance. Not just two tones of different pitch but low and high frequency bands of a single sound. This could have an application in spatial sound systems where several speakers are usually employed. Besides, if the signals were related in some way, harmonically for example, we could hypothesize that distin-

guishing the direction of arrival of each sound would be even more difficult for human hearing. Therefore, a new term, *just noticeable band splitting angle* (JNBSA), can be defined and introduced. It represents the minimum angle of separation between the high and low frequencies from which the listener begins to perceive artifacts (Direction of Arrival (DOA) error, source width), in the reproduction of a sound source by means of loudspeakers.

Using a playback system of this type, it would be possible to emit high and low frequencies from different points without any detection of difference by the listener, who would perceive the sound as coming from a point source. As commented above, the sum of the sound signals reproduced by different speakers produces coloration effects at high frequencies. If we could use such a reproduction system in which the low frequencies were reproduced by more than one speaker (using VBAP, WFS or HOA techniques) and the high frequencies were reproduced by a single speaker in a different position, we can hypothesize that the listener could perceive them as coming from one single point. If we can confirm this fact, it would open new paths to solve, or at least to improve, spatial sound reproduction systems.

7.2 Experimental approach

To conduct the experiment, a set of speakers were arranged in a circular arc shape and angularly spaced at equal intervals. The array was made with 19 self-powered Genelec speakers with virtually flat response, with a 5" woofer and a 1" tweeter. Each of them was placed separated 5 degrees in an arc of 2 meters of radius, between $-45°$ and $45°$ on either side of the listener (Figure 7.1). To evaluate the lateral sound and from behind, the listeners were rotated in the seat. Sound sources from all directions of the horizontal plane can then be evaluated but without the reflections that a complete circular array would produce.

Three signals were used in the experiment: brown noise, voice and trombone sound. These signals were split into two frequency bands, below and above 1.5 kHz. Third-order Butterworth low-pass and high-pass filters were used to make the division. This pair of filters sum correctly in amplitude and phase without producing any coloration in the transition band.

The speakers were placed in an acoustically equipped room to meet the

criteria mentioned in ITU BS.1116-3 [135] for this type of sound reproduction experiments. Therefore, the acoustic conditions of reverberation time, background noise and setup arrangement were within the criteria of this standard recommendation.

The different tests making up the experiment were performed by 9 to 11 subjects (two women and the remaining men), between 22 and 42 years old. All of them were skilled researchers with experience in critical listening, and performed the listening tests in a double-blind manner.

The perceptual location of sound sources can be strongly influenced both by the position of the listener with respect to the source and by the movement of the listener and/or sources [1, 272, 273]. In this work the perception of static sound sources was studied. For this reason, the subjects who carried out the perceptual tests were instructed not to move during the execution of the tests, and in addition, this was controlled by a supervisor by means of visual inspection.

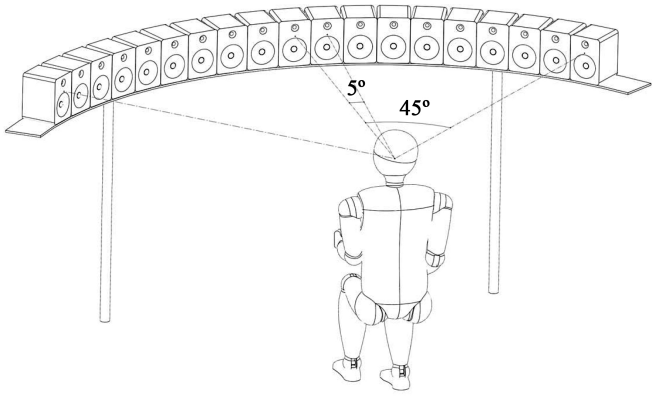


Figure 7.1. Loudspeaker set-up with 5° angular spacing

7.3 Test 1. Left/Right distinction

Test description

Following Perrott's studies [270], the aim of this experiment was to find out whether listeners are able to locate where the high frequency is, when the

reproduction of the low and high frequencies of a single sound source are separated from different angles.

The task proposed to the participants was to locate the high frequency and the question presented to them was: "Where is the high frequency?" with two possible answers: left or right. A graphical interface was developed to perform the test, as shown in Figure 7.2, which subjects used to record their responses to each stimulus. They were presented with their low and high frequency separated with 6 offset angles: 10, 15, 20, 25, 30 and 40 degrees. The setup allowed 5 degrees of separation, but in a preliminary test it was very obvious to the researchers that it was impossible to distinguish, so the 5 degree separation was omitted to reduce the duration of the listening test.

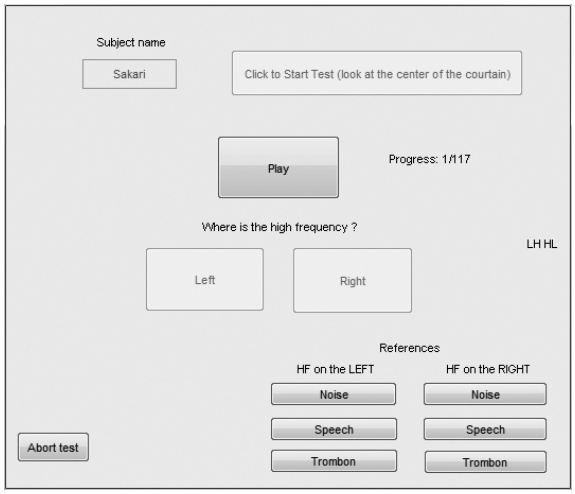


Figure 7.2. Experiment 1. MATLAB user interface for the subjective test left/right distinction

The participants were placed with 3 different orientations with respect to the center of the array: frontal, rotated laterally 90 degrees and backward 180 degrees. (see top of Figure 7.4). The intention was to check the frequency separation effect of sound sources from all directions. Because of that, the possible directions of arrival of the sound, considering the middle point angle in between the low and high frequencies, were -30, 2.5, 5, 17.5, 20, 65, 90, 115, 155, 190 and 205 degrees in the horizontal plane. The participants had to indicate where they perceived the high frequency, on

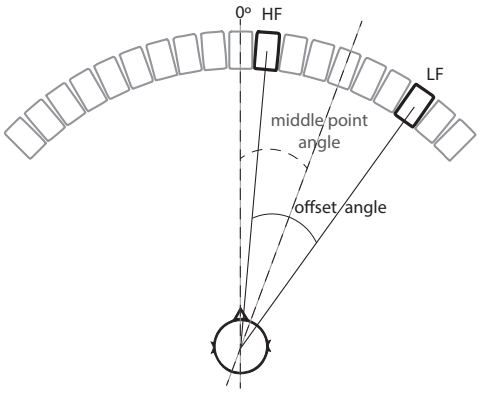


Figure 7.3. Example of middle point angle and offset angle, as used in the perceptual tests. HF and LF indicate the speakers that reproduce the high and low frequencies respectively

the left or on the right of the sound stimulus. In the case of a listening orientation rotated 90 degrees, the same coordinate system as the frontal orientation was used to indicate the answers, as an extension of the frontal orientation. In the case of a 180-degree backward listening orientation, the "left" and "right" positions were inverted to be consistent with the subject's egocentric reference. Figure 7.3 shows a diagram with an example of a middle point angle and an offset angle. Three types of sound were used: brown noise, voice and trombone sound. Table 7.1 summarizes all the stimuli characteristics. Different combinations of the previous characteristics were employed, resulting in 39 stimuli presented three times and randomly to each participant. 9 participants performed the test. To avoid disorientation, reference sounds were available for participants to listen to at any time through buttons on the interface.

High-low frequency separated	**Task: Locate the high frequency**		
Three listening orientations:	frontal	lateral	back
11 middle point angles:	-30, 2.5, 5, 17.5, 20	65, 90, 115	155, 190, 205
6 offset angles:	10, 15, 20, 25, 30	10, 20, 30, 40	10, 20, 30, 40
3 types of sound:	Voice, Trombone, Brown noise		
39 stimuli total	9 participants		

Table 7.1. Summary description of Test 1 Left/Right distinction

Results

In this test, participants were asked to indicate where they perceived the high frequency band of the sound, on the left or on the right. Figure 7.4 shows the percentage of correct answers, for each listening angle orientation and for each high and low frequency separation offset angle. It indicates the rate of correct localization of the high frequency band. Then, 50% means that the listener cannot distinguish the position of the high frequencies and therefore 50% can be considered as the baseline of correct answers. For frontal sources, listeners start to distinguish offset angles of 10 degrees but only slightly above the baseline (64%). This corresponds to the limit where it starts to be statistically significant for one person, according to [274]. As we are averaging the results of different people, it is even less significant. As the angular offset increases, the percentage of correct answers grows as expected, but surprisingly, even with the maximum separation angle (30 degrees) the detection rate is only about 80%. However, for lateral and back sources, high frequency localization is very bad and there is no clear tendency. Due to this lack of discrimination for lateral and back sources the second experiment was carried out just for frontal sources. No significant effect was found between the type of sound and the number of correct answers.

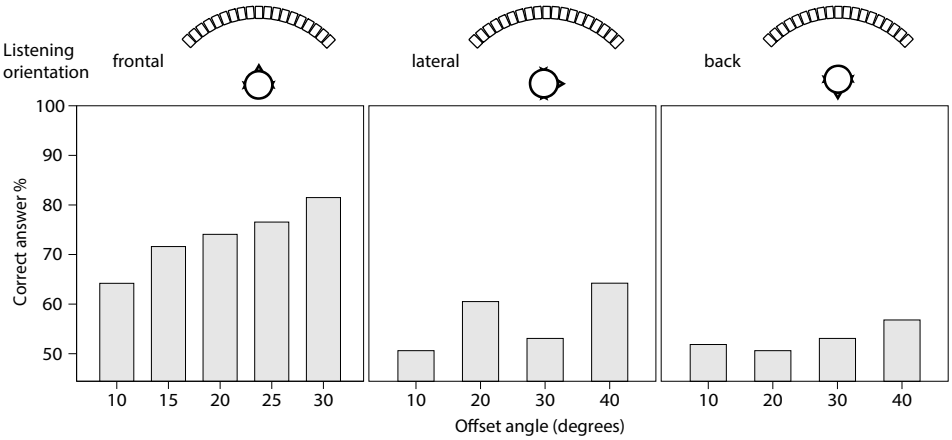


Figure 7.4. Experiment 1. Percentage of correct answers (left or right) for each of the listening orientation angles and offset angles between low-high frequencies

7.4 Test 2. Angle of arrival

Test description

Since experiment 1 indicates that the offset angle between high and low frequencies should be very large to be perceived, it seems conceivable that the possible perceived error in the direction of arrival of the split source would be small. Therefore, a second experiment was prepared to check how much error or deviation angle is perceived in the direction of arrival of a sound source, when the high and low frequencies are separated.

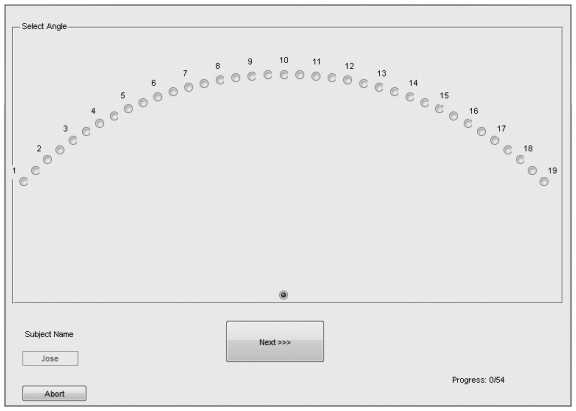


Figure 7.5. Experiment 2. MATLAB user interface for the subjective test angle of arrival

The task of the listeners was to indicate from which direction they perceived the sound, aided by some marks on an acoustically transparent curtain placed in front of the loudspeakers. The answer was to be indicated by means of a graphical user interface (Figure 7.5). In this case, 6 offset angles of low-high frequency separation were employed: 0, 5, 10, 15, 20, 25 degrees, and just one single orientation of listening position with respect to the center of the array, that is, frontal orientation. The respective position of the high and low frequencies on the left or right was randomized. As in experiment 1, participants should listen to different stimuli sounds coming from different directions of arrival, in this case we considered directions from 9 middle point angles: -35, -32.5, -30, 0, 2.5, 5, 15, 17.5, 20 degrees. The same 3 types of sound were used: brown noise, voice and trombone sound. Thus, a total of 54 stimuli (18x3 types of sound) were presented to 10 participants, the same nine as in the previous test plus one other subject.

High-low frequency separated	**Task: Locate the source direction**
One listening orientation:	frontal
9 middle point angles:	-35, -32.5, -30, 0, 2.5, 5, 15, 17.5, 20
6 offset angles:	0, 5, 10, 15, 20, 25
3 types of sound:	Voice, Trombone, Brown noise
54 stimuli total	10 participants

Table 7.2. Summary description of Test 2 Angle source separated

Results

The intention of this test was to identify the perceived error or deviation in the direction of arrival of the sound when the high and low frequencies are spatially split. These are shown in Figure 7.6, where the perceived deviation in degrees is indicated for each offset angle (also in degrees) of separation between low-high frequencies. The perceived deviation is considered as the absolute value of the difference between the answered angle and the middle point angle, meaning the deviations to the left (negative values) and to the right (positive values) together.

The graph shows that although low and high frequencies are emitted from the same loudspeaker (offset angle=0) listeners have an underlying location error of 2.5 degrees. This can be considered as a reference value of the error (or absolute deviation) in the perceived direction of arrival, and be used to quantify the other values in relative terms. At 5° of offset angle there is a similar deviation error of about 2.5 degrees, therefore the effect of separating 5 degrees is virtually non-existent. For 10° and 15° of angular separation, listeners perceive the sound source in the middle of the low and high frequencies without significant distinction with respect to the point source case (0° offset angle). Furthermore, between 0° and 15° there is a difference of just 1.14 degrees (15° offset: 3.33° deviation, 0° offset: 2.19° deviation). Then, up to 15° of offset angle, the perceived error deviates by less than ±0.6° from the reference value, and it can be considered as very small. Only beyond an offset angle of 20 degrees of separation is there a more significant error deviation over 2 or 3 degrees above the reference value.

An Analysis of Variance (ANOVA) confirms that the offset angle is a significant factor in the perception of the direction ($F = 7.577$, $df = 10$,

$p < 0.05$). However, the type of sound is irrelevant in the evaluation of the direction of arrival, according to the significance value $p = 0.170$, as well as the middle point angle between the real sources (from $-35°$ to $20°$ azimuth), for the listening oriented towards the front studied here.

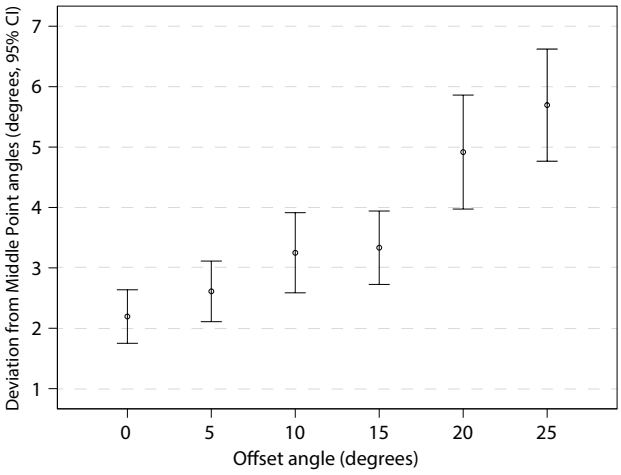


Figure 7.6. Experiment 2. Perceived deviation (in degrees) with respect to the middle point angles (between the two real sources) for each offset angle (separation of high and low frequencies)

7.5 Test 3. Source width

Test description

Results from experiments 1 and 2 can be complemented to further evaluate the spatial sensation of the listeners when the two frequency bands are separated. It would be interesting to find out if this split of frequency bands produces any noticeable defect, pursuant to the width or dispersion of the perceived sound source.

To study this, participants were asked to focus on the perceived width of the different sounds and to indicate their perceptions using a graphical interface (Figure 7.7), on a gradual scale from 1 (widest) to 5 (narrowest). The low and high frequencies were this time separated 7 offset angles: 0, 10, 15, 20, 25, 30 and 40 degrees. Participants listened oriented at three

different angles: frontal, lateral and backward orientation, as in experiment 1 (see top of Figure 7.4). Then, the sounds coming from 18 directions of arrival (considering the middle point angles between the low-high frequencies) were presented: -32.5, -30, 2.5, 5, 17.5, 20, 60, 65, 85, 90, 110, 115, 150, 155, 185, 190, 200 and 205 degrees of the horizontal plane. The same three types of sound were used again: brown noise, voice and trombone sound. 144 stimuli were presented randomly to each of the 11 participants who performed the test, the same ten as in the previous test plus one other subject. Table 7.3 summarizes the different combinations of characteristics of the stimuli. Reference examples of the widest and narrowest sounds of each type, reproduced around the central loudspeaker of the array, were available to be listened to at any time.

High-low freq separated	**Task: Rate source width**		
Three listening orientations:	frontal	lateral	back
18 middle point angles:	-32.5, -30, 2.5, 5, 17.5, 20	60, 65, 85, 90, 110, 115	150, 155, 185, 190, 200, 205
7 offset angles:	0, 10, 15, 20, 25, 30	0, 10, 20, 30, 40	0, 10, 20, 30, 40
3 types of sound:	Voice, Trombone, Brown noise		
144 stimuli total	11 participants		

Table 7.3. Summary description of Test 3 Perceived source width

Results

Using the graphical interface the participants evaluated the perceived width of the sound sources. The perceived width of the sources (1-wide to 5-narrow) for each offset angle (in degrees) is shown in Figure 7.8. There is a tendency in the perception of the width of the source that shows a little wider perception as the offset angle increases.

With an ANOVA we can see that the type of sound is significant in the perception of the width ($F = 11.80$, $df = 4$, $p < 0.05$) but this is mostly due to the perception of the brown noise, that is generally perceived slightly wider than the other type of sounds (Figures 7.9 and 7.10). Until 15° of angular separation between low-high frequencies, the voice is perceived narrower than the other sounds, and over an offset angle of 20° the voice sound is abruptly perceived as wider (Figure 7.9). In summary, the results

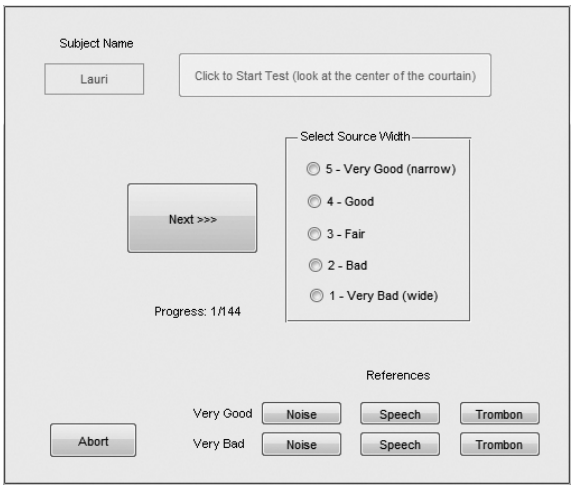


Figure 7.7. Experiment 3. MATLAB user interface for the subjective test source width perception

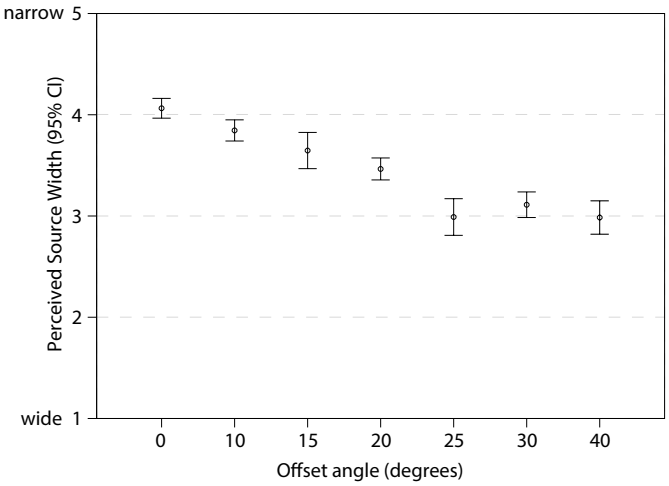


Figure 7.8. Experiment 3. Perceived width of the sources (1-wide to 5-narrow) for each offset angle (between low-high frequencies)

indicate that with offset angles of up to 10° or 15°, even 20° depending on the type of signal, the perception of the source does not begin to become significantly wider and less punctual.

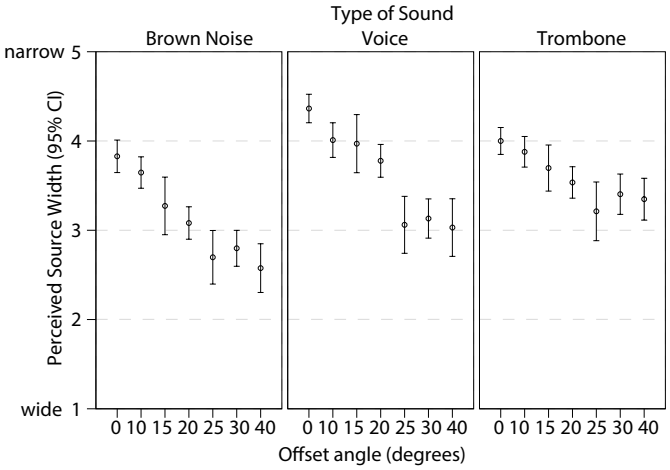


Figure 7.9. Experiment 3. Perceived width of the sources (1-wide to 5-narrow) for each offset angle and type of sound

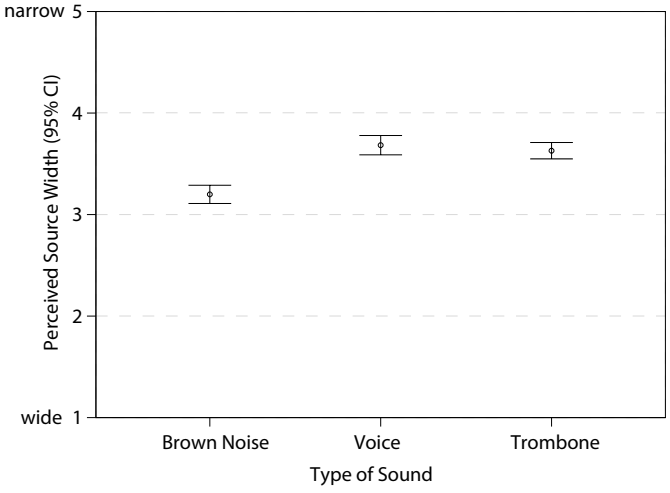


Figure 7.10. Experiment 3. Perceived width of the sources (1-wide to 5-narrow) for each type of sound

7.6 Conclusions and applications

In this chapter it has been studied the perceptual effects of reproducing high and low frequencies of one source from two different spatial positions. High frequencies were considered starting from 1.5 kHz and were reproduced in different loudspeaker positions than the low frequencies. In particular, it has been determined what the limit angles are in order that this separation is not appreciated by the listener and the sound source seems to be only one source and not two different sources. To this end, a series of subjective experiments have been carried out with an array of loudspeakers around the listener, allowing for a precise and feasible positioning.

In a first test a study has been made of what the limit angle is from which the listener is not able to discern on which side the high frequency is and where the low frequency is. It has been concluded that the orientation of the listening position is a key factor in the discrimination of the different frequencies. For frontal sources, listeners start to slightly distinguish separations of 10 degrees but just occasionally. As the offset angle increases, the percentage of correct answers increases as expected, but never reaches 100%. For lateral and backward oriented listening discrimination is very poor.

A second test has also been carried out to check the perceived direction of arrival when high and low frequencies are not reproduced from the same point in space. It has been verified that this angle of deviation is around the middle angle between the high and low frequencies. Up to 15° of offset angle the perceived error is less than ±0.6°, and it can be considered as very small. Only beyond 20 degrees of separation is there a significant error deviation.

In a final test a study has been carried out of how much the perception of source width artificially increases when separating the bands of high and low frequency. It has been found that with angular offsets of up to 10° or 15°, even 20° depending on the type of signal, the perception of the source does not begin to become significantly wider and less punctual.

As discussed in the introduction, these experiments focused on static listeners and/or sound sources. This is, then, a limitation of this work, given that to determine the dynamic angles of listening with movement, other experiments would be needed. In addition, the limited sound stimuli used in the tests may also constrain the results, especially due to their short

duration and the plainness of the content.

In [270], the term *concurrent minimum audible angle* (CMAA) was introduced by Perrott. In this work a new term, *just noticeable band splitting angle* (JNBSA) is defined and introduced. It represents the minimum angle of separation between high and low frequencies from which the listener starts perceiving artifacts (DOA error, source width), in the reproduction of a sound source using loudspeakers. Considering the overall results of the three tests, a JNBSA of 15° can be taken as a conservative value for frontal sources and much bigger for lateral and back sources. Due to the resolution of the set-up and the results of the perceptual tests, 15° of separation between high and low frequencies is the highest offset angle value that does not yet produce significantly incorrect spatial locations.

Applications and future work

The results of this work therefore allow us to study in the future the feasibility of an alternative approach for the reproduction of spatial sound, in which high frequencies are reproduced from a single loudspeaker coming from a different direction than that of the low-frequency counterpart, which can be reproduced with more sophisticated systems. This approach can be used to develop advanced sound systems using multiple loudspeakers with simplified set-ups according to perceptual considerations.

In particular, these findings can be applied in the improvement of WFS reproduction, especially in the reduction of staining artifacts that occur in high frequencies when issued from more than one speaker.

One of the drawbacks of the WFS is the spatial aliasing that occurs at high frequencies because of the separation between loudspeakers [275]. In [276] the OPSI method based on phantom sources was presented, which despite reducing aliasing adds other problems. In [277] a sub-band WFS system is proposed, where the low frequencies are reproduced by means of field synthesis, but the high frequencies are emitted by a single loudspeaker. This system presents the problem that if the listener is not in the center of the array there may be an angular error between the low and high frequency parts. The effect of this error was not evaluated and neither was the possibility of it confusing the perceived position of the source. Thanks to the study introduced here, systems such as the one presented in [277] can be tested to see if they exceed detectability limits.

In addition, due to the large number of loudspeakers that the WFS needs if a high aliasing frequency wants to be achieved, these systems are not widely used in the field of professional sound reinforcement. The current trend, partly reinforced by the introduction of object-based sound systems, is to install multiple loudspeaker arrays but with more affordable separations between 1 and 3 meters, depending on the size of the listening area. In these systems (related to different brands), the reproduction of the sources in each location is achieved by combining aspects of panning and VBAP such as the treatment of the amplitude of the signal that is sent to each speaker, with others that are characteristic of the WFS such as the delay. Hybrid systems are created empirically without a convenient objective evaluation of the result. These systems necessarily create spatial aliasing above a certain frequency. By applying amplitude panning and delays only at low frequencies and emitting the high frequency from a single loudspeaker, these aliasing effects could be reduced while keeping the perceived position of the source in the proper place, if the criteria calculated in this work are met.

Future work plans are to set up one of these hybrid systems (halfway between WFS and VBAP) and subjectively re-evaluate the advantages achieved and the angle error obtained. In addition, with such a system, it would be interesting to test a wider range of realistic sound stimuli with complex characteristics of time and frequency, as well as to test several sound sources simultaneously. This would reveal whether masking effects could condition the listening and detection of the JNBSA. Moreover, 15 degrees was determined as a conservative value of JNBSA, since the experiments have been limited to the resolution of the array set-up, which has 5 degrees separation of loudspeakers. Considering that 20 degrees of angular offset has been found to be significantly detectable, the zone between 15 and 20 degrees could be explored with more precision using an array with closer loudspeakers, and even obtain some kind of threshold bounds within this zone.

Conclusions and future work

8

The overall aim of this research was to optimize and simplify spatial sound reproduction systems, based on a perceptual criterion of the listening experience. The motivation of this research came from the need to make these technologies more accessible to the general public, given the growing demand for audiovisual experiences and devices with increasing immersion.

This chapter will summarize the results of this research work, reviewing the objectives given in the introductory chapter and describing the contributions to the knowledge obtained in the different questions studied. Then, thanks to the insight gained by conducting this research, a discussion is offered on the perspective of spatial sound and the evolution of the technologies studied. Recommendations for future research will also be addressed in a specific section. In addition, the final section contains a list of publications that have been produced from the work contained in this thesis.

8.1 Conclusions and contributions to knowledge

As commented in the introduction, the work presented in this thesis has been divided in two parts: headphones and binaural systems, and loud-

speakers based systems.

In Chapter 3 several perceptual tests have been conducted to evaluate the effects of various properties of headphone models on listening and sound immersion. This is because listening through different headphones generates diverse sound experiences, due to the different characteristics of each headphone model. The experiments outlined the following results: The frequency response has emerged as determinant in the perception of quality and the spatial impression. The binaural recordings employed do not appear to produce greater overall spatial impression than the stereo mixes, because that binaural content did not employ individualized HRTFs. Sensitivity disparities between left and right transducers, starting from a level as low as 1 dB, are detectable by listeners as localization errors, and in addition, irregularities in specific frequency bands can affect the perception of front-back discrimination (100 - 1600 Hz) or lateral position localization (4 - 7 kHz). Although it may be known that low or mid-range headphones are not the best choice for reproducing spatial sound, it had not previously been determined with accuracy how they can affect the perception of sound spatiality. Earlier studies usually only take into account high quality headphones, but here a sample of the widespread group of low and mid-range consumer headphones has also been analysed statistically and with ANOVA. The parameters evaluated and analyzed have produced specific data that can be used as direct recommendations for the construction of headphones or the selection of headphone models for specific applications. A high correlation has been found between quality perception and spatial impression perception, which is an evidence for the scientific community of the importance of the generation of faithful and accurate spatial sound for the reproduction of audiovisual material.

The correction of the frequency response of the headphones by means of equalization is presented as a possible solution to several of the undesired problems detected, but a strong equalization can force a transducer to work out of its linear condition creating non-linear distortion. Given this problem, no other previous studies had evaluated the perception of distortion caused by equalization in the case of headphones. It has been shown that non-linear distortion due to equalization is difficult to detect, but it has high correlation between the level of reproduction at which distortion is detected and the retail price of the headphones. Given the current interest in listening through headphones, these results provide valuable theoretical

and practical knowledge for academics and also for industry.

A complete HRTF measurement system has been implemented and constructed, as described in Chapter 4. The implementation covers all the hardware and software, including room acoustics conditioning, a set-up of 88 loudspeakers, miniature microphones adapted for insertion in the ear canal, an optical laser system to reference the measurements with precision, and the measurement software that allows the use of Exponential Sweep and Multiple Exponential Sweep Method, a fast measurement check and storage in SOFA format. Besides, specific post-processing has been implemented to correct the measured individual HRTF with the responses of the loudspeakers and microphones, to remove reflections of mid and high frequencies with a variant of Frequency Dependant Windowing and to reconstruct the lowest frequency band. The system built has been proven to produce highly accurate measurements in a short lapse of time of less than two and a half minutes. It has to be noted that the measurement system has been constructed on a non-anechoic room, which is an innovative approach. Moreover, a novel method has been proposed and validated to completely remove the reflections that affect the low-mid frequency band. This new method, based on additional spherical microphone array measurements and Plane Wave Decomposition, eliminates the main reflections of the room that cause comb filtering effects in that frequency band. Individual measurements of headphone responses have also been contemplated in the measurement system, along with the creation of the corresponding compensation HpTF filters. The collection of measurements obtained are of great value for the study of HRTF individualization, and the proposed method to completely suppress the reflections of the measured HRTF is an innovative solution which can make this type of measurement more affordable and accessible, promoting the use and expansion of binaural spatial sound.

The individualization of HRTF is an important landmark to be achieved for the solution of various perceptual problems of binaural sound. Chapter 5 presents the research that has been done on this topic, and as a result two tools have been obtained that can be used for the individualization of HRTF. Based on real individual measurements of HRTFs, two perceptual experiments were performed. In a first experiment an algorithm was evaluated to model parametrically the magnitude of the HRTF using PEAK filters. The HRTFs of various positions in elevation located in the median

plane of some listeners were modeled and tested, proving that the parametric model is perceptually valid even with a reduced number of filters. This tested modeling tool has revealed to be a powerful tool for investigating the perceptual mechanisms of HRTF. Psychoacoustic studies on the role of the peaks and notches in HRTF could be easily performed with this parametric model. In addition, the reduction of data that parameterization implies can make it possible to generate a more computationally efficient binaural sound. A second experiment investigated the perception of ITD and its relationship to individual measurements and anthropometric parameters. It was found that the dispersion of the responses of a perceptual test had a significant relationship with two anthropometric dimensions, the intertragus distance and the head perimeter. Regression polynomials have been calculated that adapt the ITD of a generic HRTF to the user by scaling, taking into account the individual values of these two anthropometric parameters. The approach of adapting the ITD by means of a scale factor follows the line of previous research, but here innovation has been made by using two anthropometric variables and second order polynomials for scaling real HRTFs. It is not a model-generated ITD, this allows to take advantage of ITD irregularities that usually exist around 110° and 230° of azimuth that are difficult to model using equations. In addition, polynomials are provided to individualize the ITD by scaling of two dummy heads widely used in binaural sound research and development, which are of great practical value. Furthermore, both individualization tools, the parametric magnitude HRTF modeling and the polynomials to adapt the ITD by a scaling factor, are complementary for the manipulation and individualization of HRTF, since they complete the perceptual model based on minimum phase HRTF in conjunction with the ITD, and together they can provide the complete reconstruction of the HRTF.

Some spatial perceptual attributes have been studied for reproduction systems with loudspeakers. Distance perception is examined in Chapter 6 for WFS and VBAP systems. The influence of early echoes, some types of sound, the listening angle, and the subjects were explored. Statistical analysis showed that, although WFS has a better capability to reproduce sound distance, both WFS and VBAP are able to simulate a sense of distance. This result is of special interest to better understand the perceptual mechanisms that operate with WFS and amplitude panning systems, and to better design the dimension of spatial sound systems. Depending on the application and how critical the distance perception may be, it is possible to

choose with better criterion which system should be used and which characteristics should be implemented. Spatial acuity perception is explored in Chapter 7. Using an array of loudspeakers with 5° resolution in the horizontal plane, the high frequencies (from 1.5 kHz) were reproduced at different loudspeaker positions than the low frequencies of the same sound source. Three different perceptual tests revealed that position discrimination of split high-low frequencies depends on the listening angle and is very poor for lateral and backward positions, that up to 15° of high-low frequency separation the perceived localization error of the sound source is very small, and that the source is not perceived as significantly wider up to similar values of high-low frequency separation angles. These results lead to the definition of a new term called *just noticeable band splitting angle* (JNBSA), that represents the minimum audible angle of separation between high and low frequencies from which the listener starts perceiving artifacts in the reproduction of a sound source. The innovative sub-band approach explored can be employed to develop advanced sound systems using multiple loudspeakers with simplified configurations according to perceptual considerations. WFS can be particularly benefited by an implementation that considers these findings, with promising results in reducing the staining artifacts that occur at high frequencies when emitted from more than one speaker.

Author's reflections

The contributions described are intended to promote the development and dissemination of spatial sound, making these technologies more accessible and providing a better understanding of the perceptual phenomenon of spatial listening. The technological perspective offered by this research work allows to raise questions about the evolution of spatial sound, especially about some specific details that are related to problems discussed in this work.

Creating realistic immersive sound in headphones has become a common pursuit of many headphone technologies, and recent vast efforts have been done to gain the competition of a *de facto* standard for spatial sound that includes reproduction with headphones, such are the technologies of Dolby [103] and Sony [278]. Furthermore, a personalized listening experience and the inclusion of information regarding the listening environment have also become trends in the headphone industry [17, 99]. In the same way that headphones are currently able to adapt their reproduction to

the environment for example with active noise control, the trend for headphones will be to apply signal processing techniques to dynamically adapt the reproduction to the individual, taking into account for example their morphology or their hearing pattern. These trends in headphones of adaptation to the environment and the individualization share one common objective, to render a natural hearing with headphones.

It has already been commented that individualization is a key goal to be achieved in order to improve binaural sound listening, especially through headphones. Individualization provides the solution for many perceptual problems (such as front-to-back confusion, erroneous perception of elevation and localization of sound sources, externalization). However, some authors state that a generic HRTF may be enough for binaural sound [279, 280]. This argument could be valid only if we also consider the application. Depending on the application and the maturity and cost of the technology, a different degree of individualization may be necessary. In favour of this option is the auditory learning process described in some studies [141]. In any case, the degree of error that can be introduced by listening through a non-individual HRTF will always be undetermined as it varies with the individual. In Virtual Reality this fact limits the use of generic HRTFs as it would create a different auralization in the virtual environment than in the real environment. On the contrary, it would be much more natural to use an individual HRTF to have the same auralization in both virtual and real environments, erasing in that case any learning or adaptation period. In addition, there is an unavoidable difference in the coexistence of virtual and real auralizations in the case of Augmented Reality environments. In this case of critical listening where real and virtual sounds must coexist, the integration of the virtual and real sound environments will never be complete unless both auralizations are equal, that is, unless an individualized HRTF is used, since the comparison between both auralizations will be continuous having real listening as reference. An additional perspective a step further in Augmented Reality environments would be to consider an auralization not equal to the individual's actual one to provide enhanced listening that adds new dimensions to the listening of reality. But this possibility will always be more productive and precise by starting with the individual HRTF and then going beyond it.

The merge of different spatial sound technologies to create new solutions and products is the path that research and development will follow

in this field. In this blend, specific techniques for headphones will be able to be used with speakers and customization will also be applied to systems with loudspeakers [281, 282]. In fact, compatibility between playback devices can make listening through headphones act as a Trojan horse for new spatial sound technologies, introducing new techniques that are also compatible with loudspeakers in the domestic environment, without the population being aware of it. Ambisonics is acting as an open format between spatial sound systems and will help this integration of technologies, as well as sound object coding will soon allow the expansion of spatial sound experiences to multiple media, causing the notion of sound "formats" to get blurred. The experience will be increasingly integrated and transversal between possible media or reproduction platforms, with emphasis on the personalization and adaptation of the system and content to the subject and the environment, all with increasing technological transparency towards the user.

8.2 Future work

Further to the research described in this thesis, here are presented several lines of research that remain open. Additional ideas that have not yet been developed and have emerged from the experience of the various experiments are also outlined below.

Different headphones characteristics have been tested to evaluate their effects on the perception of quality and immersion. The same methodology can be applied to more characteristics and headphone models to extend the study and find more features determinant to spatial listening. With regard to the perception of non-linear distortion caused by equalization, it has been demonstrated by a time consuming test that it is perceived in some moderately priced models in relation to the reproduction level. It would be very useful and interesting to propose a model that predicts the audibility of this non-linear distortion for headphones after equalization. For this purpose, the methodology proposed in that experiment could be used and a psychoacoustic model could be added to indicate whether or not the nonlinear distortion is perceptually masked. This model would predict the audibility of the non-linear distortion of a headphone without the need of its evaluation by means of costly perceptual tests.

An HRTF measurement system has been built and used in this work. Several improvements to the existing infrastructure have been planned: increasing the number of elevated loudspeakers to acquire a denser spatial grid of actual measurements, introducing a rotating platform to rotate the person to be measured creating a faster measurement system, or using other miniature microphone models that do not produce intermodulation distortions with MESM to also reduce measurement time. Besides, pretested interpolation techniques to increase the spatial resolution could be optimized for the specific case of the built system. Further refinements will also be desirable on the FDW postprocessing in combination with the proposed method based on PWD to eliminate reflections. A better adjustment of both methods will be very useful to obtain a complete cleaning of the reflections of the HRTFs considering all the directions measured.

Two different and complementary HRTF individualization tools have been studied and developed in this work. The parametric modeling algorithm for the magnitude of the HRTF could improve its performance by limiting the detection of certain peaks and notches to specific bands of the spectrum. It would be worthwhile to investigate with this tool the perceptual mechanisms of the different peaks and notches present in the HRTF, which could lead to a practical use of the model to directly individualize the magnitude of the HRTF. The other individualization tool is based on the adaptation of the ITD by scaling from a generic HRTF. The findings of the ITD scaling experiment demand a new perceptual experiment to evaluate the estimated ITD scale factors with subjective and objective criteria and to compare the results of both. If the objective criterion to scale ITD gives good results, a generalized method could be derived to adapt the ITD of any HRTF set by scaling it to any other person. Besides, more anthropometric measurements could be explored to increase the number of dimensions of the polynomials and use them to extend the scaling to three-dimensional ITD values. In addition, it would be very interesting to design new perceptual tests to evaluate the performance of both individualization tools together, which could easily lead to improvements in some of them. Another interesting idea to explore is to use both tools as analytical methods of the HRTF and use the obtained parameters to feed a machine learning algorithm. Then, the parameters as well as the real HRTF would be the features that could help to identify specific characteristics to localize sound sources, and therefore use them in an individualization HRTF synthesis algorithm.

With regard to loudspeaker reproduction systems, some aspects of spatial perception have been studied. It would be desirable to evaluate the effect of multiple order reflections to better compare the effect of reverberation on the perception of distance in WFS and VBAP. It would also be very interesting to study dynamic listening with moving sources to evaluate both distance and spatial acuity perception in WFS and amplitude panning systems. Complex sound scenes synthesized with various sources and movement would reveal whether masking effects could condition the perception of the distance or the spatial acuity, depending on the reproduction system. Besides, the findings of the spatial acuity experiment could be applied to a WFS system, implementing the tested sub-band approach in a way that the low frequency would be synthesized by WFS while the high frequency would be reproduced by one loudspeaker only. With a properly sized system and a well-defined listening area according to the just noticeable band splitting angle defined, the unwanted high-frequency effects typical of these multi-speaker systems could be reduced.

8.3 List of publications

The following publications have been released from the work contained in this thesis.

They are presented here not in publication order but according to the index of the thesis, in order to facilitate the identification of each publication with the different experiments presented in the thesis.

The list comprises:

3 published, 1 submitted, 1 in preparation JCR-indexed journal articles.
2 non-indexed journal articles.
8 papers in international conferences.
4 papers in national conferences.

Part I Headphones

Chapter 3 Effects of Headphones in perception

J. J. Lopez, P. Gutierrez-Parera, and E. Aguilera, "Influencia de la calidad de los auriculares en la inmersión acústica producida por grabaciones binaurales," in *Tecniacústica 2014 - 45° Congreso Español De Acústica,*

8° Congreso Ibérico De Acústica, Murcia, Spain, 2014, pp. 1028-1035. ISBN 978-84-87985-25-6. ISSN 2340-7441

P. Gutierrez-Parera, J. J. Lopez, and E. Aguilera, "On the influence of headphones quality in the spatial immersion produced by binaural recordings," in *Audio Engineering Society AES 138th Convention*, Warsaw, Poland, 2015.

P. Gutierrez-Parera, J. J. Lopez, and E. Aguilera, "Influencia de la calidad de los auriculares en la percepción espacial sonora," in *Tecniacústica 2015 - 46° Congreso español de acústica, Encuentro ibérico de acústica*, Valencia, Spain, 2015, pp. 1027-1034. ISBN 978-84-87985-26-3. ISSN 2340-7441

P. Gutierrez-Parera and J. J. Lopez, "Influence of the quality of consumer headphones in the perception of spatial audio," Journal of *Applied Sciences*, vol. 6, no. 4, p. 117, 2016. doi: 10.3390/app6040117

P. Gutierrez-Parera and J. J. Lopez, "On the influence of the frequency response over azimuth localization with consumer headphones," in *Audio Engineering Society AES, International Conference on Headphone Technology*, Aalborg, Denmark, 2016. doi: 10.17743/aesconf.2016.978-1-942220-09-1

J. J. Lopez and P. Gutierrez-Parera, "Evaluation of the frequency response of consumer headphones and its influence towards a binaural reproduction with HRTF individualization," in *ICA 2016 - 22nd International Congress on Acoustics*, Buenos Aires, Argentina, 2016.

P. Gutierrez-Parera and J. J. Lopez, "Spatial Audio with Consumer Headphones: How its quality affects the immersion," *Waves* (iTeam Research Journal), vol. 8, pp. 17-29, 2016. ISSN 1889-8297

P. Gutierrez-Parera and J. J. Lopez, "Perception of nonlinear distortion on emulation of frequency responses of headphones," *The Journal of the Acoustical Society of America*, vol. 143, no. 4, pp. 2085-2088, 2018. doi: 10.1121/1.5031030

Chapter 4 HRTF measurements

J. J. Lopez and P. Gutierrez-Parera, "Equipment for fast measurement of Head-Related Transfer Functions," in *Audio Engineering Society AES 142nd Convention*, Berlin, Germany, 2017. ISBN 978-1-5108-4352-3

J. J. Lopez, S. Martinez-Sanchez, and P. Gutierrez-Parera, "Array processing for echo cancellation in the measurement of Head-Related Transfer Functions," in *EAA Euronoise 2018*, Hersonissos, Crete, Greece, 2018, pp. 2581-2588. ISSN: 2226-5147

J. J. Lopez, P. Gutierrez-Parera, S. Martinez-Sanchez, and M. Cobos, "Head-Related Transfer Function Measurements in Non-Anechoic Environments," *IEEE/ACM Transactions on audio, speech and language processing*, 2020. (*Submitted*).

Chapter 5 HRTF individualization tools

P. Gutierrez-Parera and J. J. Lopez, "An experiment to evaluate the performance of a parametric model for the individualization of the HRTF in the median plane," in *Audio Engineering Society AES 142nd Convention*, Berlin, Germany, 2017, pp. 193-200. ISBN 978-1-5108-4352-3

P. Gutierrez-Parera and J. J. Lopez, "Interaural Time Difference individualization in HRTFs by scaling through anthropometric parameters," 2020. (*In preparation*).

Part II Loudspeakers

Chapter 6 Distance perception comparison between WFS and VBAP

J. J. Lopez, P. Gutierrez-Parera, M. Cobos, and E. Aguilera, "Sound distance perception comparison between Wave Field Synthesis and Vector Base Amplitude Panning," in *ISCCSP 2014 - 6th International Symposium on Communications, Control and Signal Processing, IEEE Proceedings*, 2014, pp. 165-168. ISBN 9781479928903. doi: 10.1109/ISCCSP.2014.6877841

P. Gutierrez-Parera and J. J. Lopez, "Comparación de la percepción sonora de distancia con los sistemas Wave Field Synthesis y Vector Base Amplitude Panning," in *Tecniacústica 2014 - 45° Congreso Español De Acústica, 8° Congreso Ibérico De Acústica*, Murcia, Spain, 2014, pp. 1052-1059. ISBN 978-84-87985-25-6. ISSN 2340-7441

P. Gutierrez-Parera, J. J. Lopez, and E. Aguilera, "On the distance perception in spatial audio system: a comparison between Wave-Field Synthesis and Panning Systems," *Waves* (iTeam Research Journal), vol. 6, pp. 51-59, 2014. ISSN 1889-8297

Chapter 7 Perceptual spatial acuity of spectrally divided sound sources

J. J. Lopez, P. Gutierrez-Parera, and L. Savioja, "Effects and applications of spatial acuity in advanced spatial audio reproduction systems with loudspeakers," Journal of *Applied Acoustics*, vol. 161, p. 107179, Apr. 2020. doi: 10.1016/j.apacoust.2019.107179

Other publications derived from work done in relation to the thesis

G. Moreno, M. Cobos, J. Lopez-Ballester, P. Gutierrez-Parera, J. Segura, and A. M. Torres, "On the Development of a MATLAB-based Tool for Real-time Spatial Audio Rendering," in *Audio Engineering Society AES 138th Convention*, Warsaw, Poland, 2015. ISBN 9781510806597

M. Cobos, G. Moreno, J. Lopez-Ballester, P. Gutierrez-Parera, I. Martin-Morato, J. Segura, and A. M. Torres, "Desarrollo de una herramienta Matlab para la reproducción de audio espacial en tiempo real," in *Tecniacústica 2015 - 46° Congreso español de acústica, Encuentro ibérico de acústica*, Valencia, Spain, 2015, pp. 1093-1102. ISBN 978-84-87985-26-3. ISSN 2340-7441

Bibliography

[1] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*, 2nd ed. Cambridge: MIT Press, 1997. ISBN 9780262024136

[2] J. Blauert, *The technology of binaural listening*, J. Blauert, Ed. Springer, 2013. ISBN 9783642377624

[3] A. Blumlein, "Improvements in and relating to sound transmission, sound recording and sound reproduction system (British Patent Specification 394325)," 1933.

[4] T. Holman, *Surround Sound: Up and Running, Second Edition.* Focal Press, 2008. ISBN 9780240808291

[5] T. Holman, "The Number of Audio Channels," in *Audio Engineering Society 100th Convention.* Copenhagen, Denmark: Audio Engineering Society, may 1996.

[6] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[7] H. Wittek, "Perceptual differences between wavefield synthesis and stereophony," Ph.D. dissertation, University of Surrey, England, 2007.

[8] S. Spors, R. Rabenstein, and J. Ahrens, "The Theory of Wave Field Synthesis Revisited," *124th AES Convention*, p. Convention Paper 7358, 2008.

[9] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, pp. 859–871, 1985.

[10] K. Farrar, "Soundfield Microphone: Design and development of microphone and control unit," *Wireless World*, pp. 48–50, 1979.

[11] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial Audio Object Coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding," in *Audio Engineering Society - 124th Audio Engineering Society Convention*, vol. 2, Amsterdam, The Netherlands, 2008, pp. 613–627. ISBN 9781605602950

[12] K. Sunder, Jianjun He, Ee Leng Tan, and Woon-Seng Gan, "Natural Sound Rendering for Headphones: Integration of signal processing techniques," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 100–113, mar 2015. doi: 10.1109/MSP.2014.2372062

[13] Global Market Insights Inc., "Earphones And Headphones Market Size By Technology, By Application, Industry Analysis Report, Regional Outlook, Growth Potential, Price Trends, Competitive Market Share & Forecast, 2017 - 2024," Global Market Insights Inc., Ocean View, USA, Tech. Rep., 2017.

[14] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

[15] X. Amatriain, J. Castellanos, T. Höllerer, J. Kuchera-Morin, S. T. Pope, G. Wakefield, and W. Wolcott, "Experiencing audio and music in a fully immersive environment," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4969 LNCS, 2008, pp. 380–400. ISBN 3540850341. ISSN 03029743. doi: 10.1007/978-3-540-85035-9_27

[16] J. Gómez Bolaños and V. Pulkki, "Immersive audiovisual environment with 3D audio playback," in *132nd Audio Engineering Society Convention 2012*, Budapest, Hungary, 2012, pp. 404–412. ISBN 9781622761180

[17] V. Välimäki, A. Franck, J. Ramo, H. Gamper, and L. Savioja, "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 92–99, 2015. doi: 10.1109/MSP.2014.2369191

[18] S. Bech and N. Zacharov, *Perceptual Audio Evaluation-Theory, Method and Application.* Chichester, England: John Wiley & Sons Ltd, 2006. ISBN 0470869232

[19] D. R. Perrott and K. Saberi, "Minimum audible angle thresholds for sources varying in both elevation and azimuth," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1728–1731, 1990. doi: 10.1121/1.399421

[20] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *Journal of the Audio Engineering Society*, vol. 50, no. 9, pp. 651–666, 2002.

[21] AES69-2015, "AES standard for file exchange - Spatial acoustic data file format," New York, NY, USA, 2015.

[22] A. R. Tilley, *The Measure of Man and Woman: Human Factors in Design (2nd Edition).* John Wiley & Sons Ltd, 2001. ISBN 978-0-471-09955-0

[23] J. W. Strutt (Lord Rayleigh), "On our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, feb 1907. doi: 10.1080/14786440709463595

[24] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1110–1122, 2001. doi: 10.1121/1.1349185

[25] R. H. Gilkey and T. R. Anderson, *Binaural and spatial hearing in real and virtual environments.* Lawrence Erlbaum Associates, 1997. ISBN 9780805816549

[26] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339–368, oct 1940. doi: 10.1037/h0054629

[27] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, sep 1999. doi: 10.1121/1.427180

[28] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz, "Auditory localization of nearby sources. II. Localization of a broadband source," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1956–1968, oct 1999. doi: 10.1121/1.427943

[29] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[30] J. Catic, S. Santurette, J. M. Buchholz, F. Gran, and T. Dau, "The effect of interaural-level-difference fluctuations on the externalization of sound," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1232–1241, aug 2013. doi: 10.1121/1.4812264

[31] J. Catic, S. Santurette, and T. Dau, "The role of reverberation-related binaural cues in the externalization of speech." *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 1154–1167, 2015. doi: 10.1121/1.4928132

[32] A. W. Bronkhorst, "Modeling auditory distance perception in rooms," in *EAA Forum Acusticum*, Sevilla, Spain, 2002.

[33] B. F. Lounsbury and R. A. Butler, "Estimation of distances of recorded sounds presented through headphones," *Scandinavian Audiology*, vol. 8, no. 3, pp. 145–149, 1979. doi: 10.3109/01050397909076315

[34] A. D. Little, D. H. Mershon, and P. H. Cox, "Spectral content as a cue to perceived auditory distance." *Perception*, vol. 21, no. 3, pp. 405–416, 1992. doi: 10.1068/p210405

[35] W. R. Thurlow, J. W. Mangels, and P. S. Runge, "Head Movements During Sound Localization," *The Journal of the Acoustical Society of America*, vol. 42, no. 2, pp. 489–493, aug 1967. doi: 10.1121/1.1910605

[36] S. Perrett and W. Noble, "The effect of head rotations on vertical plane sound localization," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2325–2332, oct 1997. doi: 10.1121/1.419642

[37] F. L. Wightman and D. J. Kistler, "Resolution of frontback ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, may 1999. doi: 10.1121/1.426899

[38] R. A. Lutfi and W. Wang, "Correlational analysis of acoustic cues for the discrimination of auditory motion," *The Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 919–928, aug 1999. doi: 10.1121/1.428033

[39] F. Rumsey, *Spatial Audio*. Oxford, UK: Focal Press, 2001, vol. 33. ISBN 0 240 51623 0

[40] J. Escolano Carrasco, "Contributions to discrete-time methods for room acoustic simulation," Ph.D. dissertation, Universitat Politècnica de València, Valencia (Spain), jul 2008.

[41] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," *Journal of Audio Engineering Society*, vol. 49, no. 4, pp. 231–249, 2001.

[42] J. J. Lopez, "Reproducción de Sonido 3-D mediante Técnicas de Control Local," Ph.D. dissertation, Universitat Politecnica de Valencia, 1999.

[43] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *Journal of Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[44] J. Daniel, R. Nicol, and S. Moreau, "Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging," in *Audio Engineering Society 114th Convention*. Amsterdam, The Netherlands: Audio Engineering Society, mar 2003.

[45] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, Aalto University, Helsinki, 2016.

[46] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, may 1993. doi: 10.1121/1.405852

[47] M. M. Boone, E. N. G. Verheijen, and P. F. van Tol, "Spatial Sound-Field Reproduction by Wave-Field Synthesis," *Journal of the Audio Engineering Society*, vol. 43, no. 12, pp. 1003–1012, dec 1995.

[48] S. Spors and R. Rabenstein, "Spatial aliasing artifacts produced by linear and circular loudspeaker arrays used for wave field synthesis," in *Audio Engineering Society - 120th Convention Spring Preprints 2006*, vol. 3. Paris, France: Audio Engineering Society, may 2006, pp. 1418–1431. ISBN 9781604235975

[49] H. Wierstorf, "Perceptual Assessment of sound field synthesis," Ph.D. dissertation, Technischen Universität Berlin, 2014.

[50] G. Theile, "Über die Lokalisation im überlagerten Schallfeld (Localization in the Superposed Sound Field)," Ph.D. dissertation, Technical University of Berlin, 1980.

[51] D. M. Leakey, "Some Measurements on the Effects of Interchannel Intensity and Time Differences in Two Channel Sound Systems," *The Journal of the Acoustical Society of America*, vol. 31, no. 7, pp. 977–986, 1959. doi: 10.1121/1.1907824

[52] ITU-R BS. 775-3, "Multichannel stereophonic sound system with and without accompanying picture," Geneva, Switzerland, 2012.

[53] E. Torick, "Highlights in the history of multichannel sound," *Journal of the Audio Engineering Society*, vol. 46, no. 1/2, pp. 27–31, 1998.

[54] A. Roginska and P. Geluso, Eds., *Immersive sound: the art and science of binaural and multi-channel audio*. Focal Press, 2017. ISBN 9781317480105

[55] K. Hamasaki, K. Hiyama, and R. Okumura, "The 22.2 multichannel sound system and its application," in *Audio Engineering Society 118th Convention*, Barcelona, Spain, 2005, p. Paper 6406. ISBN 9781604234848

[56] ITU-R BS.2051-2, "Advanced sound system for programme production," Geneva, Switzerland, 2018.

[57] ITU-R BS.2159-8, "Multichannel sound technology in home and broadcasting applications," Geneva, Switzerland, 2019.

[58] S. Kim, Y. W. Lee, and V. Pulkki, "New 10.2-channel vertical surround system (10.2-VSS); Comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers," in *129th Audio Engineering Society Convention 2010*, vol. 2. Audio Engineering Society, nov 2010, pp. 1426–1438. ISBN 9781617821943

[59] European Broadcasting Union EBU, "ETSI TS 103 223 - V1.1.1 - MDA; Object-Based Audio Immersive Sound Metadata and Bitstream," 2015.

[60] T. Sporer, J. Liebetrau, and T. Clauss, "Evaluation of object-based audio. What is the reference?" *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3464–3464, may 2017. doi: 10.1121/1.4987191

[61] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D AudioThe New Standard for Coding of Immersive Spatial Audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, aug 2015. doi: 10.1109/JSTSP.2015.2411578

[62] Auro Technologies - Barco, "AUROMAX: Next generation Immersive Sound system," Auro Technologies NV - Barco NV, Tech. Rep., 2015.

[63] V. Pulkki, "Spatial sound generation and perception by amplitude panning techniques," Ph.D. dissertation, Helsinki University of Technology. ISBN 951- 22-5531-6 2001.

[64] V. Pulkki and M. Karjalainen, "Localization of Amplitude-Panned Virtual Sources I: Stereophonic Panning," *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 739–752, 2001.

[65] V. Pulkki, "Localization of Amplitude-Panned Virtual Sources II: Two- and Three-Dimensional Panning," *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 753–767, 2001.

[66] V. Pulkki and T. Lokki, "Creating Auditory Displays with Multiple Loudspeakers Using VBAP: A Case Study with DIVA Project," in *International Conference on Auditory Display (ICAD)*, Glasgow, England, 1998.

[67] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992. doi: 10.1016/0003-682X(92)90046-U

[68] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–464, jun 1996.

[69] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, no. 6713, pp. 747–749, dec 1998. doi: 10.1038/25526

[70] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 189–194, 1998. doi: 10.1109/89.661479

[71] M. F. Simón Gálvez, T. Takeuchi, and F. M. Fazi, "Low-complexity, listener's position-adaptive binaural reproduction over a loudspeaker array," *Acta Acustica united with Acustica*, vol. 103, no. 5, pp. 847–857, sep 2017. doi: 10.3813/AAA.919112

[72] M. F. Simón Gálvez, D. Menzies, and F. M. Fazi, "Dynamic audio reproduction with linear loudspeaker arrays," *Journal of the Audio Engineering Society*, vol. 67, no. 4, pp. 190–200, apr 2019. doi: 10.17743/jaes.2019.0007

[73] G. Theile, "On the Standardization of the Frequency Response of High-Quality Studio Headphones," *Journal of the Audio Engineering Society*, vol. 34, no. 12, pp. 956–969, 1986.

[74] D. Griesinger, "Equalization and spatial equalization of dummy-head recordings for loudspeaker reproduction," *Journal of the Audio Engineering Society*, vol. 37, no. 1-2, pp. 20–29, feb 1989.

[75] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Academic Press, 1994. ISBN 0-12-084735-3

[76] P. X. Zhang and W. M. Hartmann, "On the ability of human listeners to distinguish between front and back," *Hearing Research*, vol. 260, no. 1-2, pp. 30–46, 2010. doi: 10.1016/j.heares.2009.11.001

[77] S.-M. Kim and W. Choi, "On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach," *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3657–3665, 2005. doi: 10.1121/1.1921548

[78] D. W. Batteau, "The role of the pinna in human localization." *Proceedings of the Royal Society of London. Series B. Biological sciences*, vol. 168, no. 11, pp. 158–180, aug 1967. doi: 10.1098/rspb.1967.0058

[79] D. S. Brungart, B. D. Simpson, R. L. Mckinley, A. J. Kordik, R. C. Dallman, and D. A. Ovenshire, "The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources," in *Proceedings of ICAD 04- Tenth Meeting of the International Conference on Auditory Display*, Sydney, Australia, 2004, pp. 1–7.

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

[81] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3678–3688, 1996. doi: 10.1121/1.414965

[82] F. L. Wightman and D. J. Kistler, "Headphone simulation of freefield listening. I: Stimulus synthesis," *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, feb 1989. doi: 10.1121/1.397557

[83] H. Møller, M. F. Sorensen, D. Hammershoi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *Journal of the Audio Engineering Society*, vol. 43, no. 5, pp. 300–321, 1995.

[84] S. Xu, Z. Li, and G. Salvendy, "Individualization of Head-Related Transfer Function for Three-Dimensional Virtual Auditory Display:

A Review," in *12th International Conference on Human-Computer Interaction (HCI International 2007)*, vol. 4563, 2007, pp. 397–407. ISBN 978-3-540-73334-8. ISSN 03029743. doi: 10.1007/978-3-540-73335-5_44

[85] R. Nicol, *Binaural Technology*. New York: Audio Engineering Society Inc., 2010. ISBN 9780937803721

[86] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 325–328. ISBN 9781424436798. doi: 10.1109/ASPAA.2009.5346532

[87] M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold, "HRTF customization using multiway array analysis," *European Signal Processing Conference*, no. January 2010, pp. 229–233, 2010.

[88] V. R. Algazi, R. O. Duda, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," in *113th Convention of the Audio Engineering Society*. Los Angeles, USA: Audio Engineering Society, oct 2002, pp. 1–18. ISSN 1438-2199. doi: 10.1007/s00726-011-0912-4

[89] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 2003-Janua. New Paltz, New York, USA: IEEE, 2003, pp. 157–160. ISBN 0780378504. doi: 10.1109/ASPAA.2003.1285855

[90] T. Huttunen and A. Vanne, "End-to-end process for HRTF personalization," in *Audio Engineering Society 142nd Convention*, Berlin, Germany, 2017.

[91] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1480–1492, sep 1999. doi: 10.1121/1.427176

[92] K. J. Fink and L. Ray, "Individualization of head related transfer functions using principal component analysis," *Applied Acoustics*, vol. 87, pp. 162–173, 2015. doi: 10.1016/j.apacoust.2014.07.005

[93] B. Boren and A. Roginska, "The effects of headphones on listener HRTF preference," in *131st Audio Engineering Society Convention*, vol. 1. Audio Engineering Society, oct 2011, pp. 386–393. ISBN 9781618393968

[94] D. Schonstein, L. Ferré, and B. F. Katz, "Comparison of headphones and equalization for virtual auditory source localization," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3724–3724, may 2008. doi: 10.1121/1.2935199

[95] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," *The Journal of the Acoustical Society of America*, vol. 100, no. 6, pp. 3785–3793, 1996. doi: 10.1121/1.417337

[96] A. Lindau and F. Brinkmann, "Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings," *Journal of Audio Engineering Society*, vol. 60, no. 1/2, pp. 54–61, 2012.

[97] J. Gómez Bolaños, A. Mäkivirta, and V. Pulkki, "Automatic Regularization Parameter for Headphone Transfer Function Inversion," *Journal of Audio Engineering Society*, vol. 64, no. 10, pp. 752–761, 2016. doi: 10.17743/jaes.2016.0030

[98] A. Kulkarni and H. S. Colburn, "Variability in the characterization of the headphone transfer-function," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 1071–1074, 2000. doi: 10.1121/1.428571

[99] J. Backman, T. Campbell, J. Kleimola, and M. Hiipakka, "A Self-Calibrating Earphone," in *Audio Engineering Society 142nd Convention*, Berlin, Germany, 2017.

[100] Fraunhofer IIS, "WHITE PAPER MPEG Spatial Audio Object Coding (SAOC) The MPEG Standard on Parametric Object Based Audio Coding," Fraunhofer IIS, Tech. Rep., 2012.

[101] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio - The New Standard for Universal Spatial/3D Audio Coding," *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, jan 2014. doi: 10.17743/jaes.2014.0049

[102] Dolby Laboratories Inc., "Dolby Atmos." [Online]. Available: https://www.dolby.com/xl/es/brands/dolby-atmos.html Accessed: 2020-01-08.

[103] Dolby Laboratories Inc., "Dolby Atmos Music — Immersive audio experiences that move you." [Online]. Available: https://music.dolby.com Accessed: 2019-12-13.

[104] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, jun 2017. doi: 10.1121/1.4983652

[105] C. Andersson, "Headphone Auralization of Acoustic Spaces Recorded with Spherical Microphone Arrays," Ph.D. dissertation, Chalmers University of Technology. Gothenburg, Sweden, 2017.

[106] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional audio coding: Virtual microphone-based synthesis and subjective evaluation," *AES: Journal of the Audio Engineering Society*, vol. 57, no. 9, pp. 709–724, sep 2009.

[107] F. Zotter and M. Frank, *Ambisonics A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality.* Springer, Cham, 2019. ISBN 978-3-030-17206-0

[108] F. Zotter, "Analysis and synthesis of sound-radiation with spherical arrays," Ph.D. dissertation, University of Music and Performing Arts. Graz, Austria, 2009.

[109] A. J. Berkhout, "A Holographic Approach to Acoustic Control," *Journal of the Audio Engineering Society*, vol. 36, no. 12, pp. 977–995, 1988.

[110] S. Spors, H. Buchner, R. Rabenstein, and W. Herbordt, "Active listening room compensation for massive multichannel sound reproduction systems using wave-domain adaptive filtering," *The Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 354–369, jul 2007. doi: 10.1121/1.2737669

[111] P. A. Gauthier and A. Berry, "Objective evaluation of room effects on wave field synthesis," *Acta Acustica united with Acustica*, vol. 93, no. 5, pp. 824–836, 2007.

[112] Futuresource Consulting Ltd., "Futuresource Headphones Market Report - Worldwide Oct 19," Futuresource Consulting, St Albans, Hertfordshire, GB, Tech. Rep., 2019.

[113] M. Geronazzo, J. Kleimola, E. Sikstroöm, A. de Götzen, S. Serafi, and F. Avanzini, "HOBA-VR: HRTF On Demand for Binaural Audio in immersive virtual reality environments," in *Audio Engineering Society 144th Convention*. Milan, Italy: Audio Engineering Society, may 2018, pp. e–Brief 433.

[114] ITU-T Rec. P.57, "Artificial ears," Geneva, Switzerland, 2011.

[115] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," *Journal of the Audio Engineering Society*, vol. 43, no. 4, pp. 203–217, 1995.

[116] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design criteria for headphones," *Journal of the Audio Engineering Society*, vol. 43, no. 4, pp. 218–232, apr 1995.

[117] S. E. Olive, T. Welti, and E. Mcmullin, "Listener Preference For Different Headphone Target Response Curves," in *134th Audio Engineering Society Convention*, Rome, Italy, 2013, p. paper 8867. ISBN 9781627485715

[118] S. E. Olive, T. Welti, and O. Khonsaripour, "The Influence of Program Material on Sound Quality Ratings of In-Ear Headphones," in *142nd Convention Audio Engineering Society*, Berlin, Germany, 2017, pp. 1–12.

[119] S. E. Olive, T. Welti, and O. Khonsaripour, "A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 1 Listening Test Results and Acoustic Measurements," in *Audio Engineering Society 143rd Convention*, New York, USA, 2017, p. paper 9840.

[120] S. E. Olive, T. Welti, and O. Khonsaripour, "A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones:

Part 2 Development and Validation of the Model," in *Audio Engineering Society 143rd Convention*, New York, USA, 2017, p. paper 9878.

[121] S. E. Olive, O. Khonsaripour, and T. Welti, "A Survey and Analysis of Consumer and Professional Headphones Based on Their Objective and Subjective Performances," in *AES 145th Convention*, New York, NY, USA, 2018, p. paper 10048.

[122] M. Opitz, "Headphones Listening Tests," in *Audio Engineering Society 121st Convention*, San Francisco, USA, 2006. ISBN 9781604236637

[123] H. Y. Sung, S. Jang, and J.-B. Kim, "A Method for Objective Sound Quality Evaluation of Headphones," in *Audio Engineering Society 32nd International Conference*, Hillerod, Denmark, 2007.

[124] T. Hirvonen, M. Vaalgamaa, J. Backman, and M. Karjalainen, "Listening Test Methodology for Headphone Evaluation," in *Audio Engineering Society 114th Convention*, Amsterdam, The Netherlands, 2003.

[125] G. Lorho, "Subjective Evaluation of Headphone Target Frequency Responses," in *Audio Engineering Society 126th Convention*, Munich, Germany, 2009. ISBN 9781615671663

[126] S. E. Olive, T. Welti, and E. Mcmullin, "The Influence of Listeners' Experience, Age, and Culture on Headphone Sound Quality Preferences," in *Audio Engineering Society 137th Convention*, Los Angeles, USA, 2014.

[127] S. E. Olive and T. Welti, "The Relationship between Perception and Measurement of Headphone Sound Quality," in *Audio Engineering Society 133rd Convention*, San Francisco, USA, 2012. ISBN 9781622766031

[128] F. Briolle and V. Thierry, "Transfer Function and Subjective Quality of Headphones: Part 2, Subjective Quality Evaluations," in *Audio Engineering Society Conference: 11th International Conference: Test & Measurement*, Portland, USA, 1992, pp. 254–259.

[129] S. E. Olive, T. Welti, and E. McMullin, "A Virtual Headphone Listening Test Methodology," in *Audio Engineering Society 51st International Conference*, Helsinki, Finland, 2013. ISBN 9781629933283

[130] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society 108th International Convention*, Paris, France, 2000. ISBN 0780356128

[131] A. Farina and E. Armelloni, "Emulation of not-linear, time-variant devices by the convolution technique," in *Annual Meeting 2005, Audio Engineering Society Italian Section*, Como, Italy, 2005.

[132] L. Tronchin, "The Emulation of Nonlinear Time-Invariant Audio Systems with Memory by Means of Volterra Series," *Journal of Audio Engineering Society*, vol. 60, no. 12, pp. 984–995, 2012.

[133] V. Pulkki and M. Karjalainen, *Communication Acoustics. An Introduction to Speech, Audio and Psychoacoustics.* Chichester,West Sussex, United Kingdom: JohnWiley & Sons, Ltd, 2015. ISBN 9781118866542

[134] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," Geneva, Switzerland, 2015.

[135] ITU-R BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems," Geneva, Switzerland, 2015.

[136] T. Letowski, "Sound Quality Assessment: Concepts and Criteria," in *Audio Engineering Society 87th Convention.* New York, NY, USA: Audio Engineering Society, oct 1989.

[137] N. Zacharov and K. Koivuniemi, "Unraveling the perception of spatial sound reproduction: Techniques and experimental design," in *Audio Engineering Society 19th International Conference.* Schloss Elmau, Germany: Audio Engineering Society, jun 2001.

[138] B. G. Shinn-Cunningham, "Learning reverberation: Considerations for spatial auditory displays," in *6th International Conference on Auditory Display (ICAD).* Atlanta, GA: Georgia Institute of Technology, 2000.

[139] P. Minnaar, S. K. Olesen, F. Christensen, and H. Møller, "Localization with Binaural Recordings from Artificial and Human Heads," *Journal of the Audio Engineering Society*, vol. 49, no. 5, pp. 323–336, may 2001.

[140] O. Santala and V. Pulkki, "Directional perception of distributed sound sources," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1522–1530, mar 2011. doi: 10.1121/1.3533727

[141] C. Mendonça, G. Campos, P. Dias, and J. A. Santos, "Learning Auditory Space: Generalization and Long-Term Effects," *PLoS ONE*, vol. 8, no. 10, oct 2013. doi: 10.1371/journal.pone.0077900

[142] R. H. Y. So, B. Ngan, A. Horner, J. Braasch, J. Blauert, and K. L. Leung, "Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound: cluster analysis and an experimental study," *Ergonomics*, vol. 53, no. 6, pp. 767–781, jun 2010. doi: 10.1080/00140131003675117

[143] S. Temme, S. E. Olive, S. Tatarunis, T. Welti, and E. Mcmullin, "The Correlation Between Distortion Audibility and Listener Preference in Headphones," in *137th International Convention Audio Engineering Society*, Los Angeles, USA, 2014.

[144] K. Sunder, E. Tan, and W. Gan, "Effect of Headphone Equalization on Auditory Distance Perception," in *Audio Engineering Society Convention 137*, Los Angeles, USA, 2014, pp. 1–11. ISBN 9781634397483

[145] J. Breebaart, "No correlation between headphone frequency response and retail price," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. EL526–EL530, 2017. doi: 10.1121/1.4984044

[146] D. W. Martin and L. J. Anderson, "Headphone Measurements and Their Interpretation," *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 63–70, 1947. doi: 10.1121/1.1916404

[147] J. Liski, V. Välimäki, S. Vesa, and R. Väänänen, "Real-Time Adaptive Equalization for Headphone Listening," in *25th European Signal Processing Conference (EUSIPCO)*, Kos island, Greece, 2017, pp. 638–642. ISBN 9780992862671

[148] IEC 61672-1, "Electroacoustics - Sound level meters - Part 1: Specifications," 2013.

[149] A. Novak, P. Lotton, and L. Simon, "Synchronized swept-sine: Theory, application and implementation," *Journal of the Audio Engineering Society*, vol. 63, no. 10, pp. 786–798, 2015. doi: 10.17743/jaes.2015.0071

[150] D. L. Clark, "Ten years of A/B/X testing," in *Audio Engineering Society 91st Convention*, New York, USA, 1991.

[151] "Dubset loop 140 bpm by waveplay." [Online]. Available: http://freesound.org/people/waveplay/sounds/198495/ Accessed: 2017-09-25.

[152] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, may 1990. doi: 10.1121/1.399186

[153] J. G. Richter and J. Fels, "Evaluation of localization accuracy of static sources using HRTFs from a fast measurement system," *Acta Acustica united with Acustica*, vol. 102, no. 4, pp. 763–771, jul 2016. doi: 10.3813/AAA.918992

[154] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of Head-Related Transfer Functions," *Journal of the Audio Engineering Society*, vol. 55, no. 7-8, pp. 623–636, 2007.

[155] P. Dietrich, B. Masiero, and M. Vorländer, "On the optimization of the multiple exponential sweep method," *Journal of the Audio Engineering Society*, vol. 61, no. 3, pp. 113–124, 2013.

[156] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, jun 1995. doi: 10.1121/1.412407

[157] J. Blauert, M. Brueggen, A. W. Bronkhorst, R. Drullman, G. Reynaud, L. Pellieux, W. Krebber, and R. Sottek, "The AUDIS catalog of human HRTFs," *The Journal of the Acoustical*

Society of America, vol. 103, no. 5, pp. 3082–3082, may 1998. doi: 10.1121/1.422910

[158] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, New York, USA: IEEE, 2001, pp. 99–102. ISBN 0-7803-7126-7. doi: 10.1109/ASPAA.2001.969552

[159] O. Warusfel, "Listen HRTF Database - IRCAM," 2003. [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/index.html Accessed: 2019-10-09.

[160] Acoustics Research Institute, "ARI HRTF Database," 2007. [Online]. Available: https://www.kfs.oeaw.ac.at/index.php?view=article{&}id=608{&}lang=en Accessed: 2019-10-09.

[161] N. Gupta, A. Barreto, M. Joshi, and J. C. Agudelo, "HRTF database at FIU DSP Lab," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, 2010, pp. 169–172. ISBN 9781424442966. ISSN 15206149. doi: 10.1109/ICASSP.2010.5496084

[162] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoustical Science and Technology*, vol. 35, no. 3, pp. 159–165, mar 2014. doi: 10.1250/ast.35.159

[163] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 1, pp. 1–16, sep 2019. doi: 10.17743/jaes.2019.0024

[164] International Organization for Standarization, "ISO 3382-1:2009 Acoustics - Measurement of room acoustic parameters Part 1: Performance spaces (ISO 3382-1:2009)," 2009.

[165] D. Y. Maa, "The Flutter Echoes," *Journal of the Acoustical Society of America*, vol. 13, no. 2, pp. 170–178, oct 1941. doi: 10.1121/1.1916161

[166] J. M. Berman and L. R. Fincham, "The Application of Digital Techniques to the Measurement of Loudspeakers," *Journal of the Audio Engineering Society*, vol. 25, no. 6, pp. 370–384, jun 1977.

[167] S. Müller and P. Massarani, "Transfer-Function Measurement with Sweeps," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443–471, jun 2001.

[168] D. D. Rife and J. Vanderkooy, "Transfer-Function Measurement with Maximum-Length Sequences," *Journal of the Audio Engineering Society*, vol. 37, no. 6, pp. 419–444, jun 1989.

[169] "SOFA (Spatially Oriented Format for Acoustics) - Sofaconventions." [Online]. Available: https://www.sofaconventions.org/ Accessed: 2019-11-06.

[170] D. Rudrich, "IEM Plug-in Suite." [Online]. Available: https://plugins.iem.at/ Accessed: 2019-11-06.

[171] "Personalized HRTFs in Rapture3D Advanced - Blue Ripple Sound." [Online]. Available: http://www.blueripplesound.com/personalized-hrtfs Accessed: 2019-11-06.

[172] "Harpex." [Online]. Available: https://harpex.net/ Accessed: 2019-11-06.

[173] L. Mccormack and A. Politis, "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *Audio Engineering Society Conference on Immersive and Interactive Audio 2019*, no. March, York, England, 2019.

[174] Eurecat Multimedia Technologies, "The Sfëar suite - QRUSH Immersive Audio - sound in space." [Online]. Available: http://qrush.space/ Accessed: 2019-11-07.

[175] M. Karjalainen and T. Paatero, "Frequency-dependent signal windowing," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics.* IEEE, 2001, pp. 35–38. ISBN 0-7803-7126-7. doi: 10.1109/aspaa.2001.969536

[176] F. Denk, B. Kollmeier, and S. D. Ewert, "Removing reflections in semianechoic impulse responses by frequency-dependent truncation," *AES: Journal of the Audio Engineering Society*, vol. 66, no. 3, pp. 146–153, 2018. doi: 10.17743/jaes.2018.0002

[177] S. Fontana and A. Farina, "A System for Rapid Measurement and Direct Customization of Head Related Impulse Responses," in *120th Convention Audio Engineering Society*, Paris, France, 2006. ISBN 9781604235975

[178] J. Gómez Bolaños and V. Pulkki, "HRIR database with measured actual source direction data," in *Audio Engineering Society 133rd Convention*, no. 1, 2012.

[179] C. J. Struck and S. F. Temme, "Simulated free field measurements," *Journal of the Audio Engineering Society*, vol. 42, no. 6, pp. 467–482, 1994.

[180] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2053–2064, nov 2002. doi: 10.1121/1.1508780

[181] N. A. Gumerov, A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin, "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 370–386, jan 2010. doi: 10.1121/1.3257598

[182] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Fortschritte der Akustik – AIA-DAGA 2013*, 2013, pp. 592—-595.

[183] B. Xie, "On the low frequency characteristics of head-related transfer functions," *Chinese Journal of Acoustics*, vol. 28, no. 2, pp. 116–128, 2009.

[184] V. Erbes, HagenWierstorf, M. Geier, and S. Spors, "Free database of low-frequency corrected head-related transfer functions and head-

phone compensation filter," in *142nd Convention Audio Engineering Society*, Berlin, Germany, 2017, pp. 1–5.

[185] O. Kirkeby, E. T. Seppälä, A. Kärkkäinen, L. Kärkkäinen, and T. Huttunen, "Some effects of the torso on Head-Related Transfer Functions," in *Audio Engineering Society - 122nd Audio Engineering Society Convention 2007*, vol. 2. Audio Engineering Society, may 2007, pp. 1045–1052. ISBN 9781604231403

[186] F. A. Everest and K. C. Pohlmann, *The master handbook of acoustics - sixth edition.* McGraw-Hill, 2015. ISBN 0071841040

[187] Earthworks Inc., "M30 High Definition Measurement Microphone," Mildford, NH, USA, p. 3055.

[188] M. acoustics, "em32 Eigenmike microphone." [Online]. Available: https://mhacoustics.com/products Accessed: 2019-11-14.

[189] Brüel & Kjær, "TYPE 4100 - Brüel & Kjær Sound & Vibration, sound quality Head and Torso Simulator." [Online]. Available: https://www.bksv.com/en/products/transducers/ear-simulators/head-and-torso/hats-type-4100 Accessed: 2019-09-25.

[190] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "SOFiA Sound Field Analysis Toolbox," in *International Conference on Spatial Audio (ICSA)*, 2011, pp. 7–15.

[191] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, may 2014, pp. 5177–5181. ISBN 9781479928927. ISSN 15206149. doi: 10.1109/ICASSP.2014.6854590

[192] K. I. McAnally and R. L. Martin, "Variability in the headphone-to-ear-canal transfer function," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 263–266, apr 2002.

[193] E. A. G. Shaw, "Earcanal Pressure Generated by Circumaural and Supraaural Earphones," *The Journal of the Acoustical Society of America*, vol. 39, no. 3, pp. 471–479, mar 1966. doi: 10.1121/1.1909914

[194] J. R. Sank, "Improved Real-Ear Test for Stereophones," *Journal of the Audio Engineering Society*, vol. 28, no. 4, pp. 206–218, apr 1980.

[195] F. E. Toole, "The Acoustics and Psychoacoustics of Headphones," *AES 2nd International Conference: The Art and Technology of Recording*, p. Paper Number C1006, 1984.

[196] W. L. Martens, "Individualized and generalized earphone correction filters for spatial sound reproduction," *Proceedings of the 2003 International Conference on Auditory Display (ICAD)*, pp. 263–266, 2003.

[197] I. Engel, D. L. Alon, P. W. Robinson, and R. Mehra, "The effect of generic headphone compensation on binaural renderings," in *AES Conference on Immersive and Interactive Audio*, York, UK, 2019.

[198] F. Brinkmann and A. Lindau, "On the effect of individual headphone compensation in binaural synthesis," in *Proc. of the 36th DAGA*, no. May 2014, Berlin, Germany, 2010, pp. 1055–1056.

[199] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *Journal of the Audio Engineering Society*, vol. 47, no. 7, pp. 583–595, jul 1999.

[200] Z. Schärer and A. Lindau, "Evaluation of Equalization Methods for Binaural Signals," in *AES 126th Convention*, Munich, Germany, 2009, p. 17. ISBN 9781615671663

[201] Georg Neumann GmbH, "Neumann KU100 Dummy head." [Online]. Available: https://en-de.neumann.com/ku-100 Accessed: 2019-09-25.

[202] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, jul 1993. doi: 10.1121/1.407089

[203] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency." *The Journal of the Acoustical Society of America*, vol. 106, no. 3 Pt 1, pp. 1493–1510, sep 1999. doi: 10.1121/1.427147

[204] D. R. Begault, E. M. Wenzel, A. S. Lee, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," in *Audio Engineering Society 108th Conference*. Paris, France: Audio Engineering Society, feb 2000. ISSN 00047554

[205] B. U. Seeber and H. Fastl, "Subjective selection of non-individual head-related transfer functions," in *Proceedings of the 2003 International Conference on Auditory Display*. Georgia Institute of Technology, 2003, pp. 1–4.

[206] E. A. Torres-Gallegos, F. Orduña-Bustamante, and F. Arámbula-Cosío, "Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database," *Applied Acoustics*, vol. 97, pp. 84–95, oct 2015. doi: 10.1016/j.apacoust.2015.04.009

[207] B. P. Bovbjerg, F. Christensen, P. Minnaar, and X. Chen, "Measuring the head-related transfer functions of an artificial head with a high directional resolution," in *Audio Engineering Society 109th Convention*. Los Angeles, USA: Audio Engineering Society, sep 2000, p. Preprint 5264.

[208] B. F. G. Katz, "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2440–2448, nov 2001. doi: 10.1121/1.1412440

[209] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *Journal of the Acoustical Society of America*, vol. 61, no. 6, pp. 1567–1576, jun 1977. doi: 10.1121/1.381470

[210] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *The Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167, jul 1977. doi: 10.1121/1.381498

[211] D. J. Kistler and F. L. Wightman, "A Model Of Head-Related Transfer Functions Based On Principal Components Analysis And Minimum-Phase Reconstruction," *Journal of the Acoustical Society*

of America, vol. 91, no. 3, pp. 1637–1647, mar 1992. doi: 10.1121/1.402444

[212] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "On the minimum-phase approximation of head-related transfer functions," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 1995, pp. 84–87. ISBN 0-7803-3064-1. doi: 10.1109/ASPAA.1995.482964

[213] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, mar 2000. doi: 10.1121/1.428447

[214] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 364–374, 2005. doi: 10.1121/1.1923368

[215] K. Iida, M. Yairi, and M. Morimoto, "Role of pinna cavities in median plane localization," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, p. 2844, 1998. doi: 10.1121/1.421507

[216] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *Journal of the Acoustical Society of America*, vol. 56, no. 6, pp. 1829–1834, dec 1974. doi: 10.1121/1.1903520

[217] B. C. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8khz," *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 820–836, feb 1989. doi: 10.1121/1.397554

[218] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 159–168, jul 1990. doi: 10.1121/1.399963

[219] J. Breebaart, F. Nater, and A. Kohlrausch, "Parametric binaural synthesis: Background, applications and standards," *Proc. of NAG-DAGA 2009*, no. Volume I, pp. 172–175, 2009.

[220] M. A. Blommer and G. H. Wakefield, "Pole-zero approximations for head-related transfer functions using a logarithmic error criterion," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 278–287, 1997. doi: 10.1109/89.568734

[221] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and zero modeling of head-related transfer functions," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 188–195, 1999. doi: 10.1109/89.748123

[222] C. Liu and S. Hsieh, "Common-acoustic-poles/zeros approximation of head-related transfer functions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 5. IEEE, 2001, pp. 3341–3344. ISBN 0-7803-7041-4. doi: 10.1109/ICASSP.2001.940374

[223] E. A. Durant and G. H. Wakefield, "Efficient model fitting using a genetic algorithm: Pole-zero approximations of HRTFs," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 1, pp. 18–27, 2002. doi: 10.1109/89.979382

[224] A. Kulkarni and H. S. Colburn, "Infinite-impulse-response models of the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1714–1728, apr 2004. doi: 10.1121/1.1650332

[225] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Applied Acoustics*, vol. 68, no. 8, pp. 835–850, 2007. doi: 10.1016/j.apacoust.2006.07.016

[226] G. Ramos and J. J. Lopez, "Filter design method for loudspeaker equalization based on IIR parametric filters," *Journal of the Audio Engineering Society*, vol. 54, no. 12, pp. 1162–1178, 2006.

[227] J. J. Lopez, M. Cobos, and B. Pueo, "Elevation in wave-field synthesis using HRTF cues," *Acta Acustica united with Acustica*, vol. 96, no. 2, pp. 340–350, 2010. doi: 10.3813/AAA.918283

[228] U. Zölzer, *Digital audio signal processing*, 2nd ed. Chichester, West Sussex, UK: Wiley, 2008. ISBN 9780470997857

[229] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency inter aural time differences in sound localization," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, mar 1992. doi: 10.1121/1.402445

[230] M. T. Pastore and J. Braasch, "The impact of peripheral mechanisms on the precedence effect," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 425–444, jul 2019. doi: 10.1121/1.5116680

[231] R. S. Woodworth and H. Schlosberg, *Experimental psychology, Rev. ed.* Oxford, England: Holt, 1954.

[232] V. Larcher and J.-M. Jot, "Techniques d'interpolation de filtres audio-numérique, Application à la reproduction spatiale des sons sur écouteurs," in *Proceedings of the Congrès Français d'Acoustique*, 1997.

[233] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999.

[234] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 472–479, 2001. doi: 10.1017/CBO9781107415324.004

[235] S. Busson, "Individualisation d'indices acoustiques pour la synthèse binaurale," Ph.D. dissertation, Université de la Méditerranée - Aix-Marseille II, 2006.

[236] R. O. Duda, C. Avendano, and V. R. Algazi, "An adaptable ellipsoidal head model for the interaural time difference," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2. IEEE, 1999, pp. 965–968. ISBN 0-7803-5041-3. ISSN 07367791. doi: 10.1109/ICASSP.1999.759855

[237] R. Bomhardt, M. Lins, and J. Fels, "Analytical Ellipsoidal Model of Interaural Time Differences for the Individualization of Head-Related Impulse Responses," *Journal of the Audio Engineering Society*, vol. 64, no. 11, pp. 882–893, 2016. doi: 10.17743/jaes.2016.0041

[238] M. Aussal, F. Alouges, and B. F. G. Katz, "ITD Interpolation and Personalization for Binaural Synthesis using Spherical Harmonics," in *Spatial Audio in today's 3D world - AES 25th UK Conference*, York, England, 2012.

[239] X. Zhong and B. Xie, "An individualized interaural time difference model based on spherical harmonic function expansion," *Chinese Journal of Acoustics*, vol. 32, no. 3, p. 284, 2013.

[240] X. Zhong and B. Xie, "A novel model of interaural time difference based on spatial fourier analysis," *Chinese Physics Letters*, vol. 24, no. 5, pp. 1313–1316, may 2007. doi: 10.1088/0256-307X/24/5/052

[241] A. Lindau, J. Estrella, and S. Weinzierl, "Individualization of dynamic binaural synthesis by real time manipulation of the ITD," in *Proc of the 128th AES Convention*, London, UK, 2010. ISBN 9781617387739

[242] F. Christensen, G. Martin, P. Minnaar, W. K. Song, B. Pedersen, and M. Lydolf, "A listening test system for automotive audio - Part 1: System description," in *Audio Engineering Society - 118th Convention*, vol. 1, Barcelona, Spain, 2005, pp. 163–172. ISBN 9781604234848

[243] A. Andreopoulou, D. R. Begault, and B. F. G. Katz, "Inter-Laboratory Round Robin HRTF Measurement Comparison," *IEEE Journal on Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 895–906, 2015. doi: 10.1109/JSTSP.2015.2400417

[244] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 1583, no. 4, pp. 1583–1596, oct 2002. doi: 10.1121/1.1501901

[245] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *Audio Engineering Society 98th Convention.* Paris, France: Audio Engineering Society, feb 1995.

[246] P. Minnaar, J. Plogsties, S. K. Olesen, F. Christensen, and H. Møller, "The Interaural Time Difference in Binaural Synthesis," in *Audio*

Engineering Society 108th Convention, Paris, France, 2000, pp. 1–20. ISBN 0780356128. doi: 10.1109/ASPAA.1999.810884

[247] B. F. G. Katz and M. Noisternig, "A comparative study of interaural time delay estimation methods," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3530–3540, 2014. doi: 10.1121/1.4875714

[248] A. Andreopoulou and B. F. G. Katz, "Identification of perceptually relevant methods of inter-aural time difference estimation," *The Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. 588–598, 2017. doi: 10.1121/1.4996457

[249] K. Watanabe, K. Ozawa, Y. Iwaya, Y. Suzuki, and K. Aso, "Estimation of interaural level difference based on anthropometry and its effect on sound localization," *The Journal of the Acoustical Society of America*, vol. 122, no. 5, pp. 2832–2841, 2007. doi: 10.1121/1.2785039

[250] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of HRTFs on the horizontal plane using physical features," *Applied Acoustics*, vol. 68, no. 8, pp. 897–908, 2007. doi: 10.1016/j.apacoust.2006.12.010

[251] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in hrtf individualization," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, HSCMA'11*. IEEE, may 2011, pp. 213–218. ISBN 9781457709999. doi: 10.1109/HSCMA.2011.5942401

[252] M. Romanov, P. Berghold, D. Rudrich, M. Zaunschirm, M. Frank, and F. Zotter, "Implementation and Evaluation of a Low-cost Headtracker for Binaural Synthesis," in *142nd Convention Audio Engineering Society*, Berlin, Germany, 2017, pp. 1–6.

[253] S. Werner, G. Götz, and F. Klein, "Influence of head tracking on the externalization of auditory events at divergence between synthesized and listening room using a binaural headphone system," in *Audio Engineering Society 142nd International Convention*, Berlin, Germany, 2017.

[254] B. Rosner, "Percentage points for a generalized esd many-outlier procedure," *Technometrics*, vol. 25, no. 2, pp. 165–172, may 1983. doi: 10.1080/00401706.1983.10487848

[255] A. Andreopoulou and B. F. G. Katz, "Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees," *Journal on Multimodal User Interfaces*, vol. 10, no. 3, pp. 259–271, 2016. doi: 10.1007/s12193-016-0214-y

[256] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Applied Sciences*, vol. 8, no. 11, p. 2029, oct 2018. doi: 10.3390/app8112029

[257] A. Andreopoulou and B. F. G. Katz, "Investigation on Subjective HRTF Rating Repeatability," in *Audio Engineering Society 140ConventionConvention*, Paris, France, 2016.

[258] B. G. Shinn-Cunningham, N. I. Durlach, and R. M. Held, "Adapting to supernormal auditory localization cues. I. Bias and resolution," *The Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3656–3666, 1998. doi: 10.1121/1.423088

[259] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. Wiley, 1990. ISBN 9780470316801

[260] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, "Head related transfer function personalization based on multiple regression analysis," in *2006 International Conference on Computational Intelligence and Security, ICCIAS 2006*, vol. 2. IEEE, nov 2007, pp. 1829–1832. ISBN 1424406056. doi: 10.1109/ICCIAS.2006.295380

[261] W. W. Hugeng and D. Gunawan, "Improved method for individualization of Head-Related Transfer Functions on horizontal plane using reduced number of anthropometric measurements," *Journal of Telecommunications*, vol. 2, no. 2, pp. 31–41, may 2010.

[262] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002. doi: 10.1121/1.1458027

[263] H. Wittek, S. Kerber, F. Rumsey, and G. Theile, "Spatial Perception in Wave Field Synthesis Rendered Sound Fields: Distance of Real and Virtual Nearby Sources," in *Audio Engineering Society 116th Convention*. Berlin, Germany: Audio Engineering Society, may 2004.

[264] M. Noguès, E. Corteel, and O. Warusfel, "Monitoring distance effect with wave field synthesis," *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03)*, pp. 8–11, 2003.

[265] EBU 3276, "2nd Edition. Listening conditions for the assessment of sound programme material: monophonic and twochannel stereophonic," Geneva, Switzerland, 1998.

[266] ITU-R BS.1284-2, "General methods for the subjective assessment of sound quality," Geneva, Switzerland, 2019.

[267] J. Li and Y. Yan, "Distance perception synthesis in 3D audio rendering using loudspeaker array," in *2011 International Conference on Multimedia Technology, ICMT 2011*, 2011, pp. 290–293. ISBN 9781612847740. doi: 10.1109/ICMT.2011.6001858

[268] V. Pulkki, "Coloration of Amplitude-Panned Virtual Sources," in *Audio Engineering Society 110th Convention*. Amsterdam, The Netherlands: Audio Engineering Society, may 2001, p. Paper 5402.

[269] D. Malham, "Higher order ambisonic systems for the spatialisation of sound," in *International Computer Music Conference Proceedings (ICMC 1999)*, 1999, pp. 484–487. ISSN 2223-3881

[270] D. R. Perrott, "Concurrent minimum audible angle: a re-examination of the concept of auditory spatial acuity." *The Journal of the Acoustical Society of America*, vol. 75, no. 4, pp. 1201–1206, 1984. doi: 10.1121/1.390771

[271] A. W. Mills, "On the Minimum Audible Angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958. doi: 10.1121/1.1909553

[272] H. Sato, H. Sato, M. Morimoto, and Y. Nakai, "Localization of intermittent sound with head movement: Basic study on optimum temporal characteristics of acoustic guide signals," *Applied Acoustics*, vol. 101, pp. 58–63, jan 2016. doi: 10.1016/J.APACOUST.2015.08.003

[273] J. J. Lopez and A. Gonzalez, "3-D audio with dynamic tracking for multimedia environments," in *2nd COST-G6 Workshop on Digital Audio Effects*, 1999.

[274] H. Burstein, "Approximation Formulas for Error Risk and Sample Size in ABX Testing," *Journal of the Audio Engineering Society*, vol. 36, no. 11, pp. 879–883, 1988.

[275] G. Theile, H. Wittek, and M. Reisinger, "Potential wavefield synthesis applications in the multichannel stereophonic world," in *AES 24th International Conference on Multichannel Audio*, 2003, p. Paper 35.

[276] H. Wittek, "OPSI: Optimised Phantom Source Imaging of the high frequency content of virtual sources in Wave Field Synthesis," Internal CARROUSO Paper, Tech. Rep., 2002.

[277] J. J. Lopez, S. Bleda, B. Pueo, and J. Escolano, "A Sub-band Approach to Wave-Field Synthesis Rendering," in *Audio Engineering Society 118th Convention*, Barcelona, Spain, 2005, p. Paper 6403. ISBN 9781604234848

[278] Sony Corporation, "360 Reality Audio — So Immersive. So Real." [Online]. Available: https://www.sony.com/electronics/360-reality-audio Accessed: 2019-12-13.

[279] C. C. Berger, M. Gonzalez-Franco, A. Tajadura-Jiménez, D. Florencio, and Z. Zhang, "Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity," *Frontiers in Neuroscience*, vol. 12, no. FEB, p. 21, feb 2018. doi: 10.3389/fnins.2018.00021

[280] P. Stitt, L. Picinali, and B. F. G. Katz, "Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues Through Active Learning," *Scientific Reports*, vol. 9, no. 1, p. 1063, dec 2019. doi: 10.1038/s41598-018-37873-0

[281] F. M. Fazi and M. F. Simón Gálvez, "Sound reproduction system, Patent US-2019-0090060-A1," 2019.

[282] L-Acoustics Group, "L-ISA Immersive Sound Art." [Online]. Available: http://www.l-isa-immersive.com/ Accessed: 2020-01-18.