POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

Master of Science in
Computer Engineering

Interactive binaural rendering of virtual acoustic scenes

Candidate

Mauro Molteni

Student Id. number 788577

Thesis Supervisor

Prof. Augusto Sarti

Assistant Supervisor

Dr. Antonio Canclini

Academic Year 2013/2014

POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

Laurea Magistrale in
Ingegneria Informatica

Riproduzione interattiva di scene acustiche virtuali con tecniche binaurali

Candidato

Mauro Molteni
Matricola 788577

Relatore

Prof. Augusto Sarti

Correlatore

Dr. Antonio Canclini

Anno Accademico 2013/2014

Interactive binaural rendering of virtual acoustic scenes
Master thesis. Politecnico di Milano

This thesis has been typeset by LaTeX and the smcthesis class.

Author's email: mauro1.molteni@mail.polimi.it

Dedicated to
my family

Sommario

La riproduzione di eventi sonori tramite cuffie è sempre più usata, anche per via della diffusione di dispositivi come smartphone, riproduttori portatili di contenuto multimediale e tablet computer. Questa forma di ascolto privato limita però l'esperienza dell'ascoltatore, proponendogli un suono localizzato all'interno della testa e senza informazioni riguardanti l'ambiente di ascolto, che è ben lontano dall'esperienza che si sperimenta nella vita reale.

Al fine di proporre all'utente un suono spazializzato, sono state investigate diverse tecniche che permettono di rendere un suono in una qualsiasi posizione dello spazio all'interno di un ambiente di ascolto. In particolare, l'efficienza nella simulazione di sorgenti sonore virtuali in cuffia è legata alla conoscenza di caratteristiche fisiche dell'ascoltatore, che influenzano nella vita reale l'ascolto da parte delle persone. Lo studio di queste caratteristiche richiede l'utilizzo di tecniche di misurazione costose e invasive, che producono dei risultati molto accurati (ma molto personalizzati). Per superare questi limiti, altri metodi sono stati proposti in letteratura, al fine di ottenere dei modelli non personalizzati che non si affidino a misurazioni fisiche. Anche se questi metodi risultano essere meno accurati, risultano essere interessanti per varie applicazioni.

In questa tesi proponiamo un sistema sonoro 3D che permette di riprodurre una o più sorgenti virtuali nello spazio, basate su un modello generale dell'ascoltatore. Il sistema proposto tiene anche in considerazione l'effetto di un ambiente di ascolto nel quale è possibile posizionare delle sorgenti sonore, utilizzando una tecnica basata sull'acustica geometrica (beamtracing), che permette di calcolare, in modo efficiente, l'effetto dell'ambiente sul campo acustico. Infine, il sistema è stato pensato per poter offrire all'utente un certo grado di interattività, fornendogli gli strumenti per poter esplorare l'ambiente nel quale viene posizionato. È stata dunque implementata la possibilità per l'utente di potersi muovere all'interno della stanza, potendo scegliere un punto di ascolto desiderato, e di poter inoltre muovere la testa.

Per ottenere una valutazione soggettiva delle capacità della nostra tecnologia, abbiamo condotto dei test percettivi, che ci hanno confermato alcuni limiti del nostro sistema in condizioni non favorevoli (come ad esempio quelle statiche), mentre risultano interessanti in condizioni di movimento dell'ascoltatore e di realismo offerto. I risultati ci permettono di concludere che il nostro sistema, nella sua semplicità, risulta essere efficace nell'esplorazione di un ambiente di ascolto, potendo essere utilizzato al posto di metodologie più costose a parità di condizione di movimento. Possiamo inoltre affermare che l'interazione uomo-sistema e le simulazioni di suono realistiche tramite cuffia sono tecnologie da seguire e continuare a sviluppare nel corso dei prossimi anni.

Abstract

The reproduction of sound events through headphones becoming more and more popular, mainly due to the diffusion of devices like smartphones, mobile digital media players and tablet computers. This modality of private listening enables a limited experience, providing a in-head localized sound without information about the listening environment, that is far away the listening experience of the real life.

In order to improve the listening experience, different techniques for rendering a sound in space within a listening environment were investigated. In particular, the efficiency in rendering virtual sound sources over headphones is related to the knowledge of the physical characteristics of the listener, that in real life affect the perception of sounds. The study of such characteristics require invasive and expensive measurement techniques, which produce very accurate (but highly individualized) results. In order to overcome these issues, other methods have been proposed in the literature, aimed at obtaining non-individualized models that do not rely on physical measurements. Even if such models are generally less accurate, they turn to be attractive for several applications.

In this thesis we propose a 3D sound system that allows to reproduce one or more virtual sound sources in space, based on a general model of the listener. The proposed system also renders the effect of the virtual environment in which virtual sources are located. This is accomplished by means of a beamtracing technique, which exploits the laws of geometrical acoustics for predicting the impulse responses from the virtual sources to the listener position. The system has been thought in order to provide the user a certain degree of interactivity, giving the tools to explore the environment in which he or she is located. It has been implemented the possibility for the user to move inside a room, choosing a desired listening position, and rotating the head.

In order to evaluate the proposed system, we have conducted several perceptual tests, that confirmed some limits of the system in unfavorable conditions (like for example the static ones), while interesting results were shown in conditions of motion of the listener, which turned to enhance the perceived degree of realism. The results allows us to conclude that our system in its simplicity results effective in the exploration of a listening environment and it can be used in place of more expensive methodologies under equal conditions of motion. We can state that the human-computer interaction and the rendering of realistic sounds through headphones are technologies to take into consideration or developing novel and attractive applications.

Acknowledgments

The past years at university changed me a lot and increased my desire to know and learn. I would like to thank all the people I have met during these years and who have taught me something.

A special thanks to my parents: without you I would not be here today. Thank you for giving me opportunities that maybe you could never have.
I would like to thank Professor and my supervisor Augusto Sarti for giving me the opportunity to work on a topic that I particularly appreciated. I am particularly grateful to my assistant supervisor Antonio for his assistance, the continuous support offered and his patience during the preparation of the thesis.
I'd like to thanks Giulia for her daily support and encouragement: thank you for believing in me and stand by me.
I wish to acknowledge the help provided by my friend Lucio: never stop dreaming.
Thank you to all the people involed in the listening tests.

Finally, a thought to my grandmothers: I wish you were here.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The continuous technology innovation in the electronics field and in the human-computer interaction is opening up new avenues for innovative applications in different areas. Also the sound reproduction area is affected by these changes and improvements, trying to provide always more realistic and high fidelity systems. In particular, for some years the interest in reproducing virtual auditory scenes is more and more increased.

This work of thesis is the result of a research about a system implementation aimed at the evaluation of the possibility to produce a 3D-spatial sound over headphones. In particular the final implementation is about an immersive and realistic technology, that allows the user to interact with.

Reprduction systems through loudspeakers are able to provide a surrounding signal and also a spatialized signal, but they present some limitations: they don't allow private listening, and, the correct wavefield can be reproduced only over a limited area, called sweet-spot. Simply put, the problem with using loudspeakers for 3D sound is that control over perceived spatial imagery is greatly sacrificed, since the sound will be reproduced in an unknown environment. In other words, the room and loudspeakers will impose unknown transformations that usually cannot be easily compensated for by the designer or controller of the 3D audio system. Headphone listening conditions can be roughly approximated from stereo loudspeakers using a technique called cross-talk cancellation, that allows to eliminate the crosstalk, i.e the signal that from the right channel arrives at the left ear and the signal that from the left channel arrives at the right ear.

This work is motivated by the growing interest and development of sound field rendering techniques, and by the increasingly developed applications in augmented reality. Sound reproduction through headphones is getting more and more the preferred way to consume music; as well, consumer technologies like smartphones are also music players and in this context there are many avenues of research. Mobile voice communications applications that use these new capabilities include audio teleconferencing or telepresence in a meeting. Properly reproduced over headphones, spatial sound can provide an astonishingly lifelike sense of being remotely immersed in the presence of people, musical instruments, and environmental sounds. For voice communication, spatial sound can go beyond increased realism to enhance intelligibility and can provide the natural binaural cues needed for spatial discrimination. For

environmental monitoring or games, it can provide unparalleled awareness of both the sound-generating objects and the surrounding acoustic space. Spatial sound will also be used in conjunction with video in remote monitoring to provide rapid sonic detection and orientation of events for subsequent detailed analysis by video. Finally, one can think for example at the new idea of dancing, 'Silent Disco', in which rather than using a speaker system the music is transmitted to the headphones. As well personal listening, audio for videogames, telecommunication, and so many other possible applications.

Normally, what we hear through headphones is different from the reality. In fact, without a specific binaural processing, that is a processing of the input signal in order to try to reproduce the filtering effects introduced by the listener's body, especially by the torso, the head and the outer ear, the ears receive dry signals that are perceived inside the head: this results in a limitation of the experience. In addition, the same signal is sent to both the ears. Therefore, by conveying a binaural signal to the headphones, it is possible to render a sound in space that enables a more immersive and realistic listening experience. In order to provide a spatialized sound over headphones we must be able to represent the way in which people hear sounds in real life. All these informations are collected by the transfer function that relate the sound pressure at the sound source, and the sound pressure that arrives at the ears of the listener: it is called Head Related Transfer Function (HRTF). It differs from person to person and it can be measured through a long and costly process; alternatively, it is possible to use a generic model that approximates the real ones. An example of HRTF database is freely available online [18, 19]. In this thesis we use an approach based on a structural model of the HRTF, that is a generic non personalized model. We have worked on various cues able to provide informations about simulating real-life hearing, looking for the best suited models for each of them. Using a non personalized model some problems arise, like the difficulty in hearing a sound localized outside the head and having a pronounced separation between sounds in the front and in the back of the head. We are aware of these problems but our final implementation is related to a system in which the listener is in motion: the listener is able to explore an environment and interact with the system changing the listening position and rotating the head. This allows, to some extent, to overcome the problems listed above.

In order to evaluate our work we have designed and conducted some perceptual listening tests, collecting subjective assessments about various aspects and characteristics of the system. In particular, we considered the system capability to render localized sounds in space, the system capability to render a virtual moving listener in a room, the immersivity and realism of the sounds generated. Finally, a questionnaire was proposed to the subjects undergoing the tests, which were asked to express a qualitative judgment of the overall system. This permits to have informative subjective results about the capability of the system to render a convincent auditory scene.

Outline

The thesis is organized as follow: in Chapter 2 we present the state of the art about the reproduction sound systems devoted to sound field rendering. In particular

we present a distinction between those based on loudspeakers and those based on headphones. Furthemore, we describe the limitations and the positive aspects of each one. Then we explain how it is possible to take into account the cues involved in real life hearing. After that we speak about the contribution of a virtual environment, and the cues that are exploited to perceive a position inside a room. Finally we treat the possibility to take into account the head movement.

In Chapter 3 we treat in details the aspects related to the implementation of our system. We show the model we use to represent the way humans listen to sounds, and we explain the single cues involved. Moreover we introduce the method used to represent the listening environment and consider its cues. Finally we describe the method used to take into account the head rotation.

The Chaper 4 is about the interface and the interaction with the system. In particular we show the listening environments considered, describing their geometrical and acoustical characteristics. Then we describe the way the users can use to rotate the head of the virtual listener, and finally we present the graphical user interface that includes all that is needed to understand the virtual sound scene.

In Chapter 5 we describe the criteria leading to the design and planning of formal listening tests aimed at a subjective evaluation of our system implementation. In particular, we show how we have employed standard recommendations from ITU-R, Broadcasting service (Sound) (BS) and adapted them to our specific needs.

In Chapter 6 we present all the experimental results in order to prove the effectiveness of our model. Moreover, we present a discussion of the results, aimed at highlighting the strong relation between our model and the results found in literature.

Finally, in Chapter 7 we draw some conclusions and we show possible future works.

Chapter 2

Background

From the early nineties to present days, sound reproduction techniques have changed and evolved over time [20]. This evolution was dependent from the interest in reproducing a realistic sound that surrounds the listener. In order to obtain a realistic reproduction of sounds it is necessary to position a sound in space independently from the sound reproduction system configuration: for this reason, an increasingly interest arised around techniques that allow to simulate a sound source in space. This is possible using a simple loudspeaker configuration but the desired wavefield can be correctly reproduced inside a limited listening area and in addition there is no isolation between the signals intended for the left and right ear. Conversely, a sound reproduction based on headphones provides a channel separation and there is no area outside of which it's not possible a correct sound reproduction. After the observation about the human capability of perceiving sounds in 3D using only two receivers, i.e. the ears, and the possibility to simulate a sound position changing the sound difference between the ears, researchers were persuaded toward sound synthesis over headphones. The first binaural sound system can be considered the "Théâtrophone" in 1932 [21] and today the sound reproduction through headphones is widespread.

Simulating a sound in space needs to take into consideration a set of cues present in real life that modify the sound before reaching the eardrums: indeed, if we were able to faithfully reproduce the alterations introduced by the path from the sound source to the listener, we could reproduce in a realistic way a sound through headphones. In order to take into account the sound perceived from the listeners when they ear sound in a real situation, in this Chapter we present several ways to consider the cues involved in the listening, each with its own advantages and disadvantages. In particular, we will focus on a simple but efficient way to obtain the sounds to be sent to the listener's ears. For the purposes of an entire 3D sound system, we would like to model not only the outer ear, but also the entire body involved in diffraction and reflection of sound before reaching the eardrum.

As well, in order to simulate a listening in a given room, the cues introduced by the listening environment are considered. In addition, this chapter treats the habit of humans to move the head in order to minimize cues from head when hearing a sound to localize: the reason is that several studies [10, 22, 23] have shown that allowing a listener to move the head can improve localization ability and increase the sense of realism.

In this Chapter we present all the components needed to reproduce a faithful realistic binaural sound through headphones. First, we motivate the growing interest about technologies that are able to synthesize immersive spatial sounds. Then we introduce in Section 2.2 the coordinates system used. In Section 2.3, we present the idea behind the possibility of reproducing realistic sounds over headphones: we talk about the transfer function connecting the pressure at the sound source and the pressure at the eardrums, why it's so important, what it represents. In Section 2.4, we introduce the listening environment and we describe the transfer function that takes into account not only the direct sound but also the reflections. Then, in Section 2.5, we talk more in details about the characteristics of a listening environment and the contribution that it introduces, and in Section 2.6, we describe how it is possible to take into account the contribution of the reflection paths introduced by the environment.

Furthermore, in Section 2.7, we present some existing systems that spatialize sound through headphones.

2.1 Sound reproduction

Sound reproduction is the process of reproducing sound waves (such as voice, singing, instrumental sounds and other sound effects) by means of electronic and electro-mechanical devices. Sound reproduction techniques have evolved over time and more and more interest has been focused around the possibility to create a virtual auditory scene, i.e. the effect in the listener to perceive a sound in a particular listening environment. Indeed, audio is much important in multimedia and Virtual Reality Application, and through 3D audio techniques it's possible to build a 3D audio model: in this way, the audio content is more attractive and realistic [24].
From the early days of phonograph in the late-19th century, monophonic sound reproduction was the rule for almost all audio production scenarios. Typically, these systems consist of one single loudspeaker or, in situations where an increased sound pressure is needed, multiple loudspeakers fed by a single signal. Such a system does not provide a position in space: the listener always perceives the sound as coming from the loudspeaker itself. The increasing interest in producing the effect of an appareant sound source positioned anywhere in space, led to the development of new reproduction techniques.

2.1.1 Reproduction through loudspeakers

In 1931 Blumlein [25] filled a British patent which described the basics of stereo recording and reproduction, and which is up to now the basic of all stereo recording techniques: he developed two-channel recording methods [26] in the attempt of creating an illusion of directionality and sound scene perspective. These methods prescribe the use of two loudspeakers fed by independent signals. Such techniques are referred to as *two-channel stereophony* and nowadays they are commonly employed in entertainment applications (e.g. FM radio and TV broadcasting, popular music production). In 1934 researchers at AT&T Bell labs described two major configurations of spatial audio reproduction [27, 28]: two channels and multi-channel

[29]. In experiments they proved three loudspeakers (left, center, right) provide superior quality to a larger audience compared to two loudspeakers [29].

Blumlein recognized that using a simple system with two loudspeakers, it was possible to delocalize a sound with respect to the position of the loudspeakers simply varying the sound amplitude: this technique is called *2D amplitude panning* [30]. The idea is that it is possible to position a virtual sound source along the arc connecting the two loudspeakers, changing the intensity levels of the two output signals [31]. The technique is depicted in Figure 2.1: two loudspeakers are placed at the same angle θ_0 with respect to the front of the listener and a virtual sound source that we want to simulate at an angle θ. In order to do so, the two loudspeakers are fed by the same signal s(t), but multiplied by two different gain factors g_L ang g_R for the left and the right loudspeaker, respectively. The gain factors vary depending on the source position that we want to simulate, i.e. on the angle θ. This model can be improved taking also into account the phase differences, and so adding two different delays of the input signal [30].

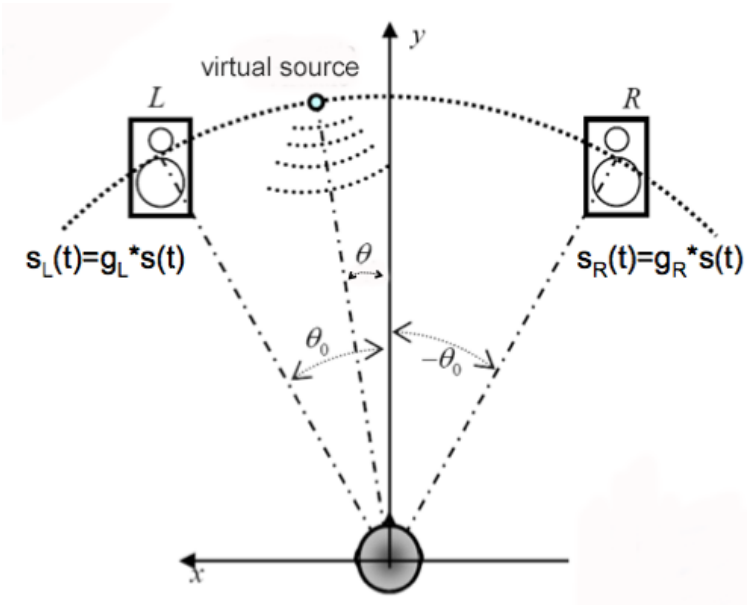

Figure 2.1. Sound reproduction system configured with two loudspeakers L and R positioned at an angle θ_0 with respect to the median plane of the listener. Varying the amplitude of the output signals it is possible simulating a virtual source positioned at any angle θ along the arc connecting the two loudspeakers. [1]

Since the *2D amplitude panning* technique allows us to only position a virtual sound source along the arc connecting the two loudspeakers, a successive step led us to the use of a technique with three loudspeakers aimed at taking into account a 3D space. The typical two-channel stereophonic listening configuration is extended with a third loudspeaker placed in an arbitrary position at the same distance from the listener as the other loudspeakers. However, the loudspeaker should not be placed on the two-dimensional plane defined by the listener and the two other loudspeakers [1] (Figure 2.2). Thus, the previous concept is expanded into a *3D amplitude panning* denoting a method for positioning a virtual sound source into a triangle formed

by three sound sources, which are driven by the same input signal with different amplitudes.

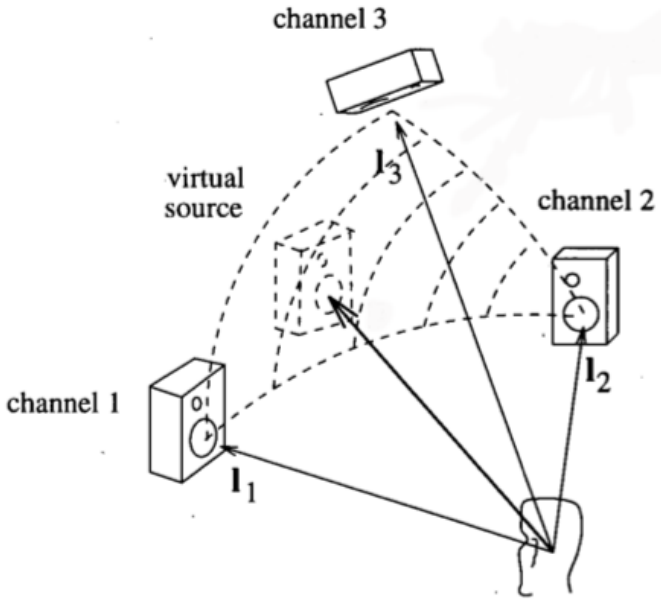

Figure 2.2. Configuration for three-dimensional amplitude panning [1].

These simple amplitude panning techniques have a particular disadvantage: a listener standing near a speaker will perceive the sound as coming from that speaker even if its level is lower than that of another speaker fed with the same signal. Thus, the listener has to be equidistant to the speakers, in a restricted area [1].
The term *stereophony* also refers to more complex systems like *surround* systems, which employ a set of loudspeakers surrounding the listeners. Nowadays cinema and soundtracks are the major applications of surround techniques. With both stereophonic and surround systems the correct reproduction of the sound scene is restricted to a narrow listening area, usually named *sweet spot* [29]. Outside this area timbral and spatial distortions occur. Moreover, in an audio system with loudspeakers we don't have two distinct channels, but the output signals mix in the air before reaching the listener; so, using loudspeakers, it's not possible sending independent signals to the ears, but it is necessary using the crosstalk cancellation technique [32] in order to present at right ear a different signal with respect to the left ear and vice versa. This technique attempts to emulate the listening experience provided by the headphones. However, the sweet spot turns to be very narrow in this case.

After a commercially non-successful extension to four channels (quadrophony) in the seventies, today 5-channel stereo, which adds 2 surround channels and a center channel is used more and more [33]. But two, three and five channel stereo systems are doomed to preserve some limitations along the time [2]:

- good spatial audio quality is limited to a small portion of the reproduction room, the so-called sweet spot;

- (virtual) sound sources can be placed at loudspeaker position, between loudspeaker positions or farer apart from the listener, but not in the gap between loudspeaker and listener;
- (virtual) sound sources placed between loudspeakers sound differently than sound sources placed on loudspeakers positions;
- sound sources placed between front and surround speaker are rather unstable, that is they are even more dependent on the exact position of the listener

In order to overcome these limitations is necessary a different approach: the idea is to use loudspeakers in a more efficient fashion, that take us to the state of the art of audio rendering systems through loudspeakers. A generic sound rendering system is a system that tries to position a virtual sound anywhere in space using loudspeakers or headphones [34]. In particular, a generic rendering system through louspeakers (Figure 2.3) is designed in order to reproduce a desired wavefield inside a listening area (the grey area in Figure 2.3): it is composed by an arbitrary distribution of M loudspeakers in positions $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M$. This system has the goal of reproducing a wavefield generated by a set of virtual sources located in $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_V$ using loudspeakers. With the *desired wavefield* term we mean the description of a wavefield to be reproduce with some rendering technique.

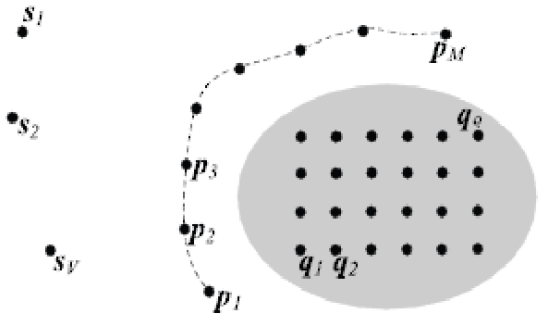

Figure 2.3. A generic rendering system model through loudspeakers.

Two main sound rendering techniques over loudspeakers exist:

- *Wave Field Synthesis* (WFS) is a sound field synthesis technique that uses an array of loudspeakers to reproduce a sound field over a large listening area. It overcomes some of the limitations of stereophonic reproduction techniques, like e. g. the sweet-spot [35]. A first concept, of what is nowadays known as WFS, was presented by Snow et al. [36] more than 50 years ago. However, technical constraints prohibited the employment of a high number of loudspeakers for sound reproduction. This technique describes sound propagation based on physical laws; it is based on Huygens-Fresnel principle [35]: each point on a wave front can be regarded as the origin of a point source [37] (Figure 2.4). For instance, we can recreate a spherical wavefront like a sum of wavefronts generated by a loudspeaker distribution.

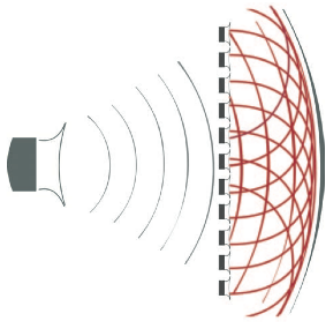

Figure 2.4. Principle of WFS: superposition of secondary sound source recreates sound-field [2]

This technique needs a large number of small and closely spaced loudspeakers form a so called loudspeaker array, in order to reconstruct the sound field within a volume [35, 38]. Each loudspeaker in the array is fed with corresponding driving signal calculated by means of algorithms based on the Kirchhoff-Helmholtz integrals and Rayleigh's representation theorems [38]. In Figure 2.5 is depicted a cinema hall in which an array of loudspeaker is implemented for the use of the Wave Field Synthesis technique.

Figure 2.5. WFS system applied for cinema [3]

- *Ambisonics* is an approach to the recording and reproduction of three-dimensional wavefields which is founded on the representation of the wavefield in polar coordinates [39]. The wavefield may be expressed as a sum of orthogonal terms with opportune polar responses, i.e spherical harmonics functions [40]. Using ambisonics we can sample the wavefront at a point in the space and starting from that reconstruct what happens around, with a configuration that is different from the way microphones are positioned. The idea is to decompose the wavefield in a series of basis function, and then get it back reproducing the basis function [41]. In practice it is common to synthesize such polar responses through array signal processing [42].

Historically first order Ambisonics recording with B-Format microphone was devised in the 1970th by Craven and Gerzon [41]. By that time the idea behind this technique was to derive from the signals of four closely spaced microphones (B-Format microphone), the pressure and the three particle velocity components at one point in space. These four signals, called B-Format signals, encode all the necessary 3D spatial information to reproduce plane waves impinging at the microphone position according to the first order 3D Ambisonics reproduction [42]. If only planar information is required (2D Ambisonics reproduction), only three microphones are necessary and the pressure and the two velocity components can be obtained directly as the output of three coincident microphones, hence the name native B-Format microphone. Nowadays first order Ambisonics recordings with B-Format microphone as well as High Order Ambisonics (HOA) [43] recordings are well founded in the theory of wavefield decomposition by means of circular and spherical harmonics functions [40].

The main difference between these two sound rendering techniques is related to the sweet-spot dimension: in Wave Field Synthesis this is the total area within the boundaries imposed by the loudspeakers, while in Ambisonics that works with spherical or circular distributions of loudspeakers, the sweet spot is in the center of the circle of the loudspeakers, and its dimension is dependent on the number of loudspeakers used. The limitations of these techniques are related to the cost of spatialization, i.e the cost of the rendering of numerous output channels [44], which can use hundreds of speakers. In addition, they are intrusive of the space, and their results are related to the number of loudspeakers used.

2.1.2 Reproduction through headphones

The classical reproduction of sounds through headphones is restricted to a listening in which the headphones are feeded with the same signal for left and right ear: this leads to a monaural listening, i.e a listening in which we have the same signal for left and right ear, and no spatialization is possible. Thus, the listener will perceive a sound internalized or near the center of his or her head. The idea of using the headphones for reproducing spatialized 3D sounds came years ago during the studies of Lord Rayleigh [45, 46] that conluded that if we were able to reproduce at the listener's eardrum the same sound pressure as in real life hearing, we could obtain realistic hearing; in particular, the fact that binaural synthesis only requires two audio channels to represent the entire three-dimensional space makes it particularly applicable for simulating a listening in a virtual environment. Headphones provide a high level of channel separation, thereby minimizing any crosstalk that arises when the signal intended for the left (or right) ear is also heard by the right (or left) ear. Headphones can also isolate the listener from external sounds and reverberation that may be present in the environment, ensuring that the acoustics of the listening environment or the listener's position in the room does not affect the listener's perception [47]. In addition, reproduction over headphones is independent from the listener's position and no sweet spot exists. On the other hand, while headphone-based systems offer potential benefits, some shortcomings exist to their use. Indeed, headphones may be uncomfortable and cumbersome to wear, especially when worn

for long periods. Additionally, unless the relevant spatial information is accounted for (inclusion of a virtual listening environment), sounds conveyed through headphones will not be properly "externalized" but will rather be perceived as originating inside the head. Finally, it is difficult to obtain good results, and the greatest risk is to obtain a sound heard inside the head, along the interaural axis, i.e the imaginary axis connecting the ears [48].

The possible of applications based on sound reproduction through headphones are a lot: if spatial sounds are properly reproduced over headphones, they guarantee a sense of realism and immersivity [49]. In particular for voice communication, spatial sound can go beyond increased realism to enhancing intellegibility and can provide the natural binaural cues needed for spatial discriminations. For all these reasons it's an interesting application for audio teleconferencing or telepresence in a meeting [49]. For music, immersive sound can go beyond reproduction through stereo set at home, giving an entire auditory image of the sound reproduction environment [50]. For environmental games, it can provide awareness of both the sound-generating objects and the surrounding acoustic space [10]. As well the advances of the computational power of consumer electronics, are making this technology available for mobile devices [9]. Furthemore, if along with auditory cues, also visual cues are provided the realism and immersivity of the listener is greatly improved [50].

2.2 Coordinate system

To specify the location of a sound source relative to the listener, we need a coordinate system. Fundamentally, spatial perception involves an egocentric frame of reference in which measurements and orientation of sound sources are given from the listener's position. Working in a 3D space, we can define three planes: the median plane is the plane which vertically cuts through the middle of the head and divides the head into right and left halves; the horizontal plane, divides the head into superior and inferior parts; the frontal plane, is responsible of front/back separation.

The head can be roughly modeled as a sphere, thus a spherical coordinate system is usually adopted. Here the standard coordinates are azimuth, elevation and range. Unfortunately, there is more than one way to define these coordinates, and different people define them in different ways.

In this thesis, the interaural-polar coordinates system [4] depicted in figure 2.6, is used.

We use the triple (azimuth θ, elevation ϕ, distance ρ) to define a point in space; indeed, azimuth and elevation descriptions indicate the perceived position only in terms of its location on the surface of a sphere surrounding the listener's head. For a more complete description the perceived distance of the sound source as another dimensional attribute is needed. Normally, azimuth and elevation are measured in degrees, where 0° degrees elevation and azimuth are at a point directly ahead of the listener, along a line bisecting the head outward from the origin point. In this coordinates system the azimuth is defined between -90 and +90 degrees. Elevation increases upward from 0 degrees to a point directly above a listener at up 90 degrees, or directly below at down 90 degrees. In figure 2.7 some values are plotted.

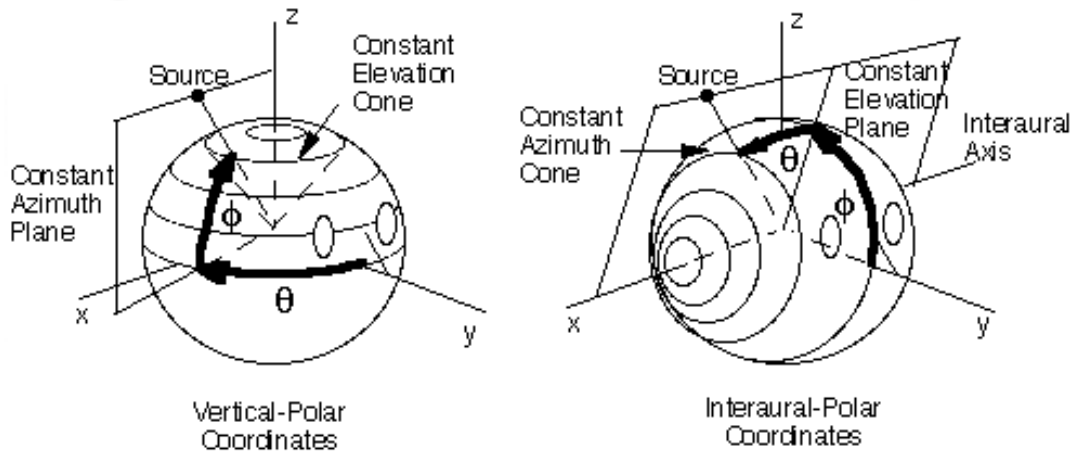

Figure 2.6. Two spherical coordinate systems: on the left the vertical-polar coordinates, on the right the interaural-polar coordinates [4].

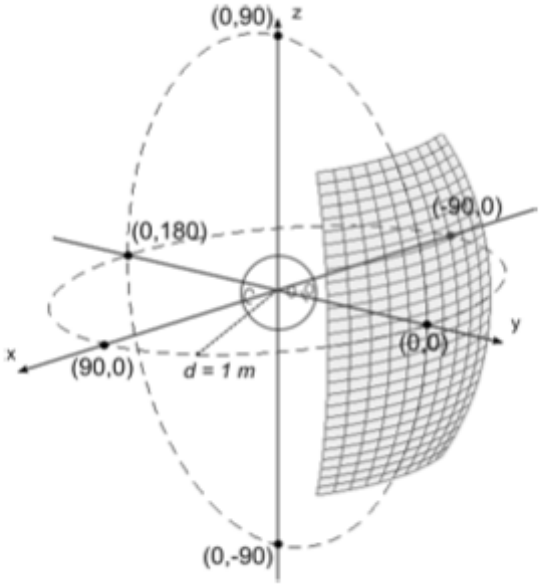

Figure 2.7. Some interaural coordinates values (six couples (θ,ϕ)) [5].

2.3 Head Related Transfer Function

In order to produce spatialized sounds over headphones is necessary to reproduce at the listener's eardrum the same sound signal he or she would hear in real life. In this context, we consider the propagation of a sound before reaching the eardrums of a listener: indeed, if we were able to take into account the alterations affecting the sound during the propagation and model them like a transfer function, then we can use it to filter the input signal and obtain the signal to be sent to the listener's ears. In this Section, we introduce the concept of Head Related Transfer Function,

that is the transfer function between the sound source and the eardrum, that is used to filter the monaural signal to obtain binaural one.
The acoustic cues for sound localization are a consequence of the physical processes of sound generation, propagation, diffraction, and scattering by objects in the environment, including the listener's own body. Being all physics processes, they can be analyzed by solving the wave equation subject to appropriate boundary conditions. In practice, the irregularities of the boundary surfaces produce extremely complex phenomena, and measuring the boundary surfaces (particularly, the pinnae) with sufficient accuracy can be challenging [49]. Analytical solutions are available only for very simple geometries. Standard numerical methods are limited by the need to have at least two spatial samples for the shortest wavelength of interest, and by execution times that grow as the cube of the number of sample points [49]. Thus, acoustic measurements are preferable over numerical methods. A suitable analysis of such measurements finally allows to separate and understand the effect of the individual body components [10].

Considering the situation depicted in Figure 2.8, we can see that the sound pressure at the eardrum is uniquely determined by the impulse response $h(t)$ from the source to the eardrum. This is called the Head-Related Impulse Response (HRIR), and its Fourier transform H(f) is called the Head Related Transfer Function (HRTF). All the information regarding the physical processes involved in the arrival of the sound to the listener's eardrum are represented in HRTF. A natural sound coming from a given direction directed to the ears will be expose to two filtering, of which the spectral and time attributes cannot be separated. As a matter of fact, in frequency domain, $X_L(f) = H_L(f)X(f)$ and $X_R(f) = H_R(f)X(f)$ where $X_L(f)$ is the signal in frequency domain at the left ear, $X_R(f)$ is the signal in frequency domain at the right ear, $X(f)$ is the source signal in frequency domain, $H_L(f)$ is the Head-Related Transfer Function for the left ear and $H_R(f)$ is the Head-Related Transfer Function for the right ear.

The HRTF enables to produce in the listener the illusion of a sound that originates at a virtual location around him. Listening to a sound signal filtered by individualized HRTFs enables to hearing that sound spatialized in an anechoic chamber. Subjects listening sounds filtered by HRTFs of other subjects show lower localization ability and report a less realistic experience.

2.3.1 Measuring the Head Related Transfer Function

Usually, a set of HRTFs is generated by measuring the Head Related Impulse Responses (HRIRs), by a time consuming procedure in an anechoic room.

Many research groups have tried to empirically measure the HRTF on human subjects or a KEMAR mannequin [51]. In such measurements, the HRTFs are usually obtained on a sphere of constant radius for a predefined set of elevation and azimuth angles. Since HRTF cannot be measured at all directions, it is important to determine the correct resolution required to achieve a sufficient spatial sampling that permits a faithful reconstruction at intermediate directions. Ajdler et al. [52] showed that an angular spacing of 5° or less in azimuth is necessary to reconstruct the data up to a bandwith of 22 kHz. For a correct reconstruction of the HRTF in elevation a less dense sampling is possible, under the 10°. Following this rule, the number of

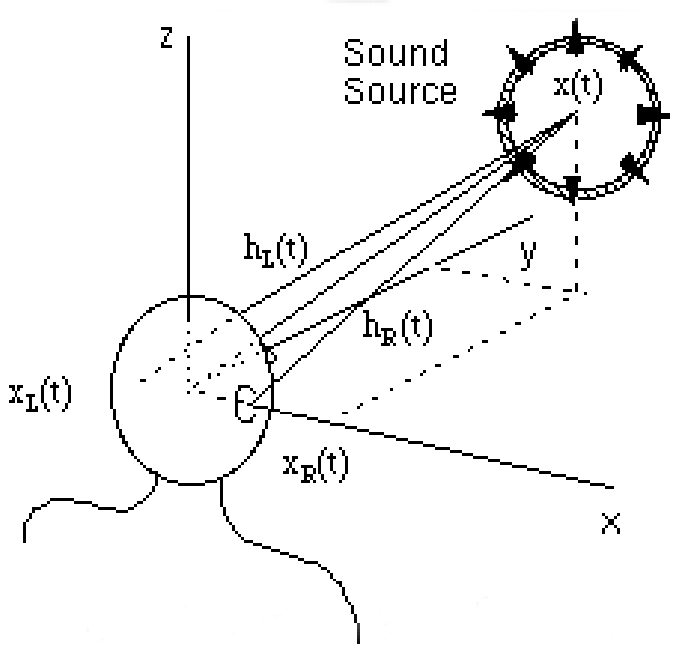

Figure 2.8. Representation of the Head Related Impulse Response, that is all we need to know the signal at the eardrum.

HRTFs in a set exceeds 1000 positions for the upper hemisphere. Because of the large number of HRTFs, the measurement duration may take tens of minutes, depending on the facilities. The problem is that during the total amount of measurements time, the subject must keep still to avoid artifacts caused by head movements. Thus, it is difficult for subjects to stay still during long HRTF measurements. Measurements result complicated due to the structure and device needed, and the time to record all the HRIRs and so, one challenge is to reduce the duration of the measurement procedure. Different HRTF measurements techniques have been proposed in time [53–55] to tackle the problem above.

We recall that the idea behind binaural rendering is that the sound pressure at the two eardrums contains all the information that is used by human listeners to elaborate they auditory perception. Therefore, if we were able to reproduce it exactly, we would also be able to reproduce the same auditory perception. In order to accurately measure the sound pressure, the recordings are commonly accomplished by using a pair of small microphones placed in the ear canals of a human listener [56] or a mannequin, at three different points:

- recording at the eardrum;
- recording at the entrance of the open ear canal;
- recording at the entrance of the ear canal, but with the ear canal physically blocked (for example with an ear plug).

Compensation has to be inserted in order to equalize the transfer function of the loudspeaker and the microphones. In addition, it has to be noted that we don't

want measurements take into account also the contribution of the ear canal (or auditory canal), because otherwise when the listener will hear a sound through heaphones it will be filtered again with his or her ear canal. So it is necessary that our measurements are missing the ear-canal resonance otherwise a compensation is also needed for the headphones. Note that these corrections are independent from direction, and merely add spectral coloration to the source [56]. After all the measuring for a person are concluded, we can store the impulse responses in a lookup table, indexed by azimuth, elevation and range.

Given the difficulty and time-consuming in measuring HRTF, another possibility has been developed: predicting the HRIR from morphology [57] based on the fact that human ears differ one from another, and so also other body parts. Some different approaches exist, based on anthropometry (i.e. the science that defines physical measures of a person's size, form, and functional capacities [58]). In particular, in order to overcome the problem of waiting for several minutes or hours before obtaining a personalized HRTF, some fast methods of spatial audio customization have been developed: these try to obtain an individualized HRTF starting from anthropometric measurements [59, 60]. On the other hand, there exist different methods that starting from some anthropometric measurements, like for example the ears as recognition element [61–64], try to extract some features in order to retrieve from an HRTFs database, the one that best matches that features [5].

HRTF is in a certain way a simplification of the reality, considering plane wave. Most HRTF measurements are made under these conditions. The far-field range dependence is easily obtained by adding the propagation delay and the inverse range dependence. The situation is more complicated when the source is distributed or is close to the head.

2.3.2 Modeling the Head Related Transfer Function

Since we know that Head Related Transfer Function is the representation in frequency of the propagation and diffraction of sounds from the source to the listener, we can capture these effects by two transfer functions, H_l (ω,θ,ϕ,ρ) and H_r (ω,θ,ϕ,ρ), that specify the relationships between the sound source and the left eardrum and between the sound source and the right eardrum, respectively. These will be two filters that vary both with the normalized frequency $\omega = \frac{2\pi f}{F_s}$ where f is the frequency and F_s is the sampling frequency, azimuth θ, elevation ϕ, and range ρ. As we have seen, obtaining correct HRTF is not an easy procedure: not always it is possible, measurements take a long time and a lot of effort and usually the procedure requires high-quality systems. Moreover, human ears change size and shape throughout life.

In order to have a cheaper and faster way to obtain the HRTF, it has been tackled the challenge of generate a generic model approximating measured HRTFs and that could be parametrizable. Some general approaches have been explored like for example the principal component analysis (PCA) [65], that describes the original data set with only a few orthogonal components and corresponding weights [66], but in particular, an approach a lot used for its simplicity [10], is to build structural models: this models coul be subdivided into submodel representing the single factors that alter the sound before reaching the eardrums.

Analyzing measured Head-Related Transfer Functions, it is possible to understand

the single contribution of each body components [10]. Obviously, it's not so simple to approximate the effects of the propagation using low-order filters. To model a correct HRTF, it's important to take into account what affects the correct localization of a sound source.

The most important cues for localizing a sound source involve the relative difference of the wavefront at the two ears on the horizontal plane [7]. The horizontal placement of the ears maximizes differences for sound events occurring around the listener [67]. Thus, we can define interaural differences as the differences between the sound reaching the two ears. In particular, we have two interaural differences: interaural time difference (ITD), that is the difference in time of arrival between the ears, and the interaural level differences (ILD) or interaural intensity differences (IID), that is the difference in loudness between the signals at the ears.

The incidence of these cues has been tested through experiments that involved the manipulation of ITD and ILD [10] and led to the discovery of the phenomenon called *lateralization*. The word "lateralized" has come to indicate a special case of localization, where [68]

- the sound is perceived and localized inside the head, mostly along the interaural axis
- the process of generating such a perception involves manipulation of interaural time or intensity differences over headphones.

Thus, manipulating the interaural differences it is possible to obtain a sound that changes in azimuth from left to right but that is always perceived inside the head.

The cues involved in the determination of direction and distance of a sound source can be grouped in coloration and interaural differences. Below, we summarize the cues, most of them discussed in details in the next sections, which are responsible for the spatialization of a sound:

- the interaural time difference (ITD)
- the interaural level difference (ILD) or interaural intensity difference (IID)
- monaural spectral cues that depend on the shape of the outer ear or pinna
- cues from torso reflection and diffraction
- the ratio between direct and reverberant energy
- cues induced by vouluntary head motion
- familiarity with the sound source

All of these cues are used by humans to localize a sound source, and for an optimum sound reproduction all of them should be considered jointly. Some of these cues are stronger and more important than others, and when two or more cues conflict, the strongest cue will often dominate [49]. However, the conflicts should be kept at bay, because they lead to confusion, causing indetermination in the localization.

It has long been demonstated that people uses dynamic cues from head motion to help localize sounds [69], which will be take into consideration in our work.

All of these cues are important for a correct localization since each affects the localization at specific frequency bands. For example, median plane directivity is governed by specular reflection from the torso at frequencies below 2.4kHz and by complex pinna phenomena for frequencies above 4kHz [70].

In order to take into account the range between the sound source and the listener, three primary cues are considered: the absolute loudness level combined with familiarity with the source, the low-frequency ILD for close sources, and direct-to-reverberant ratio for distant sources.

Head cues

The head provides the main cues for the localization of a sound in the horizontal plane [71]. Indeed, Lord Rayleigh developed the so-called *Duplex Theory* [72], according to which there are two primary cues for azimuth: he determined that the primary cue to the lateral positions of sources with frequencies greater than 500 Hz was the interaural difference in sound pressure levels (ILDs) resulting from acoustic shadowing by the head (Figure 2.9); at lower frequencies, however, the wavelength of sound is much larger than the diameter of the head, and ILDs are negligible. Rayleigh demonstrated [73] that human listeners are sensitive to interaural differences in the ongoing phase of low-frequency sounds and, thus, that interaural time differences (ITDs) could provide cues to the lateral positions of low-frequency sources. The Duplex Theory asserts that the ILD and the ITD are complementary. At low frequencies (below about 1.5 kHz), there is little ILD information, but the ITD shifts the waveform a fraction of a cycle, which is easily detected. At high frequencies (above about 1.5 kHz), there is ambiguity in the ITD, since there are several cycles of shift, but the ILD resolves this directional ambiguity. Rayleigh's Duplex Theory states that the ILD and ITD taken together provide localization information throughout the audible frequency range.

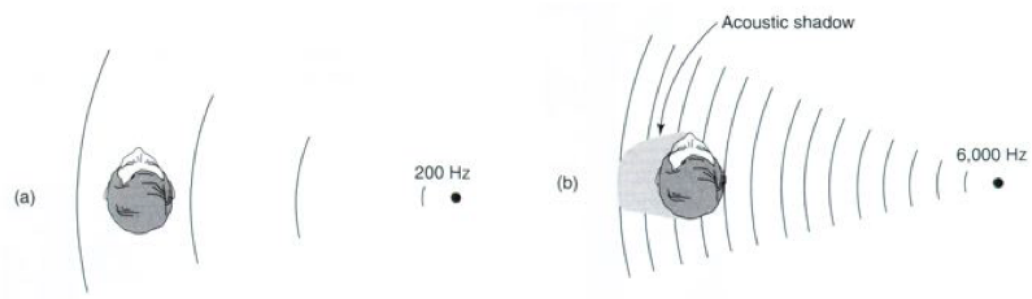

Figure 2.9. ILD acts on high frequencies in which the wavelength is comparable with the head size [6].

Lord Rayleigh, as well, obtained a frequency-domain solution for the diffraction of an acoustic wave by a rigid sphere under certain condition [7]. If we define p the free-field pressure at a distance ρ from the source, the presence of the sphere diffracts the sound wave and modifies the pressure field. We can therefore define the transfer function, that relates the pressure that would be present at the center of

the sphere in free field to that at the surface of the sphere.

Below we report some analysis extracted from [74], that explain better the range-independent approximation of the head model, i.e. the fact that we can consider the HRTF to be independent from the source distance. The authors first considered the transfer function $|H(\mu, \theta, \phi, \infty)|$ fon an infinitely distant source, where $\mu = f\frac{2\pi a}{c}$ is the normalized frequency, where a is the head radius and it is conventional to use the time $\frac{2\pi a}{c}$ that it takes for a wave to travel once around the sphere to define the normalized frequency, θ is the angle of incidence in the horizontal plane and ϕ is the angle of incidence in the vertical plane. In figure 2.10 this function is plotted at different frequencies. As you can note, *the response at low frequencies is independent from the angle of incidente* θ. For high frequencies, the response tends as the angle of incidence approaches 0° degrees, while it decreases when θ is greater than 90°.

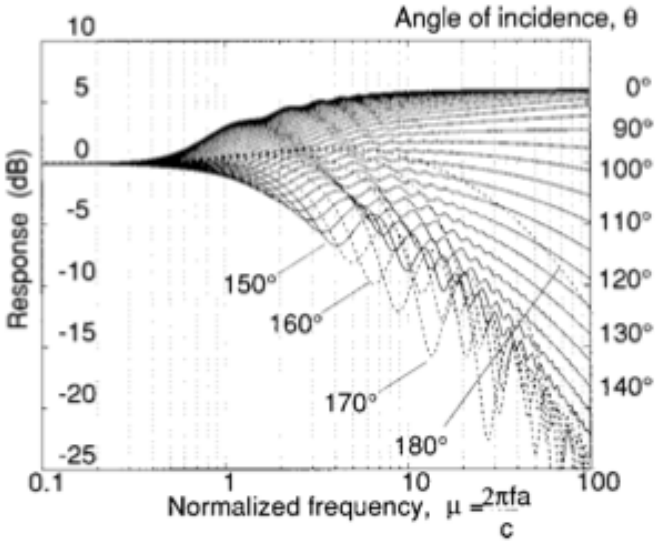

Figure 2.10. Magnitude response for an infinitely distant source. The response starts to become distinct when the normalized frequency is around 1, i.e., when the wavelength equals the circumference of the sphere. Interference effects caused by waves propagating in various directions around the sphere introduce ripples in the response that are quite prominent on the shadowed side [7].

Moreover, from Figure 2.11 we can analyze the behavior of μ at different distances ρ. In particular, for $\theta = 0°$, the response decreases as ρ increases for all the frequencies. At $\theta = 150°$, the behavior is reversed, as μ increases for increasing distances, at all the frequencies. Another general characteristic is that the difference between the responses at low and high frequencies diminishes on the near side (i.e. $\theta = 0°$) but increases on the far side. For example, when $\rho = 1.25$, the extra high-frequency rise at the front of the sphere, instead of being 6 dB, is only about 2 dB. This is consistent with the informal experience of a relative increase in the low-frequency content of close sound sources.

These two effects combined imply that the low-frequency interaural level difference (ILD) becomes even further exaggerated as the source approaches one ear.

Figure 2.12 shows that ILD at an azimuth of 100 degrees becomes very large as

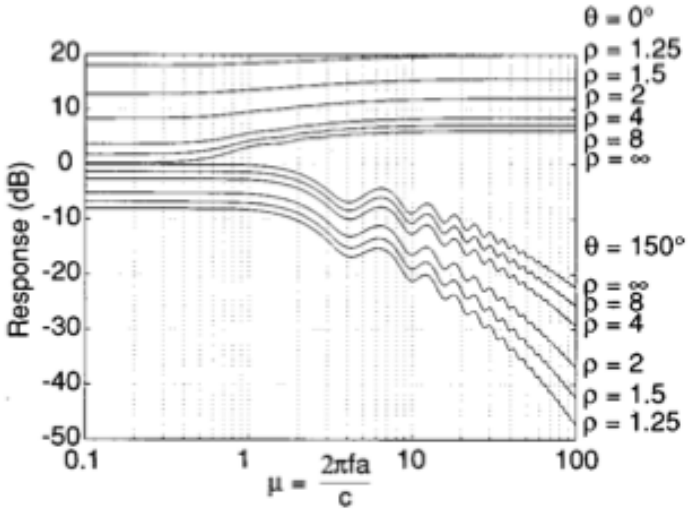

Figure 2.11. Effect of range on the magnitude response. Note that as the source approaches the sphere, the response increases on the near side and decreases on the far side. This results in the possibility of having large interaural level differences at low frequencies [7].

ρ approaches unity, even at low frequencies. This development of a large ILD at low frequencies would seem to be a major cue indicating that a sound source is very close. These observations thus confirm that the variation of low-frequency ILD with range is significant for ranges smaller than about five times the sphere radius and that the intensity of a sound is a cue for distance.

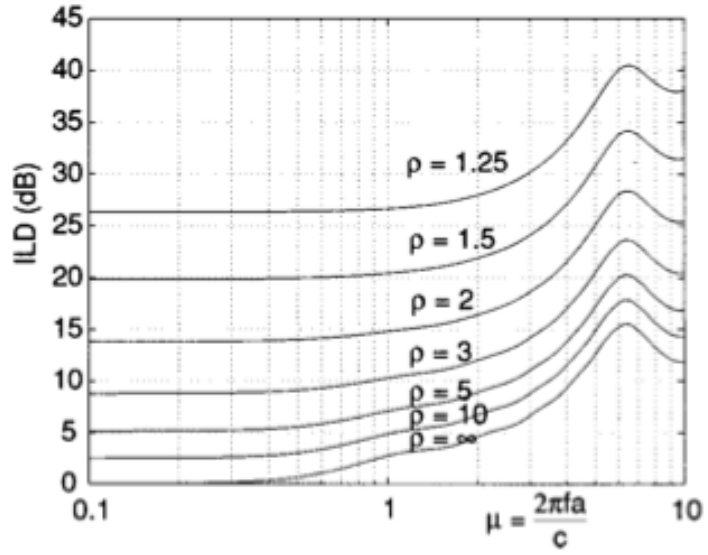

Figure 2.12. The ILD when the azimuth to the sound source is 100 degrees. Note that very substantial low-frequency ILD's occur as the source approaches the sphere. [7].

Using ray-tracing, Woodworth obtained an approximated equation for the ITD [7]. In figure 2.13 are plotted the bounds of the ITD computed following this

approximation: the upper bound corresponds to a source at the surface of the sphere, and the lower bound corresponds to a source at infinity distance. Bringing the source closer to the sphere increases the ITD of more or less 146 μs for the 8.75-cm standard head radius; Brungart and Rabinowitz pointed out that humans are insensitive to time delays above 700 μs [75], and so they hypothesized that changes in the ITD do not provide significant information about range. The results shown here support their conjecture and considering the ITD independent from the range cannot be considered an error.

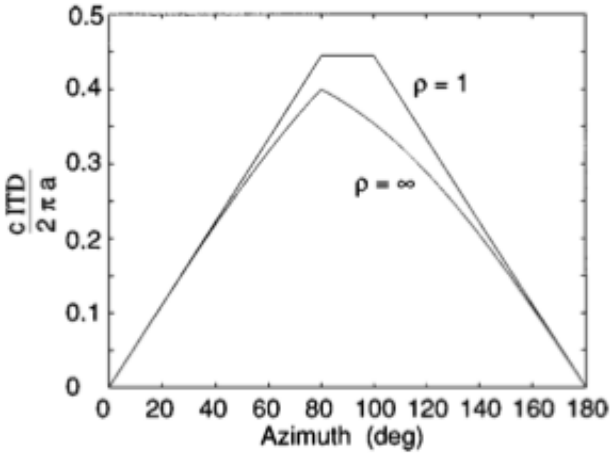

Figure 2.13. Bounds on the normalized interaural time difference computed from the Woodworth/Schlosberg formula: $\Delta\tau = \frac{c\Delta t}{2\pi a}$ is the normalized time difference between the time that the wave reaches the observation point and the time that it would reach the center of the sphere in free field where a is the head radius and c the speed of sound. In general, the ITD is not very sensitive to range [7].

Relevant work has been made [76] to investigate the role of ITD and ILD through listening tests in which the stimuli were synthesized with ITDs corresponding to one direction and ILDs corresponding to a different direction. For an ITD fixed and an ILD that varyies, localization estimate were toward the direction of the ITD as long as low-frequency components were present in the stimulus.

For simplicity, we have reasoned in simple spherical head model terms, that is an approximation of a real human head. Researches about the use of different head models have been conducted, and some alternatives have been analyzed [77], like for example the ellipsoidal model proposed in [78], where ITD varies also in elevation.

Pinnae cues

The greatest differences among different people's HRTFs are due to the massive subject-to-subject pinna shape variation [79]. As a matter of fact, the pinna plays a primary role in determining the frequency content of the HRTF thanks to two primary acoustic phenomena [5]:

- reflection over pinna edges. Sound waves are typically reflected by the outer ear, as long as their wavelength is small enough compared to the pinna dimen-

sions, and the interference between the direct and reflected waves causes sharp notches to appear in the high-frequency side of the received signal's spectrum [80].

- resonant modes in pinna cavities. In [81], Shaw argued that since the concha acts as a resonator some frequency bands of both the direct and reflected sound waves are significantly enhanced, depending on the elevation of the source.

Consequently, the part of the HRTF due to the pinna's contribution presents a sequence of peaks and notches in its magnitude. As depicted in figure 2.14, these alteration of the HRTF are dependent from the direction of arrival.

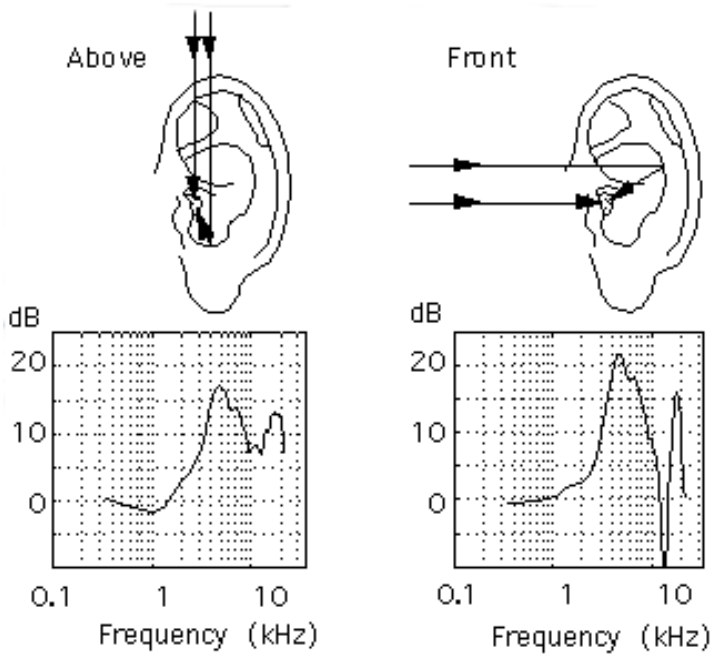

Figure 2.14. Measured frequency responses for two different directions of arrival. In each case we can see that there are two paths from the source to the ear canal - a direct path and a longer path following a reflection from the pinna. At moderately low frequencies, the pinna essentially collects additional sound energy, and the signals from the two paths arrive in phase. However, at high frequencies, the delayed signal is out of phase with the direct signal, and destructive interference occurs. [8]

The effect of pinnae is considerable around 3kHz, where the wavelength becomes comparable to the pinna size, and do not appear to contribute on sound below 3 kHz: in particular it introduces the so-called "pinna notch" within the octave from 6 to 12kHz [70]. Since monaural spectral modifications introduced by the pinnae provide the primary cues for vertical localization [70], localization in elevation requires a wide-band sources with substantial high-frequency energy.

Torso cues

The contribution of the torso is less important than head and pinna cues and tipically is linked to reflection and attenuation or shadowing of the sound [14]. In particular it operates on low frequencies components [82]. This cue doesn't provide significant information for front/back discrimination although correct low-frequency spectrum synthesis is important for a better localization: indeed, removal of the torso results in a loss of specular reflections that provide weakest elevation cues [14]. Surprisingly, it has been found [82] that when the stimuli were low-pass filtered at 5kHz, the judgements of elevations remained accurate. This contrasts the notion that elevation perception is necessarily based on higher-frequency, monaural spectral features of the HRTF. Delays introduced are maximum for sound source located above the subjects ($\phi = 90°$).

In the frequency domain the torso reflections act as a comb filter, introducing periodic notches in the spectrum.The frequencies at which the notches occurs are inversely related to the delays, and thus produce a pattern that varies with elevation (figure 2.15).

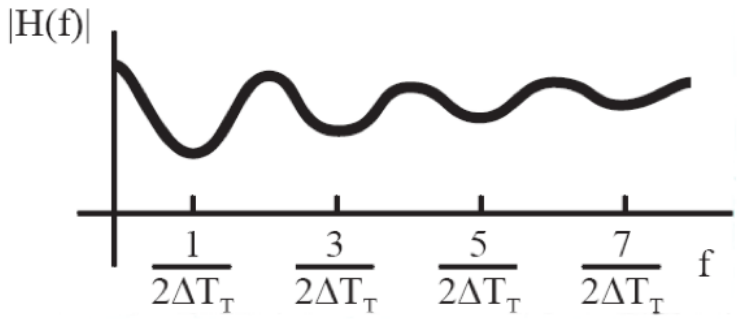

Figure 2.15. Effect of the torso in the frequency domain: the periodic notches in the spectrum are located at frequencies inversely related to the delays that varying with the elevation ϕ.

2.3.3 Localization problems

Working with non-individualized HRTFs, some problems in localization arise [10]. In particular, when we are dealing with structural models, and not measured transfer functions, we are introducing some approximations.

In this context, a common experience for the listener is the front-back confusion that results from ambiguities caused by the roughly spherical shape of the head and the primary role of ILD and ITD as localization cues [10, 83]. This problem arises due to the so-called cone of confusion, that is given by the simplification of the head model and from the choice to use an interaural system coordinates [10]. To better understand what is it, we can refert to figure 2.16.

As explained in section 2.2, the azimuth values varies from $-90°$ to $+90°$, and the difference from a point in front of or behind the listener is given by the elevation. But, since our head model is independent from the elevation, the anterior and posterior points can be confused. Indeed, we remember that, for example, a point in front of the listener is identified by $(\theta = 0°, \phi = 0°)$ and a point behind is identified

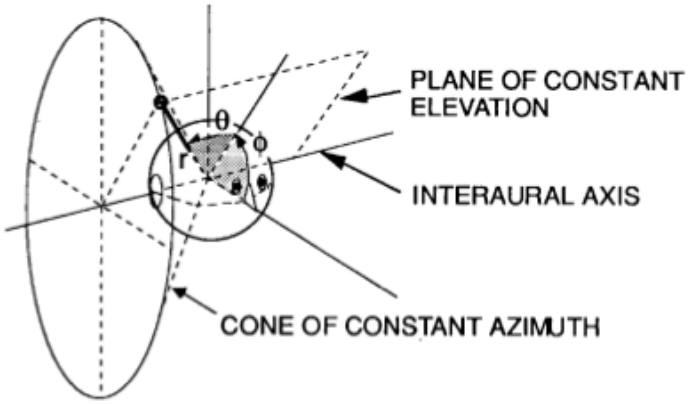

Figure 2.16. Representation of the cone of confusion: there is a cone on which ITD is constant independently from the elevation. [7]

by $(\theta = 0°, \phi = 180°)$, where elevation ϕ is measured from the horizontal plane in the usual vertical-polar system, and the azimuth θ is measured from the median plane in the interaural-polar system. Thus, a surface of constant azimuth is a cone of essentially constant ITD. Notice that, this is also dependent from the coordinates system concerned: if we used a vertical-polar coordinates system we would have a cone of confusion for the elevation, where the azimuth θ varies from $-180°$ and $+180°$, and elevation ϕ varies from $-90°$ and $+90°$ [84]. So in this case, we could have confusion with up locations heard as down, and viceversa [84]. The listener cannot separate the difference of each direction in the cone of confusion with just a difference of time or difference of level [84].

One cue thought to help in disambiguating the cone of confusion is the complex spectral shaping provided by the HRTF [10]. The notion is that the shape of stimuli spectrum at one ear, not just the interaural differences, provides information about position. For example, high-frequencies tend to be more atteanuated for sources in the rear than for sources in the front [67]. Experiments with non individualized HRTF showed that stimuli with energy concentrated at higher spectral frequency regions tended to be heard to the front, and stimuli concentrated at lower spectral frequency regions were heard to the rear [10]. In addition, detailed features of these monaural cues, such as peaks and notches in the spectrum are thought to be important for the perception of source elevation as well as front-back confusion. It has been also hypothesized that the listener can disambiguate front-back localization by tracking changes in the size and direction of the interaural cues over time [10]. In particular, Wallach demonstrated that motion cues dominate pinna cues in resolving front/back confusion [85].

Another problem, is that achieving convincing externalization, i.e. the perception of a sound not located inside the head, with headphone-based sound reproduction has proved to be a difficult challenge, particularly for sources directly in front of or directly behind the listener [86]. The most severe problem is to perceive the virtual source localized inside the listener's head [87]. This is usually called intracranial, or inside-the-head-locatedness. Externalization is related to the perception of auditory distance, and to resolve this problem is necessary to take into account the cues that

are involved in the distance perception. Hence, it's important that the listener is surrounded by an environment, in which a sense of distance to the source can be perceived [10].

2.3.4 Head motion

The listener experience is limited when the listener's head is fixed. In particular some principal localization cues are absent if the head motion is not considered [49]. As well, it is demonstrated [88] that enabling head motion in binaural synthesis dramatically improves localization even if non individualized HRTF are used, and may also increase the externalization of the sound. Mackensen et al. [89] conducted listening experiments where horizontal as well as vertical head movements were possible. The results indicate only a slight improvement when both horizontal and vertical head movements is enabled, when compared to horizonal head movement only. In [90] further listening tests were conducted and results that much more realistic simulations can be made if the changes in the sounds at the ears due to the head movements are implemented. On the other hand, head movements doesn't improve signficantly the distance perception. Also, cues due to head motion tend to dominate pinna cues when they are mutually conflicting, suggesting that information based on spectral shape is less salient that interaural cues.

Technology that accounts head motion are recent due to the computational power that has increased over the last years and the new technologies accessible to everybody; in this context, particular interest has increased due to the possibility to achieve augmented realism and versatility when the signals respond dynamically to the motion of the listener [69].

There are two possible ways of exploiting dynamic cues of head motion [49]. The first approach start from a lookup table of measured HRTF and then uses HRTF interpolation to account for head motion. The second approach, the motion-tracked binaural (MTB) [69], is based on sampling the sound field sparsely in the space around a real or virtual dummy head. MTB requires knowing the signals at multiple points around the head and uses interpolation of the signals from these microphones to account for head motion.

HRTF interpolation

When head-motion is taken into account, playback requires rapid interpolation between the entries of the tables containing the HRTFs values stored. This approach has been widely used for high-quality systems, in particular in computer games and military training systems [91]. Systems using this approach expects measurements of the HRIRs for every few degrees of head rotation and then, during playback, the signal from the head tracker is used to control an interpolator that, for each source, combines adjacent impulse responses to produce left-ear and right-ear responses that vary continously with head rotation. The results of the convolution of the source signals with the HRIRs are then fed to headphones.

Here separate signals are available for each source, the spatial locations of the sources are all known, and a head tracker is used to determine the location and orientation of the listener in the room. If the system is properly implemented, this approach produces very high-quality spatial sound. Obviously, a first quality

depending factor is the choice to measure a dummy head or to store measured HRIR. A large number of impulse responses must be measured and the error introduced by the interpolation algorithm must be unnoticeable. Finally, the combined process of head tracking, interpolation, and convolution cannot introduce detectable latency.

So, the two major issues in the implementation of HRTF-based rendering are computational cost and latency. Computational requirements depend on the complexity of the auditory scene, the allowed motion of the listener, and the efficiency of the implementation of the algorithms. To tackle this problem, approximations can be introduced. In conclusion, individualized HRIRs are used for high-performance systems, and generic HRIRs are used for consumer-grade products.

Motion-Tracked Binaural

For many applications, we would like to be able to capture a natural sound field, with no prior knowledge of the number or locations of the sources, or the structure of the acoustic environment. For this reason, and for the computational cost of the HRTF interpolation, the following method is so important.

The resulting generalization of binaural recording is called Motion-Tracked Binaural. The idea behind this technology is that one could account for head motion by sampling at additional points and interpolating. Sounds in the recording space are captured by microphones that are mounted around the diameter of a sphere or cylinder that is roughly the size of a human head. These signals can either be sent directly to the listener, or recorded for subsequent playback. The head tracker is used to control the interpolation between signals from the microphones that bridge the listener's ears.

The listener wears a head tracker, so that at any instant the system can determine the locations of the listener's ears. If the ears happen to be coincident with a pair of microphones, the signals from those microphones are directed to the listener's headphones. In general, the listener's ears will be between two microphones, and an interpolation procedure is used to determine the headphone signals.

Different applications have different requirements for spatial sampling. Also, the microphones distribution is dependent from the applications: for example they could be mounted uniformly around the equator of a sphere or uniformly distributed around a horizontal equator (panoramic), as depicted in Figure 2.17.

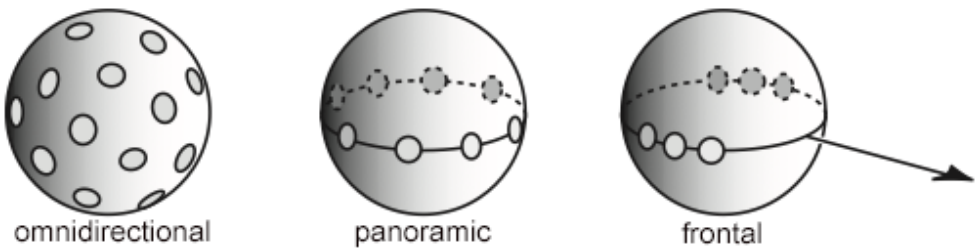

Figure 2.17. Different microphones distributions for the Motion-Tracked Binaural (MTB). [9]

The convenience of this procedure, is that signal interpolation is much simpler

than HRIR interpolation followed by convolution. However, for exact waveform reconstruction, Nyquist sampling theory requires the microphones to be no more than half a wavelength apart. If signals from adjacent microphones are directly interpolated, when the wavelength is shorter than half the intermicrophone distance, interference notches will appear in the spectrum. If r is the radius of the microphone array, N is the number of microphones, and c is the speed of sound, direct interpolation will produce deep spectral notches at odd multiples of the frequency $f_{max} = \frac{N_c}{4\pi r}$ [49]. In principle, one should have at least two samples per wavelength and so, to cover the full 20-kHz bandwidth without suffering a significant spectral notch would require distributing about 128 microphones around a typical dummy head [9]. Fortunately, exact waveform reconstruction is not necessary. The phase sensitivity needed for reconstruction is most important for the low-frequency ITD. What the authors concluded is that eight microphones produce results that are acceptable for speech, and sixteen seem to be sufficient for music. Once microphones are distributed, varying the head rotation involves making an interpolation between the output signals.

It's important to notice that however, the conversion of legacy stereo recordings through convolution leads to the same kinds of computational demands faced by HRTF-based methods, with the exception that the number of HRTFs required may be small. An alternative to real-time rendering is to perform the computations off-line for each sound source, and to store the resulting sound files for playback. A complex spatial soundscape is then created by a superposition of sounds files. Real-time computations are eliminated in exchange for an increase of the storage needed for sound files.

A general procedure for taking into account the head rotation using MTB is illustrated in figure 2.18. Instead of using two microphones in a dummy head, MTB employs an array of M microphones. The listener wears a head tracker, so that at any instant the system can determine the locations of the listener's ears. The signals from those microphones nearest the ears are interpolated and then directed to the listener's headphones.

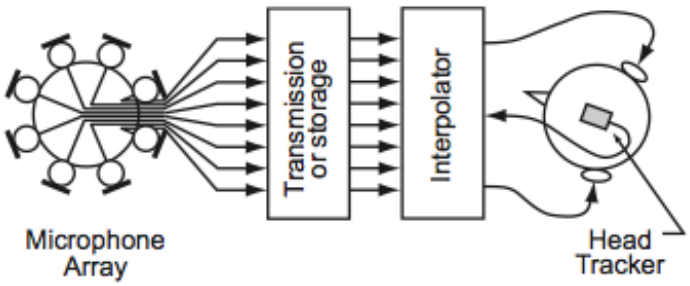

Figure 2.18. A scheme for motion-tracked binaural.[9]

Up to this point, we have assumed that physical microphones would be used to capture the sounds of interest. However, a simulated MTB microphone array can be used to "capture" the sounds of synthesized sources. Let $h_m(t)$ be the impulse response of the m-th microphone in an MTB array to a sound source S. Then if s(t) is the sound signal from the source, the microphone output is given by the

convolution $x_m(t) = h_m(t) * s(t)$. For good externalization and musically pleasant sound, $h_m(t)$ should include the effects of the room. In simple cases (such as a spherical MTB array in a rectangular room), $h_m(t)$ can be computed. Alternatively, $h_m(t)$ can simply be measured. The process of computing the microphone outputs can be computationally demanding. If there are N sources and M microphones, the process requires NM convolutions. However, this procedure provides an effective way to generate virtual auditory space, and it is particularly attractive for mixed-reality applications in which a small number of synthetic sources are combined with live or recorded spatial sound.

The HRTF approach and the MTB approach have complementary strengths and weaknesses. MTB is computationally simple, it's highly effective for live sound and faithfully captures the acoustics of the recording space. It efficiently supports multiple simultaneously head-tracked listeners in broadcast or streamed applications, and to some extent it can be individualized to specific listeners. It does not allow the listener to move around in the recording space, and it does not readily support conventional recording practices, such as the use of spot microphones. Although MTB produces highly-realistic, well externalized spatial sound, the signals produced by this method only approximate the exact experience, and critical listening tests have revealed various audible defects.

2.4 Binaural Room Impulse Response

Listening to sound filtered by individualized or non-individualized HRTF produces a sound that could be well or less well positioned in the space, but that it's in an anechoic chamber. Obviously, this isn't an usual listening environment and it's also not enjoyable. What is missing, it's the reflections of sound energy from objects in the environment that have a profound effect on the quality of the sound that we hear. In particular, the sound reflected energy is an important distance cue and when the reflected sounds are missing, the perception is that the source be very close. Then it's important to take into considerations the environment in which the sound propagates. A way to do so is to measure the impulse response in a room, including all of the early reflections and subsequent reverberation caused by multiple reflections. The impulse response represents the intensity and differences time in arrival of direct and reflected sounds. When separate measurements of the HRTF are made for each ear in a reverberant room, we are talking about the Binaural Room Impulse Response (BRIR). All the acoustical information related to the listening environment can be represented by convolving the sound with the complete BRIR.

Obviously, all the problems listed for the HRTF measurements still apply in this case, and in addition, BRIR are much longer than HRIR, so the cost of the convolution is higher.

2.5 Localization cues induced by the environments

Generally, an impulse response can be divided in three parts: the direct path contribution, the early reflections, and the reverberation tail, produced by higher order reflection whose contribute become diffusive and no more deterministic. This

is due to the fact that we are used to hear sounds in reverberant condition, where the signal that reaches our hears is the combination of the direct sound and the reflected sound, as depicted in Figure 2.19.

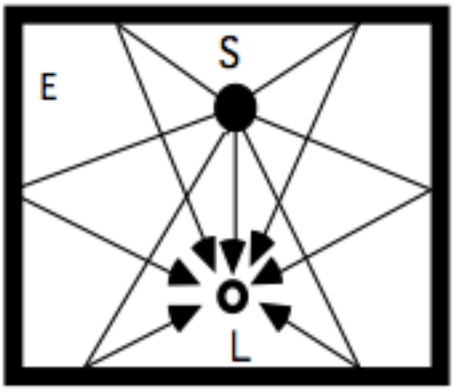

Figure 2.19. Represantion of multipath sound reaching the listener. [10]

Each part of the impulse response has a specific impact on the perception of the sound in space. This remains true for virtual spaces rendered through binaural systems. In particular, in the following we will focus on the localization cues introduced by the early reflections and by the reverberation tail.

As far as early reflections are concerned, they can give some informations about the room: in particular, the time delay with which they arrive is proportional to the distance to be covered, and so, bigger the room, bigger the delay. The amplitude of the early reflections also depend on the size of the room, and in particular it is inversely proportional to the distance traveled. In addition, the effects of early reflections can be modified by changes in the room geometry which reorder the sequence of reflections. However, we are not able to discriminate between the exact amplitude, timing and direction of early reflections [92], and so actually they don't give us precise details to room size perception. Another aspect related to early reflections is the possibility to yield a more realistic listening experience providing a "stereophonic" effect. At the same time, however, they provide also confusion in localization. So, the presence for early reflections most likely has nothing to do with localizability of sources but rather with the sense of surround which comes simply from interaural incoherence. In addition the environment in which the listener is located can affect the front-back confusion. For example, if a wall is located in front of a listener at closer proximity that the intended sound source location it has been demonstrated that the listener tends to favour the rear. Instead the early reflections which came from the same direction as the direct sound reinforce the sense of localization of the source [93].

On the other hand the reverberation tail, in particular the total level of reverberation (or the direct/reverberant sound ratio) and the length of the reverberation tail, give us strong cues about the size of the room. In most reverberant environments, the intensity of reverberation is roughly the same everywhere. Since the intensity of the sound received directly from the source varies with distance, the ratio of

direct-to-reverberant energy is a cue for distance. In addition differences between the timbre of direct and reverberant energy provides another localization cue, one that might be important for front/back discrimination as well. When we are surrounded by sound from all directions, we are not able to extract information regarding the position of the sound source from the reverberation.

Begault examined [10] the effect of artificial reverberations combined with the effects of head tracking: reduced azimuth error, but raised elevation judgements. Additionally, using the head tracking sounds externalization was much more often perceived with respect to when only reverberation was used. Localization was not affected by the late-reverb as compared using only early reflections. Front/back reversals were almost completely eliminated when head movements were allowed. Direct-to-reverberant energy ratio dominates auditory distance cues for unfamiliar sounds while intensity is the most important distance cue in speech signals. It is well known that reverberation contributes to the externalization of sound source, particularly for non-individualized HRTF. In summary, a welcome improvement in externalization can be reached with the addition of synthetic reverberation, but there is a decrease in localization accuracy. Experiments [94] showed the importance of head rotation in the horizontal plane for accurate localization. To take these results into account, the convolution in an auralization system needs to be calculated in dependence of the actual head orientation. The convolution with impulse responses is memory and processing power intense. For each relevant head position and orientation the related transfer functions need to be stored when using a databased auralization system. Thus it would be useful to reduce the stored data to a minimum required set. As found in [95], a reduction of dynamic cues doesn't translate into audible changes in room impression.

2.6 Room acoustics simulation

A primary challenge in acoustic modeling is computation of reflection paths from a sound source to a listener (receiver). As sound may travel from source to receiver via a multitude of reflection, transmission, and diffraction paths, accurate simulation is extremely computational intensive. For instance, consider the simple example shown in figure 2.20. In order to present an accurate model of a sound source (labeled 'S') at a receiver location (labeled 'R'), we must account for an infinite number of possible reflection paths (some of which are shown). The knowledge of reflection paths, along with the reflective properties of the walls and the directivity functions of the source and the receiver, lead to an accurate prediction of the impulse response from S to R.

The challenge is further complicated when multiple sound sources are present, since psychoacoustic effects are involved. Within nonanechoic environmental contexts, sound arrives to a listener by both direct and indirect paths. In this case, the sound source can be said to arrive at the listener as a diffuse field, due to the effect of the environmental context [10].

If we play the same auditory stimulus in different environments, we would perceive different auditory responses, depending on the geometry of the environment, on the position of the listener, and on the position of the source. Room acoustic modeling is

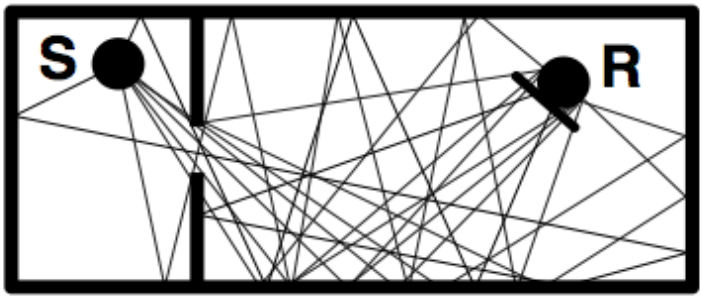

Figure 2.20. Example of reflection paths. [11]

often done using geometric acoustic method like ray-tracing, beamtracer or otherwise image-source method. Two ways of auralizing virtual or real listening environments via headphones are commonly in use: model based and data based systems. The first group uses a more or less complex model to reproduce the acoustical specifications of a room. With higher complexity of the model used, the quality of the results rises, but the processing power needed increases rapidly. In contrast, a data based system will store either "pre-synthesized" impulse response or measured data of real existing room. The ear signals are produced in the same way in both systems, namely a convolution of an impulse response with the incoming audio signal.

The reverberation of a room can be taken into account measuring the impulse response of the empty room implementing then by a filter, otherwise it's possible to realize an artifical reverberator through different approaches. A physical model reverb using, for example, the waveguide approach can be used and it has some advantages, but there is a computational problem: they are too expensive for real-time computation. A simplified approach is to construct a perceptual model: this is based on the fact that we do not perceive the full complexity of reverberation, but only the most important perceptual aspects. Thus, a reverberant circuit can be synthesized which is able to produce an acoustically and perceptually indistinguishible natural reverb. This approach is much more efficient and it has the advantage to be parameterized and controllable.

2.7 Rendering systems on heaphones

Many virtual sound systems over headphones exist in the literature. Here, we introduce some of them, that are worth to be mentioned.

- *Holophonics* is a binaural recording system created by Hugo Zuccarelli [96] This recording technology doesn't use traditional microphones, but uses a sound processing technique that captures the full spectrum of information travelling from the ear to the brain in the recording environment. During playback, the information recorded reaches the auditory cortex of the listener's brain and recreates the same sensation as if one was listening to the original event [97]. In this technology, it isn't possible to divide the information about the direct sound and that of the environment information. As his inventor stated, 'it provides for the auditory system what laser holography provides for

the visual system and thus can be considered holography of sound' [97]. It has been used in some Stevie Wonder and Lionel Ritchie albums.

- *Rondo Player* is a music player designed for headphones, developed from Dysonics [98], designed for mobile communication systems. Based on researches summarized in [50], it reproduces the sounds positioned in spatial location, in a panoramic configuration, adding the capability to respond to the rotation of the head, using the gyroscope of the mobile phone. In fact, the system places the listener in the sweet spot of a virtual reproduction system with two loudspeakers, and it uses the orientation of the listener's head to experience the sound playing over a pair of loudspeakers. In particular, there is a version called Rondo Motion that intergrates a wireless motion sensing on the headphones to capture the rotation of the head [99]. The system allows the user to choose among three listening settings (front row, stadium concert, middle of a club concert) in which the listener position is fixed [100].

- *QSound Lab* has developed some technologies for videogaming [101], in which an appareant location of a sound source is controlled in azimuth and range by the user, through a range control block that has variable amplitude scalers and a time delay and by an azimuth control block that also has variable amplitude scalers and time delays. The values of the scalers and the various delays are read out of look-up tables in a controller that is addressed by an azimuth index value corresponding to any location on a circle surrounding the headphone wearer. Several range control blocks and azimuth control blocks can be provided depending on the number of input audio signals to be located [102]. An appareant moving sound source location can be reproduced since, for example, video games involve video movement with an accompanying sound program in which the apparent sound source also moves.

- *Dolby Headphones* tries to create the sensation of multiple loudspeakers in a room [103]. Principally integrated in headphones for gaming, it gained partnership from multiple brands like Nokia that supports this technology. In addition, it can be included in almost any device that can process stereo or multichannel audio and has an headphone output. It can be embedded into DSP chips or implemented in software for use in A/V receivers and preamp/processors, TVs, and PCs. The advertising slogan states that 'Dolby Headphone technology provides a real benefit to your customers—the ability to put on any pair of headphones and experience 5.1-channel surround sound' [103].

Chapter 3

System Description

A simpler and more flexible system can be achieved using a structural model instead of measuring the HRTF. A structural model consists in a combination of elementary submodels, each representing different parts of a general human body. In particular, each submodel takes into account specific localization cues, which contribute obtain a synthetic approximation of the HRTF [10]. This approach has some drawbacks: it can be affects by front/back reversal i.e azimuth errors between the front and rear hemispheres, and a possible missing sense of sound externalization, i.e the possibility to hear the sound inside the head [15]. An aspect to be taken into account is that not only the localization is important, but also a sense of realism and immersivity [69].

A less individualized, but more computationally efficient implementation using a model-based HRTF is presented in this Chapter, taking into account all the localization cues needed for a correct sound localization. In order to give a listening experience in which are present both a sense of immersivity that a capability to delocalize sounds in space, the system implemented takes into account not only the direct but also the reflected sounds, simulating the experience of hearing in a room. By its nature, binaural techniques can capture and reproduce most of the sound attributes of a specific location in a specific venue. The goal of our system is to solve the front/back reversal and externalization problems. In order to overcome the limitations that arise with the using of a non-individualized HRTF, a system that integrates the capability to rotate the virtual listener's head is taken into account. Indeed, by providing frontal externalization, removing front/back confusion and stabilizing the sound field with head rotation, also a structural model can enhances greatly the listening experience.

The system that we present can be used to transfom any sounds into the format presented for any subsequently play back in order to reproduce spatial effects, or it can be used in real-time like for instance a spatial music player.
In this Chapter, we present the implemented model: it produces horizontal and externalization cues. Furthemore, we take into account the possibility to rotate the virtual listener's head.

In the first Section we present the structural model used, explaining what localization cues are taken into account and what bandwith of sound each submodel alters. For this reason we treat the head model, that introduces azimuth localization, providing the well-known ITD and ILD cues; the torso model is also considered,

and we specify geometric simplification about its form and the dependence on the elevation, and the pinnae model is introduced, well-known for its contribution in elevation.

To address the externalization of the sounds i.e. the perception to hear sounds outside the head, in Section 3.2 we introduce the needs of a listening environment that is necessary in order to reach this goal when working with non-individualized HRTFs, what are the main characteristics that affects the hearing and how taking into account the reflections. The room model introduces early reflections to provide externalization, and is noteworthy mostly for its extreme simplicity. In particular in each reflection a new wavefront is created, and then reflection can be modeled like a new sound source. We also present an overview of the fast beamtracing method [104] used to model the propagation of the wavefield as acoustic rays, following the laws of geometrical acoustics.

In Section 3.3, we present the approach used to consider the possibility for the user to turn the virtual head.

Finally, in Section 3.4, we describe some our decisions about the implementation of the model.

3.1 HRTF structural model

We have chosen to use a structural model for the HRTF, even if, as we saw in Chapter 2, working with non-individualized HRTFs some problems arise. To mitigate these side-effects, we introduce the possibility of rotating the virtual head of the listener, which has been proved to improve the ability to correctly localize the source [49].

The general structural model we use to generate a spatial sound is depicted in figure 3.1: in the following we present each submodel.
The output obtained from the system is azimuth dependent, i.e. according to the orientation of the head the input is filtered with an appropriate set of filters. Thus, what we describe in the following is the procedure whereby starting from a monaural input we obtain a binaural output.

It has been shown [12] analyzing the HRTFs of individuals with particularly protudring ears, that an accentuated spectral difference is presented between symmetric about the median plane locations. This could reinforce the cues in solving the cone of confusions with respect using HRTFs of individuals with "normal ears". In particular this study shows that an additional shadowing of some frequency bands could aid in back localization. A comparison of HRTFs with small protrusion angle to those with large protrusion angle revealed attenuation in selected frequency bands for some azimuths in the posterior of the head. In contrast, a slight boost in sound level was observed in the low frequency range for the frontal hemisphere. Following these results, the implemented system provides spectral modifications of the sound to be rendered, defined as a function of the azimuth. Therefore, the monaural input is pre-filtered with filters shown in Figure 3.2 before being passed to the structural model. Since no analytical references to the filters there exist, we visually sample some points from the plots and then we use an interpolation method to obtain the remaining values in frequency. An inverse Fourier transform is then applied to obtain the filters in the time domain.

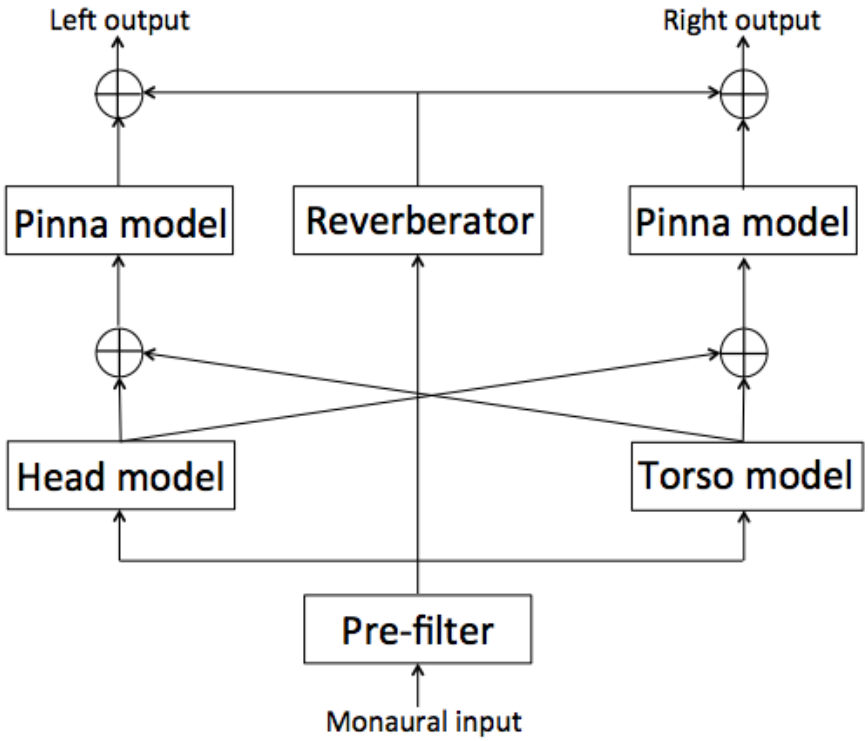

Figure 3.1. The model used in this thesis for generating a spatial sound.

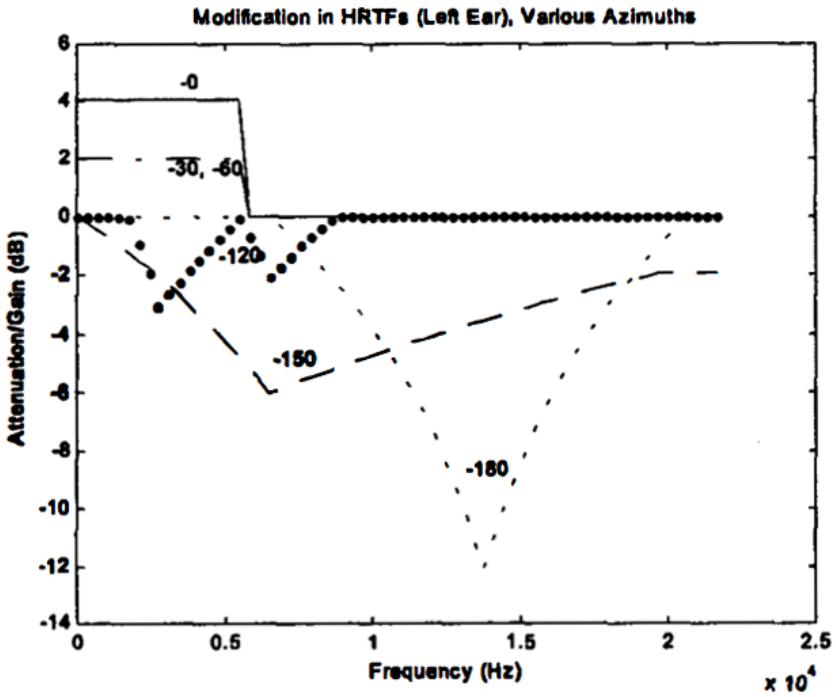

Figure 3.2. Gains and attenuations introduced by protruding simulated "ears" [12].

3.1.1 Head and torso model

Here, an approximate snowman model, depicted in figure 3.3, in which the body is approximated by a spherical head directly above a spherical torso, is taken into account, in order to understand the effect introduced by the head and the torso. The snowman model is defined by three parameters: the head radius a, the torso radius b and the neck height h.

Figure 3.3. A snowman model, composed by two spheres of different radius: this is an approximation that permits to better understand the diffraction of the sound and simplify the behavior. [13]

A simple spherical model is used for the head. We know [10] that using such model the cone of confusion is introduced. However, this side effect is almost negligible thanks to the introduction of a rotating virtual head [69]. The HRTF for the sphere is obtained from Rayleigh's infinite series solution [74] to the equations for the diffraction of sound by a sphere. To compute the transfer function from the source to the ear, two quantities are needed: the angle of incidence and the head radius a.

The structural filter model that we use for taking into account the ILD, i.e. the differences in intensity levels, is based on the results showed in [5]. In particular, the filter proposed is computationally simple, and it's the bilinear transform of the filter described in [7].

$$\mathbf{H}_{head}(z) = \frac{\frac{\beta+\alpha f_s}{\beta+f_s} + \frac{\beta-\alpha f_s}{\beta+f_s} z^{-1}}{1 + \frac{\beta-f_s}{\beta+f_s} z^{-1}} \tag{3.1}$$

where f_s is the sampling frequency, β depends on the head radius parameter a as $\beta = \frac{c}{a}$ where c is the sound speed, and α is defined as

$$\alpha(\theta_{inc}) = 1 + \frac{\alpha_{min}}{2} + (1 - \frac{\alpha_{min}}{2}) \cos(\frac{\theta_{inc}}{\theta_{min}}\pi) \tag{3.2}$$

Note that θ_{inc} is the incidence angle that, assuming the interaural axis to coincide with the x-axis for sake of brevity, relates to azimuth θ as $\theta_{inc} = 90° - \theta$ for the right ear and $\theta_{inc} = 90° + \theta$ for the left ear. A good approximation is heuristically found for parameters $\alpha_{min} = 0.1$ and $\theta_{min} = 180°$.

With regard to the propagation delay introduced by the head, this is computed by the formula $\Delta t = (\frac{a}{c})(\sin\theta + \theta)$ where depending on the azimuth θ the delay is applied to the right or to the left ear. In this formula a is the head radius, and c is the speed of sound.

Although the human torso does not have a regular shape, it can be approximated by a simple ellipsodial model [13], based on analytical simplicity and its small number of parameters (height, widht, depth). An ellipsoid is a closed quadratic surface that is a three-dimensional analogue of an ellipse [105]. The standard equation of an ellipsoid centered at the origin of a Cartesian coordinate system and aligned with the axes is

$$\frac{x^2}{R_1^2} + \frac{y^2}{R_2^2} + \frac{z^2}{R_3^2} = 1 \tag{3.3}$$

where R_1, R_2 and R_3 represent the length of the semi-axes.

Although an ellipsoid fits the human torso better than does a sphere, the geometry of a spherical torso is easier to analyze [14]. The effect of the torso is twofold as it is possible observe in Figure 3.4 and it's dependent by the sound source position: it can provide both reflections and shadowing of the sound. Indeed, as the source descends in elevation, a point of grazing incidence is reached, below which torso reflections disappear and torso shadowing emerges. Rays drawn from the ear to points of tangency around the upper torso define a cone that we call the torso-shadow cone within which no reflection appear [13]. So the model switches between a torso-reflection behavior (figure 3.4 a) when the source is outside the torso-shadow cone and a torso-shadow sub-model (figure 3.4 b) when the source is inside the torso-shadow cone. In order to determine the transition condition, we consider $\bar{s}$ to be a unit vector pointing in the direction of the infinitely distant source, and $\bar{d}$ the vector of length d from the center of the torso to the ear (note that the analysis has to be done separately for each ear). The separation condition between the two different effects is given by a ray from the source to the ear tangent to the torso and it is depicted in Figure 3.5: in this case the projection of the vector $\bar{d}$ is given by $\sqrt{d^2 - b^2}$. Thus, the source is inside the torso-shadow cone if $\bar{d} \cdot \bar{s} < -\sqrt{d^2 - b^2}$ and it is outside if $\bar{d} \cdot \bar{s} > -\sqrt{d^2 - b^2}$.

Regarding the torso-reflection, the problem is to compute the point on the surface of the sphere where the reflection will occur, and use it to calculate the difference in path lengths to the ears for the direct and the reflected sound waves. So, the torso introduces an additional version of the sound that is delayed and attenuated: this corresponds to the introduction of some notches in the HRTF. Note that the reflections are delayed by a time delay that varies by elevation and not by azimuth: indeed if the sound source moves on a circumference in the horizontal plane (showed

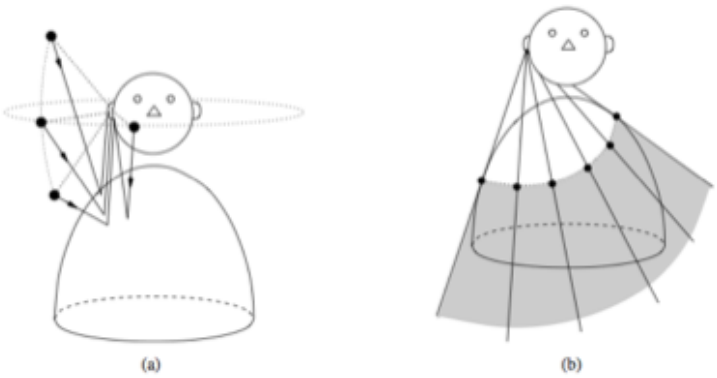

Figure 3.4. Representation of the two main effects of the torso, dependent on the position of the sound source. Sound reflections appear when the source is outside the shadow cone and sound shadowing for sources below the listener.

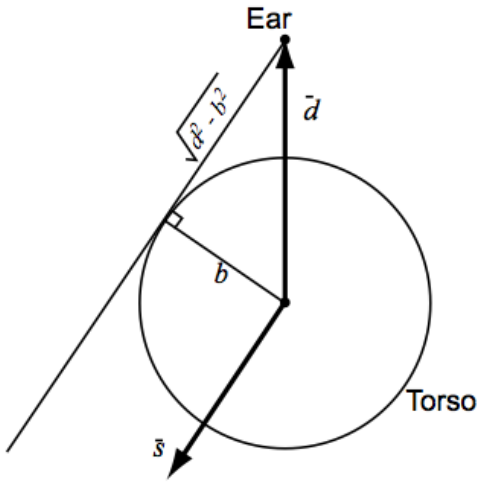

Figure 3.5. The geometry for a tangent ray to the torso [14].

in figure 3.4 a) the delay is negligible [106]. In particular, the reflected pulses are maximally delayed for sound source locations right above the listener (figure 3.6).

In order to derive the filters that represent the effect of the torso, we follow the work done in [14], where it is considered only the behavior in the frontal plane. The torso reflection model assumes that at the ears arrive two components: a direct component and a reflected component that arrives after being reflected from the torso. The direct component is influenced by the head, while the reflected component experiences a further attenuation because of the torso reflection coefficient γ and an additional propagation delay τ because of the greater length of the reflected path. For simplicity we assume a torso reflection coefficient $\gamma = 0.3$ constant, following the choice made in [14]. The reflected component arrives at the head from a different direction than the direct component, resulting in a different observation angle θ_R. The observation angle is the angle between a ray from the center of the sphere to the sound source and a ray from the center of the sphere to an observation point P

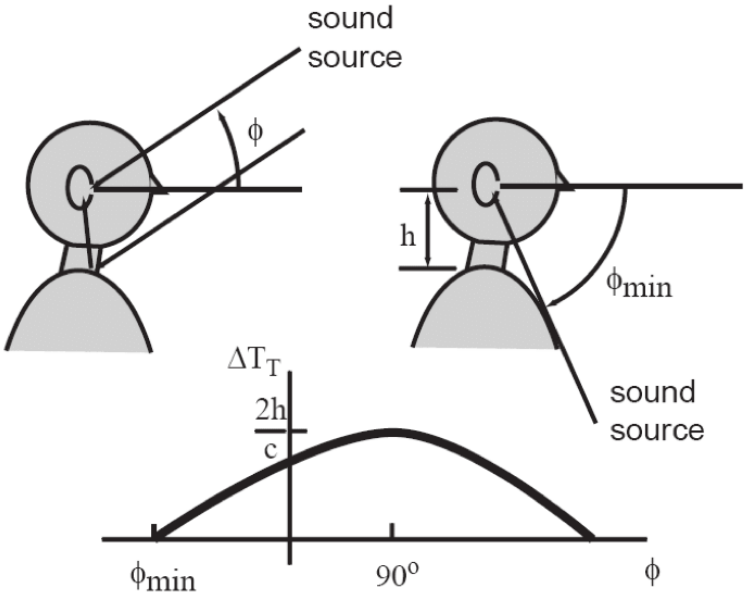

Figure 3.6. Plot of the delay values introduced by the torso: it varies with the elevation and its maximum is for sound source ahead the listener.

on the surface of the sphere. For an HRTF model, the point P represents an ear location.

The analysis carried out in [14], use a ray-tracing argument to derive formulas for the angle of arrival and the additional propagation delay τ. Using this analysis we want to find $\bar{b}$ that is a vector of length b from the center of the torso to the point of reflection, given the vectors $\bar{s}$ and $\bar{d}$.

Observing the Figure 3.7, it is possible to note that we can specify the vector $\bar{s}$ in terms of the angle ζ from $\bar{d}$ to $\bar{s}$ or in terms of the complementary angle ϵ given by

$$\epsilon = \frac{\pi}{2} - \zeta = \frac{\pi}{2} - \arccos\left(\frac{\bar{d}\cdot\bar{s}}{d}\right). \tag{3.4}$$

Given the torso radius b, the vector $\bar{d}$ and the vector $\bar{s}$, we can then find the amount of time $\tau = \frac{d_f}{c}(1+\cos(2\psi))$, where $\psi = \xi + \beta$ is the sum of the angle ξ between the vector $\bar{b}$ and the vector $\bar{d}$, and the angle β between the reflected ray and the vector $\bar{d}$, and d_f is the distance from the point of reflection to the ear. It is possible to define the angle $\beta = \arctan(\frac{\sin(\xi)}{A-\cos(\xi)})$ in terms of the angle ξ, where A is the ratio $A = \frac{d}{b}$. The distance d_f from the point of reflection to the ear is obtained using the *law of cosines* and is given by $d_f = \sqrt{b^2 + d^2 - 2bd\cos\xi}$. In order to obtain ξ it is exploited the fact that at the point of reflection the angle of incidence must equal the angle of reflection: this links ξ and ϵ. Since it is not possible to obtain an analytic solution for ξ as an explicit function of ϵ and A, the equation is found numerically. A linear approximation of ξ is found to be

$$\xi \approx \begin{cases} \xi_0 - (1 - \frac{\xi_0}{\xi_{max}})\epsilon, & \text{if } -\xi_{max} \le \epsilon \le 0 \\ \xi_0(1 - \frac{\epsilon}{\frac{\pi}{2}}), & \text{if } 0 \le \epsilon \le \frac{\pi}{2} \end{cases} \tag{3.5}$$

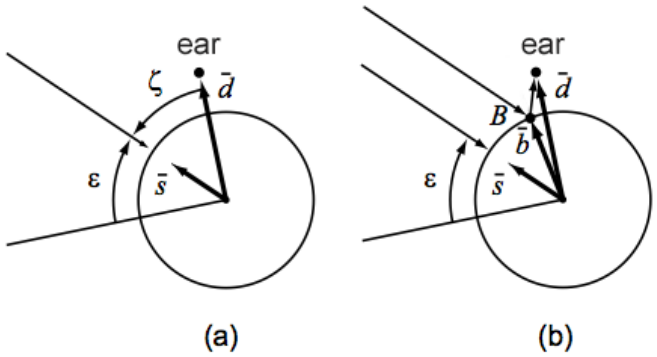

Figure 3.7. Normal view of the plane defined by the vector $\bar{d}$ from the center of the torso to the ear and the vector $\bar{s}$ pointing to the source. In (a) is shown the elevation angle ϵ. In (b) is represented the vector $\bar{b}$ to be found. [14]

where $\xi_0 = \frac{A-1}{2A-1}\frac{\pi}{2}$ is the value for $\epsilon = 0$ and $\xi_{max} = \arccos\frac{1}{A}$ is the maximum value of ξ reached when the reflected ray is tangent to the torso sphere.

Finally, we need to derive the elevation angle θ_R, that is the angle of arrival at the head after the reflection. In order to compute this angle, we need to specify the unit vector $\bar{r}$ that points in the direction of the incoming reflected wave, where $\bar{r} = \frac{\bar{b}-\bar{d}}{\|\bar{b}-\bar{d}\|}$. Then the angle is equal to $\theta_R = \arccos\frac{\bar{d}\cdot\bar{r}}{a}$, and we can use this angle to compute the filters for the head effects using the formula seen before.

When the source is in the torso-shadow cone, waves from the source must travel around the torso before reaching the head. The physical situation is rather complex, with wave components that take different paths around the torso to the ear traveling different distances. We approximate this behavior by assuming that all the components are first shadowed by the torso acting as an isolated sphere and that these components all arrive at the ear at some effective head angle θ_H. This results in the simple cascade of a torso-shadow filter and a head-shadow filter as represented in Figure 3.8.

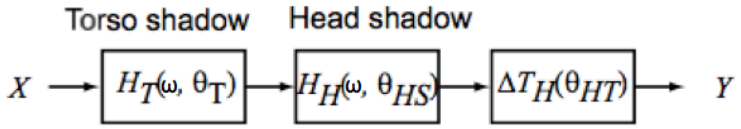

Figure 3.8. The torso shadow sub-model. Here the filters are represented depending only on the frequency ω and the azimuth angle. [14]

Here the basic goal is to find the vector $\bar{b}$ from the center of the torso to the point of tangent incidence. Once this vector is determined, the vector $\bar{r}$ is obtained at once from the equation $\bar{r} = \frac{\bar{b}-\bar{d}}{\|\bar{b}-\bar{d}\|}$ where the unit vector $\bar{r}$ points in the direction of the incoming wave. We can write the vector $\bar{b}$ in an orthogonal decomposition form

$$\bar{b} = \eta_1\bar{d} + \eta_2\bar{d}_2 \tag{3.6}$$

where $\bar{d}$ is the vector to the ear, and $\bar{d}_2$, is the vector orthogonal to $\bar{d}$. Using the two conditions (1) the vector $\bar{b} - \bar{d}$ must be orthogonal to $\bar{b}$ (tangent incidence), and (2) the length of $\bar{b}$ must be the radius b of the torso, we can determine the coefficients η_1 and η_2. From condition (1) we have $(\bar{d} - \bar{b}) \cdot \bar{b} = 0$ or $\bar{d} \cdot \bar{b} = b^2$. But because $\bar{b}$ and $\bar{d}_2$ are orthogonal, we obtain $\bar{d} \cdot \bar{b} = \bar{d} \cdot (\eta_1 \bar{d} + \eta_2 \bar{d}_2) = \eta_1 d^2 = b^2$ so that

$$\eta_1 = (\frac{b}{d})^2. \tag{3.7}$$

From Eq. 3.6, the requirement that $\|\bar{b}\| = b$, and the orthogonality of $\bar{d}$ and $\bar{d}_2$, we have $\|\bar{b}\|^2 = \eta_1^2 d^2 + \eta_2^2 \|\bar{d}_2\|^2 = b^2 = \eta_1 d^2$ so that

$$\eta_2 = \pm\sqrt{\eta_1(1 - \eta_1)}\frac{d}{\|\bar{d}_2\|}. \tag{3.8}$$

Since $\bar{d}_2$ can be written as

$$\bar{d}_2 = d^2\bar{s} - (\bar{d} \cdot \bar{s})\bar{d}, \tag{3.9}$$

we have $\|\bar{d}\|^2 = \|d^2\bar{s} - (\bar{d} \cdot \bar{s})\bar{d}\|^2 = d^4 - 2d^2(\bar{d} \cdot s)^2 + (\bar{d} \cdot \bar{s})^2 d^2 = d^2(d^2 - (\bar{d} \cdot \bar{s})^2)$. Substituting this in Eq. 3.8 leads to

$$\eta_2 = \pm\sqrt{\frac{\eta_1(1 - \eta_1)}{d^2 - (\bar{d} \cdot \bar{s})^2}}. \tag{3.10}$$

Thus, we obtain two solutions for $\bar{b}$, one for the case where the source is on the ipsilateral side of the torso and one for the contralateral side. From Eq. 3.9, it is clear that $\bar{d}_2$ always points to that side that $\bar{s}$ is on. Thus, we always choose the positive sign, and the direction of $\bar{s}$ will automatically determine the proper solution. Now that we have the desired unit vector $\bar{r}$ from d to b, we can obtain the observation angle θ_{HS} for the head, that is $\theta_{HS} = \arccos \frac{\bar{d} \cdot \bar{r}}{a}$. Thus, we can use this angle in the Equation 3.2 in order to obtain the shadow effect of the head.

In [14] the angle θ_{HT} is obtained considering the torso absent for simplicity, and it is $\theta_{HT} = \arccos \frac{\bar{d} \cdot \bar{s}}{a}$. Thus the filter results to be

$$\Delta T_H(\theta_{HT}) = \begin{cases} -\frac{a}{c}\cos\theta_{HT}, & 0 \le |\theta_{HT}| < \frac{\pi}{2} \\ \frac{a}{c}[|\theta_{HT}| - \frac{\pi}{2}], & \frac{\pi}{2} \le |\theta_{HT}| < \pi \end{cases} \tag{3.11}$$

For the torso-shadow filter H_T, the problem is to determine the observation angle θ_T for determining the torso shadow. From Figure 3.9 we note that we can use the angle ζ between $\bar{s}$ and $\bar{d}$. However this choice produces a discontinuity at the boundary of the torso shadow cone.

To reduce this discontinuity we compute θ_T by interpolating between θ_{flat} and π, which is the value of ζ when the source switches from the ipsilateral to the contralateral side (Figure 3.9). The critical angle θ_{flat} is computed for $\alpha = 1$, where the zero and pole of the shadow filter cancel, and the frequency response is flat, and it is given by

$$\theta_{flat} = \theta_{min}[\frac{1}{2} + \frac{1}{\pi}\arcsin\frac{\alpha_{min}}{2 - \alpha_{min}}] \tag{3.12}$$

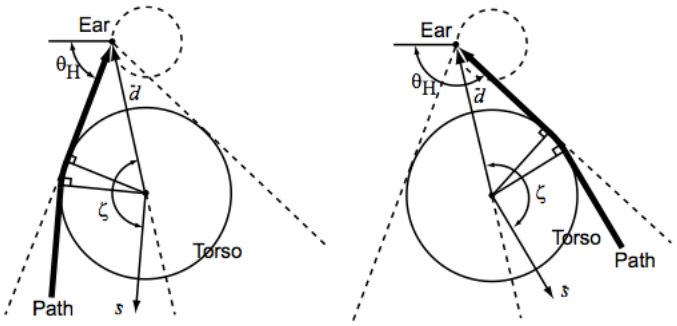

Figure 3.9. The dominant paths from the source to the ear when the source is in the torso-shadow cone. The filter model assumes that all of the energy arrives at the observation angle θ_H. Note that θ_H is smaller in (a) when the source is on the ipsilateral side than it is in (b) when it is on the contralateral side. [14]

To be specific, let ζ_{min} be the value of ζ at grazing incidence. Then $\zeta_{min} = \frac{\pi}{2} + \arccos\frac{b}{d}$, and interpolation yields

$$\theta_T = \frac{\pi(\zeta - \zeta_{min}) - \theta_{flat}(\zeta - \pi)}{\pi - \zeta_{min}} \tag{3.13}$$

The following parameters are used for the snowman model in our implementation: the head radius $a = 11$ cm, torso radius $b = 16.9$ cm, neck height $h = 5.3$ cm.

3.1.2 Pinna model

It has been acknowledged that the pinna introduces notches into the signal at frequencies dependent from the angle of incidence. We follow the work done in [15] and in the following we summarize the most important concepts.

To model the pinna, we use the schema in figure 3.10 where ρ_k represents the attenuations and τ_k the time delays associated with the k^{th} peak or notch.

It has been found through listening experiments that considering more than 5 peaks or notches the improvements are not so perceivable. In addition, another simplification is introduced: the amplitudes are considered to be constant and not varying with the azimuth and elevation. Examining measured HRIR the authors of [15] observed that the time delays could be well approximated as:

$$\tau_k(\theta, \phi) = A_k \cos(\frac{\theta}{2}) \sin(D_k(\frac{\pi}{2} - \phi) + B_k) \tag{3.14}$$

In particular, as the authors stated, this is a generic model in which only the parameters D_k should be adapted to individual listeners. In the table in figure 3.11 are listed the numerical values for the parameters: A_k and B_k are given in samples at 44.1 kHz sampling rate, D_{k1} is adapted for two examined subjects, and D_{k2} is for the third. In our model we choose to use the values D_{k1} since in [15] these values were adapted for more people than the other ones.

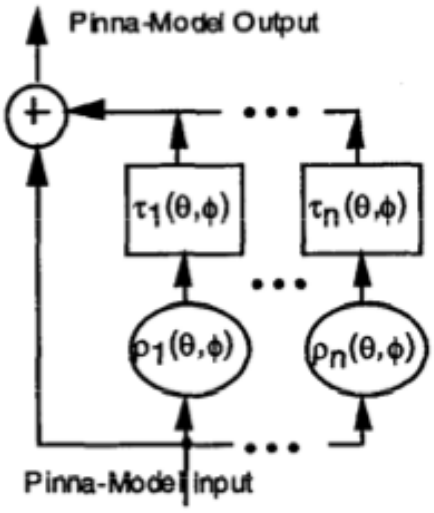

Figure 3.10. The pinna model used in our system implementation. [15]

k	ρ_k	A_k	B_k	D_{k1}	D_{k2}
1	.5	1	2	1	.85
2	-1	5	4	.5	.35
3	.5	5	7	.5	.35
4	-.25	5	11	.5	.35
5	.25	5	13	.5	.35

Figure 3.11. Pinna model coefficients. [15]

3.2 Listening environment

Interactive virtual environment systems simulate the experience of immersive exploration of a three-dimensional virtual world by rendering the environment as perceived from the viewpoint of an observer moving under real-time control by the user. We must pay more attention to produce realistic sound in order to create a complete immersive experience in which aural cues reinforce the localization cues, to support more natural interaction, navigation and sense of presence within a virtual environment. For example, changes in sound reverberation, can enhance and reinforce the immersivity of the system but at the same time degrade the sound localization capability of the users.

In order to take into account the reflected sound reaching the virtual listener, we refer to the example in Figure 3.12, in which a direct sound along with a reflected wave reaching the listener's ears. To model the sound reaching the listener, we need to know not only the direct path but also the sound reflection path: the simple method that can be used to find all possible paths is the image sources methods [107]. In this method, we choose a point source in an environment. When a wall

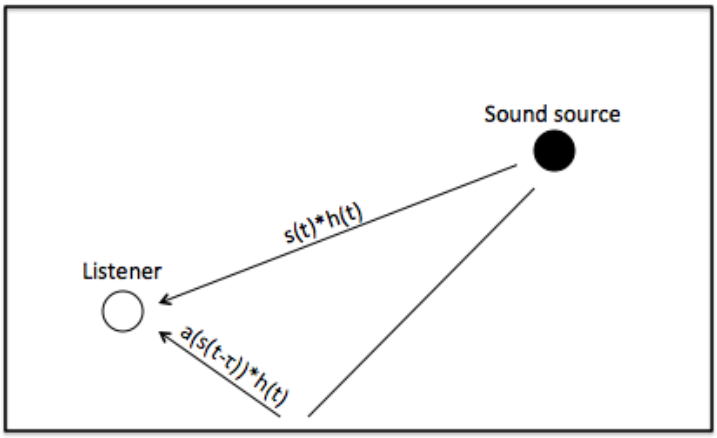

Figure 3.12. The direct sound and a reflection reach the listener: the first is filtered by the hrtf, the second is delayed and attenuated before being filtered.

is encountered, the wall boundary condition may be satisfied by placing an image symmetrically on the far side of the wall [107]. Using this method, a specular reflection can be constructed geometrically by mirroring the source in the plane of the reflecting surface: reflected paths from the real source are replaced by direct paths from reflected mirror images of the source. In the example we take into consideration, we place in the reflected point an image source and we model the reflected sound as if it were a direct sound. The important information used here is that in the time-domain, each image contributes only a pure impulse of known strength and delay.

Once we know the delay and amplitude attenuation for each reflection path we can model the reflected signal like a delayed and attenuated sound source signal. In Figure 3.12, the direct signal is the convolution between the sound source signal and the impulse response, while the reflected signal is the convolution between the sound source signal delayed by an amount of time τ and it is attenuated by an amplitude factor a. The time delay τ is given by the inverse of the distance, that represents the longer path that the reflection travels before arriving at the listener; the attenuation a is given by the multiplication between the inverse of the distance times the wall reflection coefficient on which it bounces. If we continue to consider the example in figure 3.12, and we suppose a constant wall reflection coefficient γ, and we define d_r the additional distance for the reflected sound, then $\tau = \frac{1}{d_r}$ and $a = \frac{1}{d_r}\gamma$.

This simple method works fine for simple geometric models like a rectangular room model, but when the geometric complexity of the room increases, it becomes difficult to position all the image sources. Indeed, when the number of walls increase the situation becomes more complicated because each image is itself images. For this reason, it's necessary to use a geometric technique that tracing the sound ray path returns the sound image positions, like for example the beamtracer, that enables to model a reverberant environment, describing the early reflections like the ensemble of all the reflected paths between the sound source and the listener. In this way what we need is to precompute and store spatial data structures that encode all possible transmission and specular reflection paths from each audio source, and then use

these data structures to compute reflection paths to an arbitrarily moving observer view-point for real-time auralization during an interactive user session.

3.2.1 Beamtracer

Beam tracing is an efficient geometric solution to the modeling of sound propagation based on acoustic beams. This method was originally developed in [108] for image rendering applications and later it has been extended in [11] to audio rendering. Beam tracing methods classify reflection paths from a source by recursively tracing pyramidal beams (i.e., sets of rays) through the environment. Briefly, a set of pyramidal beams are constructed that completely cover the 2D space of directions from the source. For each beam, polygons are considered for intersection in order from front to back. As intersecting polygons are detected, the original beam is clipped to remove the shadow region, a transmission beam is constructed matching the shadow region, and a reflection beam is constructed by mirroring the transmission beam over the polygon's plane (figure 3.13).

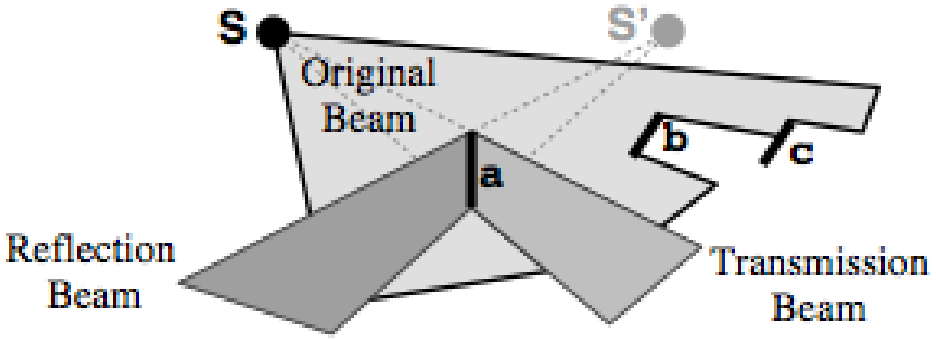

Figure 3.13. Beam tracing method. [11]

The beam tracer adopted in this work is that introduced in [104]. It is a fast implementation of the orginal algorithm, and it operates in two stages. First, a visibility map of the reflectors is computed offline, independent from the positions of the source and the receiver, and these are stored in a suitable data structure (beam-tree). Then, the rays that propagate from the source to the receiver are obtained through a simple lookup (path-tracing) of the beam-tree. This operation turns to be very efficient, and thus suitable for real-time purpose. Indeed, beamtracing turns to be effective in interactive sessions where the user navigates in the virtual environment. In particular, beamtracing makes possible to predict in real-time the reflective paths from the virtual loudspeakers and the listener positions.

The attenuation, length, and directional vectors for the corresponding reflections path can be derived quickly from the data stored. In figure 3.14, it is shown an example representing the specular reflection path to a particular receiver point (labeled 'R'), where the label 'A','B','C','D' and 'E' represents the spatial subdivision of the space operated by the beamtracing algorithm. Once a set of reflection paths from a source to the receiver has been computed, the source-receiver impulse response is generated by adding one pulse corresponding to each distinct path from the source

to the receiver. The delay associated with each pulse is given by l/c, where l is the length of the corresponding path, and c is the speed of sound. Since the pulse is attenuated by every reflection and dispersion, the amplitude of each pulse is given by Γ/l, where Γ is the product of all the frequency independent reflectivity coefficients for each of the reflecting and transmitting surfaces along the corresponding reflection path.

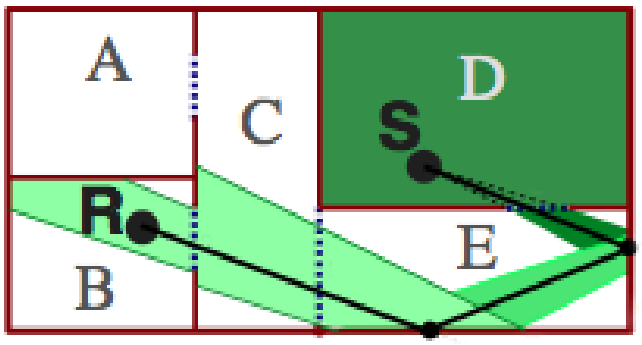

Figure 3.14. Reflection path to receiver point ('R') for source point ('S') computed by the beamtracer. [11]

The system with this integration supports real-time auralization of sound based on realistic acoustic modeling in virtual environments. The advantage of this method is that it doesn't model only the specular reflections as well as the fact to be independent to a particular receiver position, and so to be applicable in virtual environment applications like our system in which the receiver may move.

3.2.2 Modeling the environment

We have developed a system that uses precomputed spatial subdivision and beam data structures to enable real-time acoustic modeling and auralization in interactive virtual environments.
Our virtual environment system takes as input a description of the geometry and acoustic surface properties of the environment and a set of anechoic audio source signals at fixed locations. As the user moves in the virtual environment, the system uses the beamtracer to compute the positions of the images sources seen from the current user position, along with a stereo audio signal spatialized according to the computed reflection paths from each audio source to the observer location. In particular, we constructed off-line a lookup table containing all the information regarding the reflection paths from any source to any given position inside the listener room. The beamtracer gives us a table in which for any possible listener position, it contains the position of any image source for any virtual loudspeaker and the walls on which it bounces off before reaching the listener. Using these data, we can retrieve the distance from any image source and the ears of the listener, that in addition to the attenuation given from the walls absorption permits us to delay and attenuate the dry sound. The delay for any reflections is given by the distance between an image source and the listener's ears and the attenuation is

given by the distance between an image source and the listener's ears and by the reflection coefficients of the walls. So we can save a lookup table indexed by any possible listener position containing for any rotation of the head the delays and the attenuations to be applied to the sound in order to model the reflections for each ear, like that in figure 3.15.

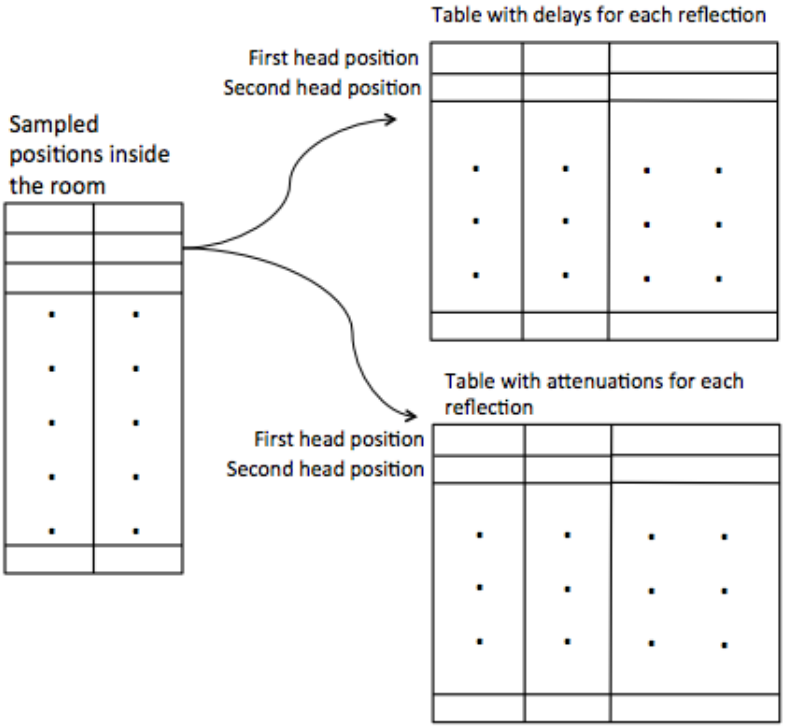

Figure 3.15. Representation of the lookup table returned by the beamtracer, in which are saved all the informations needed to implement the room reflections.

The version of the beamtracer used here works in 2D, so it considers only lateral reflections.

Since our model is able to consider not only azimuthal cues but also elevation cues, and since lateral reflections provide a realistic listening but are detrimental to azimuth localization, we can follow the work in [109], where it is found that a listening environment comprised of a physical floor can enhance the localization accuracy. Thus, we take in consideration in addition the floor reflections, using the model presented in [16], in which the authors determine the elevation angles of arrival through geometric ray-tracing. Only the first order floor reflection is considered, and it is supposed arriving from the same azimuth θ of the direct sond, in order to improve the localization. The virtual floor was defined to be 1.6 m below the subject's head and it was modeled with a frequency-independent reflection coefficient of 0.5. The azimuth angle of the floor reflection was the same as that of the direct sound. The elevation angle ϕ was computed using the equation

$$\phi = \arctan\left(\frac{2h_l + \rho\sin(\pi_e)}{\rho\cos(\phi_e)}\right). \tag{3.15}$$

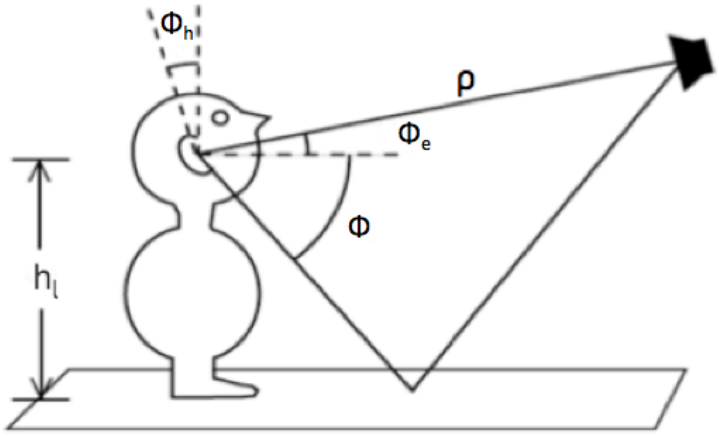

Figure 3.16. Virtual floor reflection. [16]

In the Figure 3.16, h_l is the distance between the floor and the subject's head, ρ is the distance of the source, ϕ_e is the elevation of the source relative to the horizontal plane, ϕ_h is the elevation of the subject's head relative to the horizontal plane and ϕ is the elevation of the reflected sound. As for lateral reflections, even here we save a lookup table in which for every listening position in the room, we can retrieve the delay to be applied at the sound that depends on the distance from the sound source, and the scaling factor that depends on the reflection coefficient and inversely with the range of the total distance traveled by the sound.

After modeling the early reflections, it's needed to model the diffuseness of the reflections or late reverberation. It has been found that a model of this reverberant component can be achieved by using a pseudo-random binary sequence (PRBS) with exponential attenuation or decay [16]. In the following, we describe the part of the impulse response related to the late reverberation: $h_L[n]$ represents the impulse response where the contribution of early reflections is neglegted, and g[n] represents a sample from the normal distribution with mean μ and standard deviation σ.

$$h_L[n] = \begin{cases} a \exp(\frac{(n-t_d)}{b}) g[n], & \text{for } n \geq t_d \\ 0, & \text{for } n < t_d \end{cases} \tag{3.16}$$

The three parameters of the late reverb (a, b, t_d) control the total energy in the late reverb, the decay-rate of the late reverb, and the time between the direct sound and start of the late reverb, respectively. The ratio of direct-to-reverb energy (DRR), i.e the ratio between the sound energy of the direct sound and the reflected sound energy [110], is considered a measure of the loudness of the reverb, and it is defined as

$$DRR = \sqrt{\frac{\sum_{n=0}^{N_D} h(n)^2}{\sum_{n=N_D+1}^{N} h(n)^2}} \tag{3.17}$$

where h(n) is the impulse response, N_D is the last sample corresponding to the direct sound, and N is the length of the impulse response.

The starting time of the late reverb portion of a room response is usually considered to be about 80 msec after the arrival of the direct sound.

3.3 Head rotation

We choose to introduce also the dynamic cues represented by the virtual head rotation, trying to improve the median-plane externalization and to reduce the front-back confusion, in order to give a more realistic experience. For taking into account it, we decide to implement the interpolation of the signals using the motion-tracked binaural explained in Chapter 2.

We have chosen to sample the space in 16 angular positions in azimuth that is sufficient for music [49], and using an interpolation procedure that overcomes the problem caused by spatial undersampling. With a practical number of listening positions in space, the simplest approach of merely switching between these ones produces positional discontinuities and distracting clicks [111]. Direct linear interpolation of listening positions signals produces comb-filter spectral notches, where the frequency of the first notch is inversely proportional to the spacing between the sampled rotations of the head. This produces annoying "flanging" sounds when the listener turns the head [49]. To eliminate the flanging sounds associated with the spectral notches, the listening positions signals are split into low-frequency components (below $0.5 f_{max}$) and high-frequency components (above $0.5 f_{max}$), where f_{max} is the maximum considered frequency given by $f_{max} = \frac{F_s}{2}$ where F_s is the sampling frequency, beyond which we are no longer able to correctly reconstruct the signal. The low-frequency components are interpolated, and then the high-frequency components are restored. Several methods have been investigated for restoring the high frequencies [111]. The interpolation procedure that we use is illustrated in figure 3.17, where the restoring of high frequencies is done using the high-pass filtered nearest listening positons signal, w_0 represents the distance in degrees betwen two listening positions and w represents the distance from the nearest listening position.

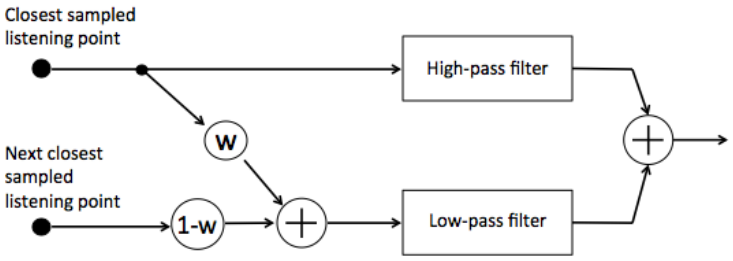

Figure 3.17. Scheme of the signal interpolation method used in our implementation to take into account the head motion.

3.4 Implementation choices

The audio processing procedure is depicted in figure 3.18. Since the output audio is generated at discrete intervals, the amount of data required by the output audio determines the frequency of execution for the other processing tasks. For example, outputting 2048 audio samples at a sample rate of 44100 Hz corresponds to about 46 ms of audio data. So approximately every 46 ms the audio pipeline will render a new set of 2048 audio samples. The size of the output buffer (in our implementation is 8192 samples) is a crucial parameter for the real-time audio processing. Since the audio pipeline must respond to changes in the virtual listener's position and/or virtual listener's head rotation, the delay between when the orientation change is made and when this change is heard is critical. This is referred to as the update latency and if it is too large the listener will be aware of the delay, that is the virtual listener will not appear to be moving as the listener moves the virtual listener from the user-interface. The amount of allowable latency is relative and may vary, but values between 30-100 ms are typically tolerated.

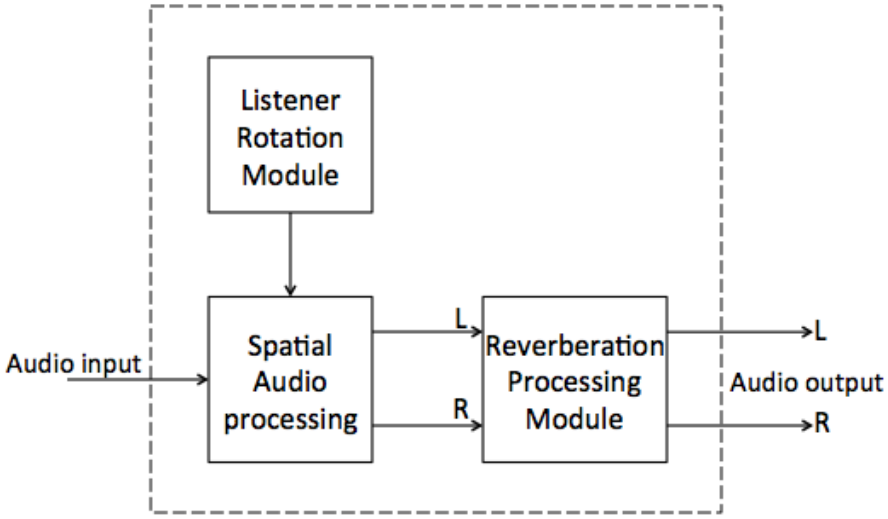

Figure 3.18. Block diagram of audio processing module

An important off-line computation is accomplished: for each sampled position in the room, in a lookup table the direction of arrival for each direct and reflected sound, as well the distance to the listener and the attenuation given by the reflection coefficients of the walls and the distance to be traveled, are stored. What we do, after deciding to use a 16-virtual listening positions related to the rotation of the head, is to consider 16 possible angular directions of the sound arrival (as depicted in Figure 3.19 for only some angles), each of which is associated with a set of cues filters that together compose the structural HRTF. When a virtual listener is located in a given point of the room, we can retrieve from a lookup table all the features for the reflections about that listening point; then, dependent on the rotation of the head, each reflections will reach the listener's ears at different angles of arrival. The

range of these angles is divided in 16 regions, each of which is associated with the corresponding filters. For example, considering the situation depicted in Figure 3.19 in which the virtual listener's head is rotated of zero degrees, and in which only few of regions are represented: our system retrieves from the table stored all the sounds reaching the listener in that listening position, and in particular it retrieves for each region the reflections that fall into. In this way, we know that for this situation all the sound waves reaching the listener at an angle of arrival that is in the range of the region 1, will be filtered using the filters associated with an angle of 0 degrees.

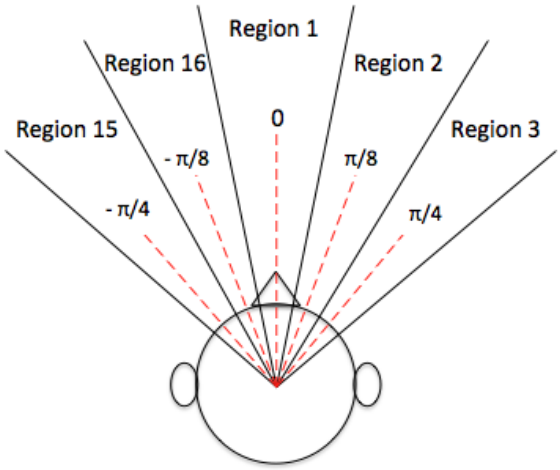

Figure 3.19. Representation of the division of the space in some regions and their central values used in the implementation of the system.

Using this approach, every sound that arrives in a certain range of angles of arrival, will be filtered with the same HRTF. Note that it's an important choice, because in this way we can consider any order of reflections that we will always have a maximum of 16 HRTF filterings.

The monaural sound to be spatialized is sent to the spatial audio processing: if it is a mono signal then the virtual loudspeakers will be fed by the same input, otherwise one channel will be send to the right virtual loudspeaker and the other to the left one. Before audio processing, these signals are windowed by 50% overlapped Hanning windows of length 4096 samples. Then, once the position in the room and the head rotation are selected, the filters related to the nearest and next nearest virtual listening positions are retrieved in order to filter the signal. Before apply the model, each sounds in each position is delayed and attenuated according to the distance between the sound source and the listener's ears position, and to the any bounces on the walls. Then, for every sound in each portion, they are all summed up to obtain the inputs that will be filtered.

In this way, not only the direct sound is filtered by the HRTF but also the early reflections. We always obtain two signals, no matter if the virtual listener is rotating or not, and then we interpolate them using a simple linear interpolation presented in Section 3.3 in Figure 3.17: the weight w used is related to the distance from the virtual listener to the sampled listening point (Figure 3.20).
As stated in [49], 'discrete room reflections of 50 ms in total duration may be

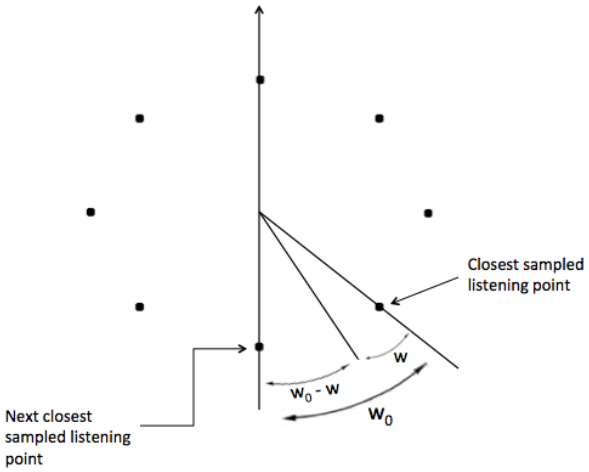

Figure 3.20. The interpolation between two outputs: the weigths are dependent on the rotation angle.

sufficient'. For this reason, we model the early reflections up to the third order of reflections, and then we add the late reverberation. Since we know that the late reverb does not contribute to the localization of the sound source, we use the monaural sound like input to the late reverb filter that is a choice made also by other structural models in literature [15]. The output is delayed by an amount of time τ representing the time between the direct sound and start of the late reverb, and then mixed with the output of the HRTF model by a gain factor in order to take into account the direct to reverberant ratio.

Finally, in order to take into account a moving virtual listener in the room, an additional interpolation is required. Some listening positions inside the room are sampled using a linear grid, like in the example in Figure 3.21, where the step Δ_X from a point and the next one in the horizonal direction and the step Δ_Y from a point and the next one in the vertical direction are considered to be equal to 0.15 m. Considering the Figure 3.21, when a listener moves in the room and is located in a given position (LP_X, LP_Y) that is depicted like the black circle: we compute the signal ouputs from the four surrounding listening points, that are represented like crosses. With respect of them, we define the ouput signals for each of these like y_{P1},y_{P2},y_{P3},y_{P4}. Then, the output signal for that specific position (LP_X, LP_Y) is obtained using a bilinear interpolation, that is a linear interpolation over two variables.

The key idea is to perform linear interpolation first in one direction, and then again in the other direction. First we interpolate on the x-axis and in this direction more influence is given by the sampled position closest to the actual listening position; thus, we obtain the two points

$$V1 = \frac{\Delta_X - LP_X}{\Delta_X} y_{P1} + \frac{LP_X}{\Delta_X} y_{P2} \tag{3.18}$$

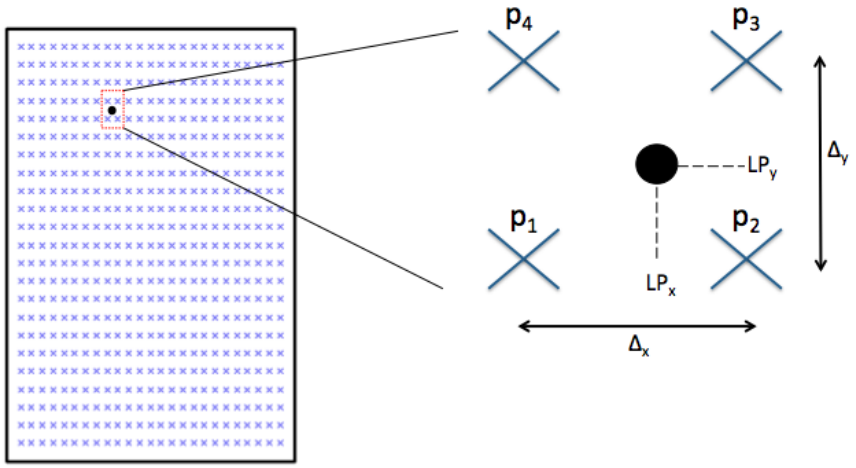

Figure 3.21. Bilinear interpolation.

$$V2 = \frac{\Delta_X - LP_X}{\Delta_X} y_{P4} + \frac{LP_X}{\Delta_X} y_{P3} \tag{3.19}$$

Then we interpolate in the vertical direction, and this leads to

$$LP = \frac{\Delta_Y - LP_Y}{\Delta_Y} V1 + \frac{LP_Y}{\Delta_Y} V2. \tag{3.20}$$

We define $C_X = \frac{LP_X}{\Delta_X}$ and $C_Y = \frac{LP_Y}{\Delta_Y}$ so that the signal output at the position (LP_X, LP_Y) is

$$y_{LP} = (1-C_X)(1-C_Y)y_{P1} + C_X(1-C_Y)y_{P2} + C_X C_Y y_{P3} + (1-C_X)C_Y y_{P4} \tag{3.21}$$

In this way a bilinear interpolation between the output signals of the four sampled listening points surrounding the virtual listening position selected by the user is introduced.

Chapter 4

Interface Description

In the previous Chapter we introduced the description of our system. Now, in order to obtain an interactive implementation, it is needed to introduce some aspects that permit at the user to interact with the system and enjoy a full immersive experience. We designed three virtual environments, in which the user can experience different listening experiences. In this Chapter we describe how these environments differ in their geometrical and reflective properties. Moreover, we explain how we take into account the possibility for the user to rotate the virtual listener's head. Finally, we present the implemented Graphical User Interface that provides the user with some important information about the listening setting.

4.1 Listening Environments

We have chosen to model three different listening environments with different size and reverberation characteristics.

The first is the simple rectangular room depicted in figure 4.1, in which the reflection coefficients are constant for all the walls, set to a value of 0.48. This scenario corresponds to a conference hall of size 4 x 6 meters, and the user can select a listening position inside the room, where the red circles represent two virtual sources. In this room the reverberation tail lasts 0.7 seconds.

The second is the representation of model for a room in the *Sound and Music Computing Lab* of Politecnico di Milano, Polo Regionale di Como, in which the reflection coefficients are the same for all the walls, set to 0.6. The model is depicted in figure 4.2 and the red circles represent two virtual loudspeakers. The reverberation tail of this room lasts 1.8 seconds, corresponding to the value that best fits the impulse response measured in the real room.

The third is the model of a concert hall in which the reflection coefficients are different from wall to wall. The floor map of this room is shown in Figure 4.3, where the different reflection coefficients are indicated beside each wall and the virtual sources are depicted like red circles. In this room the reverberation tail lasts 2.7 seconds since it is bigger than the other two rooms.

In order to keep the computational cost affordable, we limited ourself to compute early reflections up to the third order, by means of the beamtracing algorithm. As explained in Chapter 3, the rest of the response is obtained adding a non-deterministic

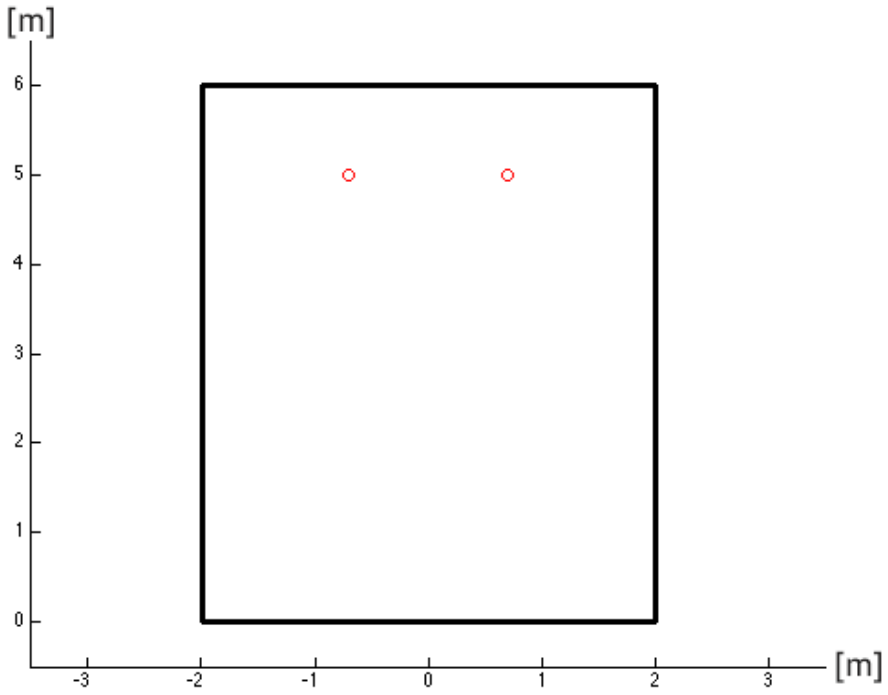

Figure 4.1. A simple rectangular room model in which two virtual loudspeakers (red circles) are located.

reverberation tail, modeled like an exponential decaying normal distribution whose parameters are mean $\mu = 0$ and standard deviation $\sigma = \sqrt{12}$. The amount of time τ representing the time between the direct sound and the starting of the late reverb changes among the three room models: it is set to $\tau = 50$ ms for the rectangular room, $\tau = 60$ ms for the model of the Sound and Music Computing Lab of Politecnico di Milano, Polo Regionale di Como, room and $\tau = 70$ ms for the concert hall, before to be mixed with the output of the HRTF model. These values are such that the resulting model of the impulse responses resemble the shape of feasible responses. The intensity level of the reverberation tail is scaled depending on the distance ρ of the virtual listener from the virtual sources, according to $1 - \frac{1}{\rho}$ [112].

4.2 Virtual head rotation

Our system provides the possibility for the user to rotate the virtual listener's head, and also to navigate inside the listening room. This facility is enabled by means of a device that permits to interact with the system. A simple choice is made using a MIDI controller that communicates with the personal computer sending messages in a specific format, which is precisely the MIDI format. In particular, we use an *AKAI LPD8* where the user can use a knob (circled in red in Figure 4.4) in order to rotate the virtual head.

In addition to this control, the user can navigate inside a listening room selecting a listening point using the mouse.

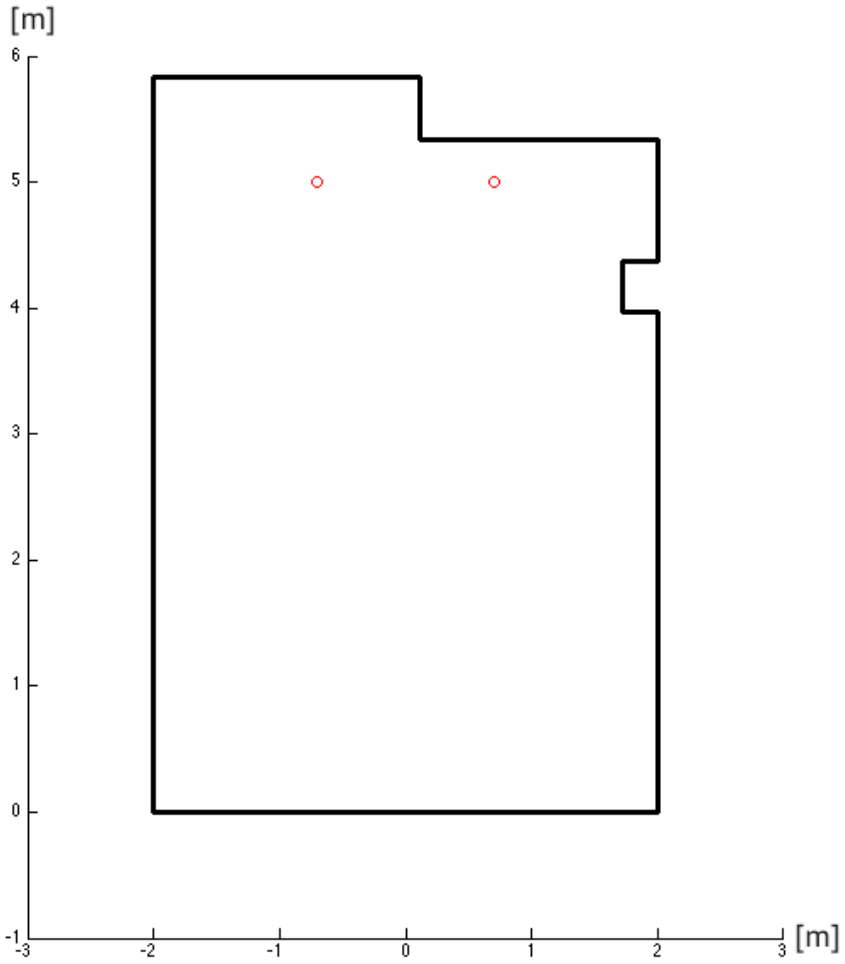

Figure 4.2. The *Sound and Music Computing Lab* of Politecnico di Milano, Polo Regionale di Como, room model, in which two virtual loudspeakers (red circles) are located.

4.3 Graphical User Interface

In order to provide an interactive system and reinforce the auditory cues with a indication of the orientation of the head, it's important to develop a system that provides a user interface in which the user can observe the position of the virtual listener and the orientation of the virtual head. In this context our simple graphical user interface provides three objects, that guarantee the user the control of the system. The implemented interface is depicted in Figure 4.5, where a number marks each single object, so that we can refer to this number in the following explanation.

The object marked as 1 is the representation of the virtual listening environment in which the listener is hearing: two virtual sound sources are depicted, like red circles, and the position of the virtual listener, like a blue circle. In this way the listener can understand in which position is listening to the music. Morover, by clicking with the mouse in a point of the room model, it is possible to change the listening position. The pop-up menu allows the user to change the environment,

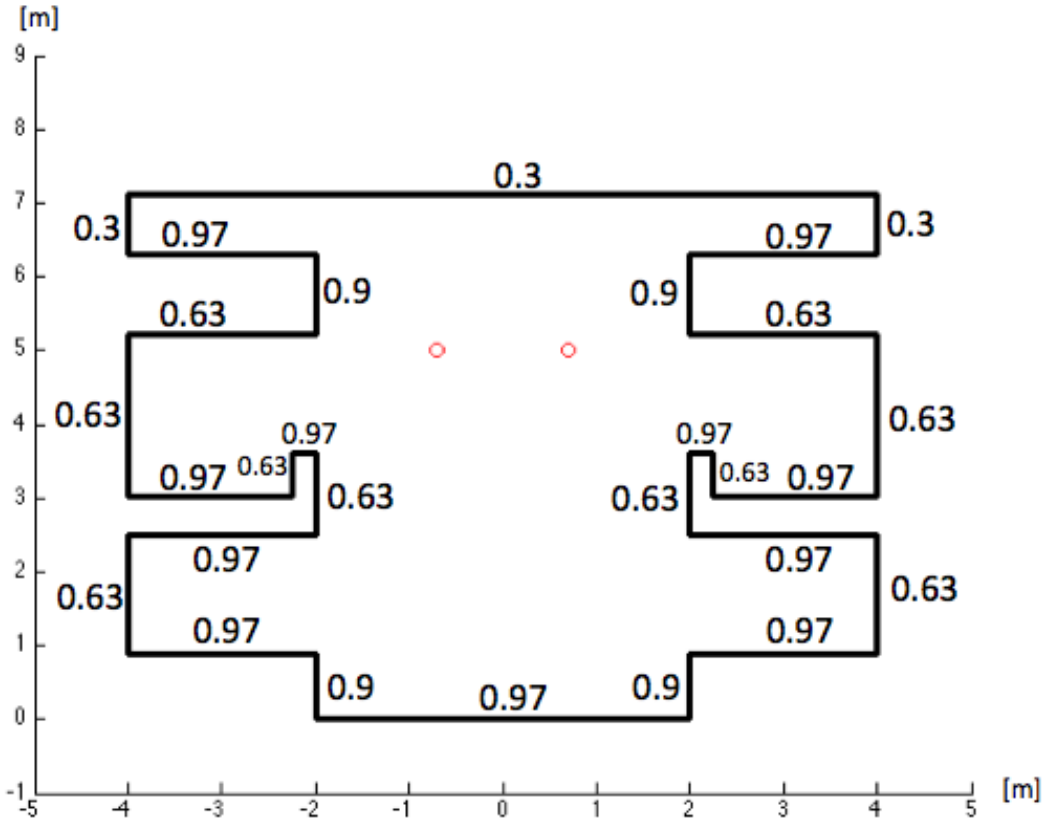

Figure 4.3. A concert hall model in which two virtual loudspeakers (red circles) are located.

Figure 4.4. The AKAI LPD8 MIDI controller [17]. The knob circled in red permits to the user to rotate the virtual head.

choosing from a list of available options. In the object marked as 2, we show the orientation of the virtual head: it is updated each time the user rotates the corresponding knob and it is needed to provide a visual indication about the angle of orientation of the virtual listener with respect the two virtual sound sources. The third object, marked as 3, is a button that permits the user to browse the list of available songs, and to select the desired one.

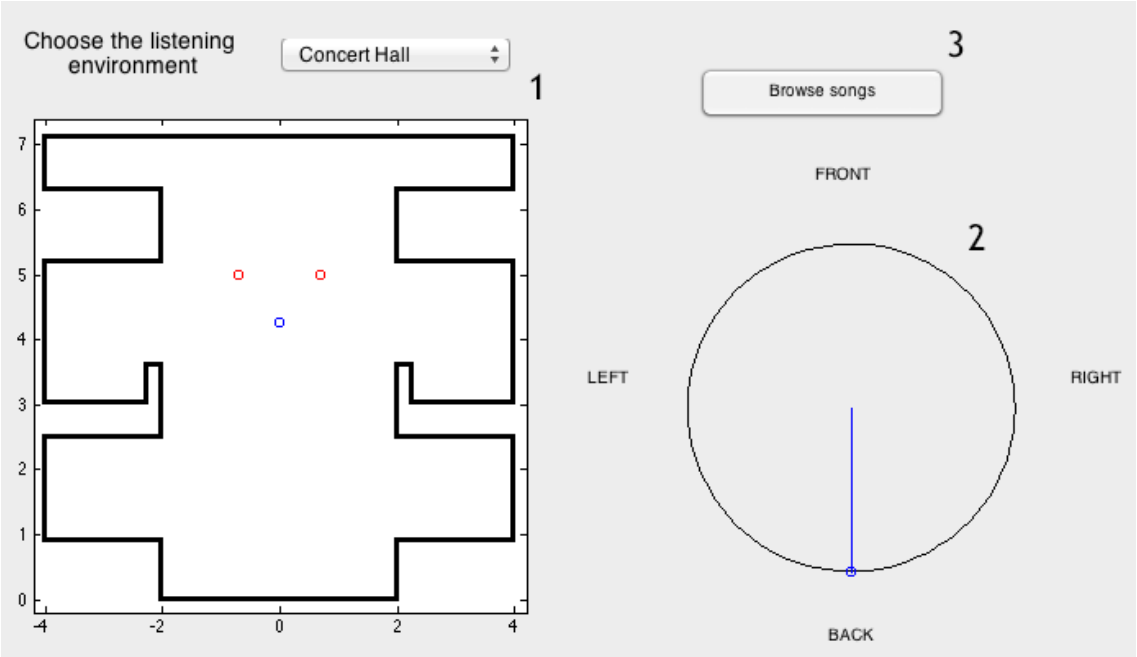

Figure 4.5. A simple implemented Graphical User Interface for our system.

Chapter 5

Evaluation methodology

In Chapter 3 we have presented a system implementation in order to produce a spatialized sound, using a structural model. We know (Section 2.3.3) that it is an approximation of the HRTF and that some localization and externalization problem can arise. In this chapter we propose formal listening tests aimed at providing a subjective evaluation of the previously introduced implementation. In particular, we will discuss in detail the design phase of listening tests, presenting the criteria that most suite the need of our work.

The most common criteria used for assessing the quality of a binaural system are localization or perceived separation of multiple instruments or the width of the sound stage, the sense to be immersed in sounds, typically encountered in a concert hall, the sense of distance or depth of the sound field, the perception or identification of a real performance space, and the sense of naturalness or presence [113]; these are all characteristics that a good system should have.

In order to provide a subjective evaluation of the effectiveness of the system, we need to rely on results gathered from listening tests. In this chapter we present a formal methodology for devising a set of suitable listening tests, aimed at producing reliable results.

We analyze separately the cases of localization of an acoustic virtual source, localization of a moving virtual listener and the sound rendering system effectiveness. With the latter, we mean the correct ability of the system to simulate the correct sound source position and listening environment, at each time, depending on the virtual listener's head rotation. This choice is due to the need of evaluating different aspects of the system. On one hand the correct localization of a sound in the space, in presence and in absence of a listening environment, and the identification of a sound trajectory can be tested using a specific test procedure. On the other hand, the rendering capability is a macroscopic feature and being also subjective is difficult to test. In this context, we do not introduce strong control conditions and in particular we are interested in evaluating the immersivity of the listener and his or her personal evaluation of the system.

In Section 5.1 we present the conditions under which the tests are conducted, and in particular the people involved and the material used. In later Sections, the various tests are presented: we show the design of each one, the method used and the chosen analysis of the results.

5.1 Test conditions

In this section we present the choice of general test parameters, which are valid for both the three tests. In particular, we discuss the selection of the test panel (i.e. the characterization of subjects involved in the listening tests) and the choice of the sound material adopted in the tests.

5.1.1 Selection of the Test Panel

The subjects involved in our listening tests are not experienced in listening to sound in a critical way in contrast with ITU-R BS.1284-1 that states that expert listeners are preferred. However the system is not intended for high-quality sound broadcasting or reproduction and so this choice is not a problem.

5.1.2 Test material

In order to verify the ability of the system to correctly reproduce a virtual sound source, we devise a set of listening tests related to the perceived sound location. To do so, we adopt a male speech extracted from the European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) CD [114]. Such sound sample comes from easily accessible sources and has already been adopted for listening tests in the context of sound field rendering evaluation [115]. We choose a vocal signal because human voice is considered critical to evaluate the sense of audio quality and it is known that localization is most sensitive with speech or singing [115].

As far as the evaluation of the system immersivity is concerned, we adopt musical samples, since music is a broadband signal which in addition has a clear temporal structure; moreover, a considered application of the system is to be a music player. In particular, the choice is an excerpt from Suzanne's Vega *Tom's Diner* for the third test. The mono test sequences are rendered as virtual sources (according to the model explained in Chapter 3) at different fixed or moving positions according to the test procedure explained later.

5.2 Sound localization

In this Section we describe the test method adopted for a subjective evaluation of the localization of a virtual sound source. In particular, we restrict ourselves to the case of a single virtual source rendered at the same distance from the listener but with different angles of incidence and in different listening environments. At the end of this Section, we discuss the statistical analysis procedure adopted in order to identify the average behavior of the system and the reliability of the results.

5.2.1 Experimental design

The perceptual evaluation of the sound source localization capability requires a formal test method, with a well defined protocol.

In particular, in the design phase of such experiments we need to ensure that uncontrolled factors will not deviate the results of the listening tests. As an example, we propose to each subject the same stimuli but in a different and random order. In

this way we can ensure that the judgements made by the subjects are independent from the actual sequence of stimuli. Similarly, listening tests need to be designed so that subjects are not overloaded to the point of lessened accuracy of judgement. A grading session should not last for more than 20-30 minutes, and that no more than 10 to 15 trials per session should be scheduled to achieve the desired session length [116]. The sound signal should be controlled so that the amplitudes of the peaks only rarely exceed the peak amplitude of the permitted maximum signal defined in Recommendation ITU-R BS.645.

Another important factor we have considered during the design phase of our listening test is that the listeners must not be overloaded, in order to prevent a loss of accuracy in the judgements.

Finally, a particular attention should be paid about the choice of a response technique. There are different available way of reporting a response [117, 118]: for example it could be used the head-pointing technique, in which the subjects hold their heads still during the sound presentation, and then turn their heads to point their noses in the direction from which they believe the sounds are presented; the position of the subject's head can be monitored using an head-tracker. In verbal pointing techniques the perceived localization is indicated through verbal reports, like for example indicating the spherical coordinates. Other different pointing techniques are available like for example that described in [119].

The accuracy of judgements collected with the head-pointing technique is high but, it may requires subject training; instead, the verbal reporting technique is simple but less accurate [120]. For our localization experiments, the listeners are asked to place a point on a computer screen indicating the position of the perceived sound: it is more intuitive than the verbal reporting, and the judgements are collected at fast rates.

5.2.2 Test method

In order to get a subjective evaluation of the localization of an acoustic virtual source we adopt a simple test method in which we present seven stimuli without any reference. One subject at time is involved in the listening test and seven stimuli are presented: the subject is asked to indicate the perceived angle of incidence of the virtual source assigned to the stimuli, and once answered the next stimulus is presented, as suggested in ITU-R BS.1284. We decide to discretize the angle of incidence of 5° degrees: this is a reasonable choice due to the errors on the minimum audible angle, i.e. a paradigm designed to measure the accuracy which human listeners can localize a source of sound [121], of the listeners using non-individualized HRTF [10, 122, 123]. As well, the listener is asked to indicate a confidence value of the response in a range from zero to five, where zero means that he or she is not sure at all about the given answer, while five means he or she is perfectly sure about the response given.

Only one virtual loudspeaker is used and the test is repeated under two different scenarios: in an anechoic chamber and in a reverberant room. In particular the reverberant room considered is a model of the *Sound and Music Computing Lab* of Politecnico di Milano, Polo Regionale di Como.

The azimuth values for perceived location were recorded as well as the azimuth

values used for rendering the sound source. Subjects are not get any feedback about their pointing accuracy. The stimuli presentation is not repeated. The interval between each stimuli depends on the time that subjects take to indicate the angle of arrival; however, it's greater than some seconds. The subjects are provided with oral instructions that include a brief introduction to the scope of the test and a description of the technique of presentation of the stimuli. After that, it is checked that such task is clear for the subject.

5.2.3 Statistical analysis

A statistical analysis of the results is needed in order to identify the average performance of the system and the reliability of the results. For our test we refer to the analysis introduce in the *Recommendation* ITU-R BS.1116.

We define x_i the score of subject i. Then, the mean score for each of the presentation is computed as

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5.1}$$

where N is the number of subjects involved in the test.

The results of the test are provided along with the related confidence interval, in order to provide an explicit indication of the reliability of the results. The confidence interval is derived from the standard deviation and the size of the listening panel. In particular, we use a 95% confidence interval which is given by

$$[\bar{z} - \delta, \bar{z} + \delta] \tag{5.2}$$

A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. In this case δ is determined by

$$\delta = t_{0.05} \frac{s_s}{\sqrt{N}} \tag{5.3}$$

where $t_{0.05}$ is the t-value [124] for a two-sided confidence interval (i.e., the range is defined in terms of lower and upper limits computed from the sample of data) level of 95%, as reported in Table 5.1, and s_s is the standard deviation for all subjects in session s.

From Table 5.1, we can retrieve the t-value for a confidence interval of 95% starting from the number of scores (i.e. the number of degrees of freedom, equals to $N - 1$), that depends on the size of the sample.

5.3 Trajectories localization

In this Section we discuss how to evaluate the rendering system when the listening point changes over time. To do so, we devise a set of tests considering the rendering

Number of scores	*t*value	Number of scores	*t*value
1	12.70	15	2.131
2	4.303	16	2.120
3	3.182	17	2.110
4	2.776	18	2.101
5	2.571	19	2.093
6	2.447	20	2.086
7	2.365	25	2.060
8	2.306	30	2.042
9	2.262	40	2.021
10	2.228	50	2.009
11	2.201	100	1.984
12	2.179	120	1.980
13	2.160	$1000 \sim \infty$	1.962
14	2.145		

Table 5.1. Value of Student's t distribution for two-sided confidence interval of 95%.

of a virtual source within a listening environment. The listener point is smoothly moved in the environment to define a set of trajectories to be identified by the subjects. At the end of this section we discuss the statistical analysis procedure adopted for determining the capability of the users to identify the correct trajectory.

5.3.1 Experimental design

The spatial impression, here restricted to the localization of a moving virtual listener, is a macroscopic feature delivered by a sound field rendering system. In order to gather reliable information from a subjective evaluation, in this scenario it is convenient to introduce a visual reference. Moreover, as we made for the previous test, we propose to each subject the same stimuli but in a different and random order. In this way we can ensure that the judgments made by the subjects are independent from the actual sequence of stimuli.

5.3.2 Test method

In order to get a subjective evaluation of the localization of a moving virtual listener, we adopt a simple test method. The test data consist of a set of four different trajectories that are selected randomly when the test starts.

The participants are asked to carefully hear the sound on the headphone. During the listening, they are provided with three graphical representations of possible trajectories, only one of them corresponding to the heard one. The stimulis are repeated until the subjects indicate their choice.

As we said in Chapter 2, distance cues are closely associated with listening in a reverberant room. Indeed, as well the changes in intensity, the main cue that people use to discriminate the distance of a sound source, is the direct-to-reverberant ratio [68, 112], that is the ratio of the energy in the first wave front to the reflected sound

energy. In this context, it is needed that our listening environment is not an anechoic chamber so that listener could use the reflections like cue to better understand the distance from the source. On the other hand, judgements about angle of arrivals of sounds in a reverberant room are difficult with respect an anechoic chamber; indeed, reflections can be confusing for the listener and the localization task can results harder. The identification of trajectories requires an attention toward the distance cues and also the angle of arrival of the sound, and in this context a medium size room like the concert hall model (see Chapter 4) is the best choice, due to its reverberant characteristics that provide good distance cues but at the same time do not reduce so much the localization capability of the user.

In order to proceed with the test, the participants are asked to seat comfortably in the chair in front of the computer used for the test. Before starting, the objective and the procedure of the experiment were explained to each individual participant, and a training example is carried out: this consists in presenting a sound trajectory on headphones along with its correct graphical representation.

5.3.3 Statistical analysis

For this test, we choose to report the results in a table form in which each row shows how many trajectories from each test example were recognized or identified like a different trajectory. In this way, we can extract the overall recognition rate for each trajectory and rank the overall classification.

5.4 Rendering system

The goal of this test is to specifically evaluate the combined effect of realism and immersivity for a sound rendering system implemented using the model presented in Chapter 3, as well as the system's usability. The evaluation of such subjective goals is difficult, because no reference response exists and the range of different responses can be wide. In this context, we do not introduce strong control conditions. However, we need to take into account some aspects design phase of such experiments. In particular, as we made for the previous tests, we propose to each subject the same stimuli but in a different and random order. In this way we can ensure that the judgments made by the subjects are independent from the actual sequence of stimuli. The aim of the concerning test is twofold: on one hand we evaluate some quality features, on the other hand we receive a judgement about the user experience with respect to a realistic music player implementation of the system in which is possible rotating the listener's virtual head.

5.4.1 Experimental design

It is difficult to define measurement metrics of the perceived quality of a spatial sound field, and to evaluate its degree of immersivity and realism. The general consensus of researchers in the field of spatial audio is to adopt two quality features, namely the apparent source width (ASW) and the listener envelopment (LEV). ASW describes the perceived width of a sound scene, while LEV is associated with the feeling of being enveloped by sound, i.e. the effect of feeling 'inside' and 'surrounded

by' the reverberant sound of the room [125]. Since the perception of ASW and LEV depends on the room properties, changes in the presented configurations are taken into account. The shape and the acoustical properties of the enclosure, namely its absorption and reflection properties have a major influence on ASW and LEV [125].

On the other hand, objective findings such as physiological changes are difficult to measure for emotional response and immersivity to sound and music [126] as it does not take into account cognitive aspects. Therefore for a phenomenological experience (experienced from the first-person point of view), subjective reports were used for the assessments.

5.4.2 Test method

The test is divided into two sub-tests: in the first the evaluation of the ASW and LEV, in the second the evaluation of the system in a qualitative point of view.

Appearent source width and listener envelopment evaluation

The test is presented in the form of single presentations with reference. One subject at time is involved in the listening test and two stimuli ("A","B") are presented in sequence. The stimuli are presented in Table 5.2, where *early part* means we are considering only the early reflections in order to simulate a listening environment and *late part* means that only the late reverberation is considered. The subject is asked to rate the stimulus "B" with respect to the stimulus "A", in terms of ASW and LEV. The sound source is modeled using two virtual loudspeakers. Since the properties of the enclosure are found to have a major influence on the perception of ASW and LEV, we decided to consider concert-hall described in Section 4.1 as the environment for this test. Indeed, this kind of environment has been designed with a transition time between the early and late reflections reflections of 70 ms, according to the fact that the room reflections related to the sound field arriving later than 50 ms after the direct sound influence the feeling of envelopment [127]. In addition it is a medium size room with reflection coefficients that ranging from wall to wall, that is a feature that could improve the sense of Appearent Source Width; moreover, among our listening environments, it is the room with a longer reverberation tail, that is an aspect that influences the perception of the two criteria used in this test. The reflected properties of the different room configurations have been chosen accordingly and are listed in Table 5.2. The simulated room has been designed according to the characteristics mentioned in Chapter 4.

Room case name	Early part	Late part
no room	no	no
early room	yes	no
late room	no	yes
full room	yes	yes

Table 5.2. Properties of the presented room cases.

The first test deals with the influence of all four room simulation configurations being tested against each other. Table 5.3 summarizes all the tested combinations of reference-test pairs. The listening position is held constant for all configurations.

Configuration	Reference stimulus	Test stimulus
1	no room	early room
2	no room	late room
3	no room	full room
4	early room	late room
5	late room	full room
6	early room	full room

Table 5.3. Test stimuli for the evaluation of test condition 1.

In the second test we want to investigate the influence of increasing the distance from the virtual sources, while adding late reverberation to the sound field at the same time. To do so, we define a set of listening positions in which the listener moves away from the sound source on the vertical axis: the distances are shown in Table 5.4.

Configuration	Reference stimulus	Test stimulus	Distance
7	early room	late room	0.75 m
8	early room	late room	2.75 m
9	early room	late room	4.15 m

Table 5.4. Test stimuli for the evaluation of test condition 2.

The listening test is divided into a training session and the main test, which is twofold. In the training session, the distance and room configurations are presented in their maximum characteristics, in order to train the test subjects to the perception of ASW and LEV, respectively. In the main test the audio sample is presented as a test signal. A paired comparison listening test, also referred to as A/B-test method, is used [128]. The test subjects are instructed to rate ASW and LEV of the *test stimulus*, compared to as a *reference stimulus*. The subject is asked to rate according to the seven-grade comparison scale reported in Table 5.5, which is introduced in *Recommendation* ITU-R BS.1284, where 3 means that the test stimulus has a higher ASW or LEV compared to the reference stimulus. A rating of -3 indicates a smaller ASW or LEV of the test stimulus, respectively. If no difference can be perceived the value 0 has to be chosen.

Comparison	
3	Much better
2	Better
1	Slightly better
0	The same
-1	Slightly worse
-2	Worse
-3	Much worse

Table 5.5. Seven grade comparison scale.

Evaluation of a Virtual Sound Source

Our evaluation includes a post-experiment questionnaire. The questionnaire is developed to identify the effect of realism as well as the user's perception of the utility of the system, and the level of user satisfaction with the interaction.

We classify our observations by learnability and ease of use, utility, user satisfaction, interface features and realism. In terms of

- *learnability and ease of use*, we mean if the system is easy to learn, how the user quantifies the usability and if he or she is able to use the tool without difficulty;
- *utility of the tool*, we want to know from the users if the system is useful and it would improve the listening experience with respect to traditional music reproduction techniques;
- *user satisfaction*, we intend if the user likes the tool and if he or she is satisfied by the system response, in terms of response velocity and in terms of realism;
- *interface features*, the interest concerns the interface indication consistency with the hearing and if it helps the user in the judgement;
- *realism*, we concern if the sound respects the overall ambience and the head rotation.

It is known from psychoacoustics that head movements improve the performance in localization experiments [88]. To evaluate the stability of the reproduced sound-image against movements of the listener's head as well the immersivity an additional test is conducted. The implemented system expects that the user listens to a single music sample in a fixed position in a virtual room in which the head rotation is permitted. After that, the procedure has a quantitative quality evaluation in which data are collected with the questionnaire.

The following five-grade scale (Table 5.6) is used for the subjective assessment of sound quality as suggested from the ITU-R BS.1284-1.

	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 5.6. Five grade quality scale.

5.4.3 Statistical analysis

The statistical analysis procedure adopted for the first test is identical to that presented in Section 5.2.3 in the case of sound localization. With regard to the questionnaire, we decide to represent the mean assessments using an histogram.

Chapter 6

Experimental results

In this Chapter we show the results of the perceptual tests presented in Chapter 5 to obtain a subjective evaluation of the system. After the user responses were collected following the presented test methods, we need to obtain some statistical results and draw some conclusions.

In the previous Chapters we described some problems that arise when a structural non-individualized HRTF is used, and so we expect that some of these will be present in our results. The tests aimed at evaluating different aspects: the capability to locate a sound source in space, the efficiency of the system to synthesize a moving virtual listener, the immersivity of the output signals and the response to the head rotation.

The tests covered the aspects presented in Chapter 5; some are related to static conditions, in which the position of the listener, as well as his or her head, are fixed in the virtual environment; some others, are related to dynamic conditions, i.e. the listener moves inside the environments, possibly rotating the virtual head. The aspects interested in the tests are presented in Chapter 5, and some are related to static conditions in which the listener is fixed while others are related to a dynamic conditions, in which the listener is allowed to move.
This Chapter is structured as follows. In Section 6.1 we show the results of the listening tests, including a detailed statistical analysis. In Section 6.2, on the base of the complete set of results, we point out the strengths and the weaknesses of the system.

6.1 Results

In this Section we present the results of the formal listening tests aimed at assessing the capability of the system to simulate sound sources in space. The results are collected through the tests conducted according to the protocols presented in Chapter 5. The number of people involved was 21 and the average age was 26.47 years old; they were not experienced in listening to sound in a critical way.

6.1.1 Sound localization results

In the following we analyze the results obtained from the responses of the users involved in the sound localization test. This test aims at assessing the capa-

bility of the system to render a virtual source in a virtual environment. This test has been conducted according to the guidelines presented in Section 5.2.2. For this purpose, a speech audio signal (a male speech extracted from the European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) CD [114]) was generated from seven positions around the virtual listener at $\theta = -45°, 0°, -160°, 45°, 65°, -20°, 135°$.

The different angles of arrival of the sound corresponding to the different stimuli are depicted in Figure 6.1.

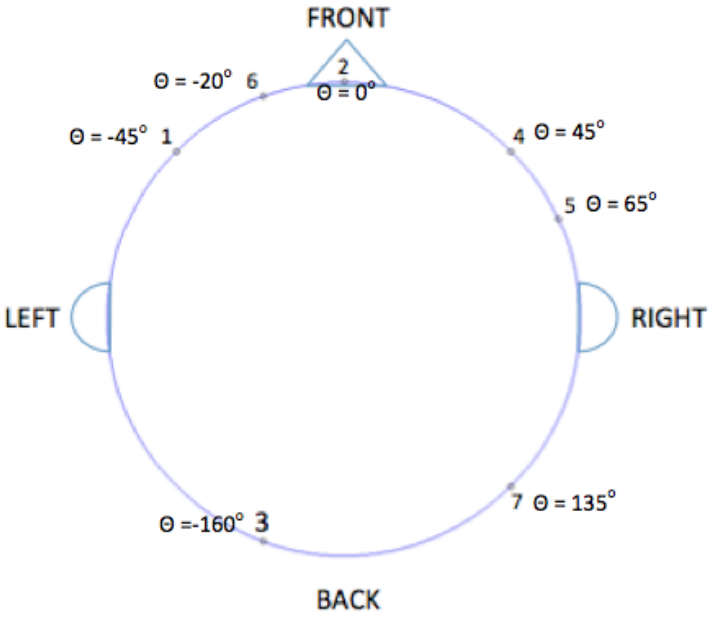

Figure 6.1. The points depicted represent the different angle of arrivals tested. The numbers near the points indicate the numerical designation of each virtual sound.

The results analyzed are presented by concentrating to the difference between the localization of the synthesized sound source and the perceived angle of arrival by the users. The test has been conducted in two different conditions: in the first the environment was not modeled, i.e. only the direct path from the source to the listener positions was rendered; the test was then repeated including the reverberation. In particular, for this test we considered the environment representing a concert hall, described in Chapter 4. The subjects were asked to indicate the perceived angle of arrival.

The Figures 6.2 and 6.3 show the mean of the judgments expressed by all the listeners and the relative confidence interval, computed with the procedure presented in Section 5.2.3. In particular, we show the mean results depicted like blue dashes, the confidence intervals depicted between two black dashes representing the upper limit and the lower limit, and the synthesized angle depicted like red dashes, for the two listening enviroment configurations that have been presented to the listeners.

In Figure 6.2 we show the results of the subjective test in an anechoic room. Some subjects perform better than the others: we are interested in the mean results and the confidence intervals, computed with the procedure presented in Section 5.2.3.

Along with these, for each stimulus also the mean confidence value of the responses are shown.

The first stimulus was synthesized at -45 degrees, in the front hemisphere: the listeners were able to perceive that the sound arrived from the left side but they were pretty sure that the angle was -91 degrees.

In the second stimulus the mean confidence value is smaller and the confidence interval is wider: indeed, the results relative to this stimulus have been affected by the known front-back confusion effect. Despite this the most of the users were able to identify that the sound source was in the front, and indeed the mean value is -2°.

The results of the third stimulus shown that with sufficient confidence the users perceived that the sound was located in the back hemisphere, however the mean value of the responses is -113°, i.e. the average error is of 47°.

For the sound source positioned at 45 degrees, i.e in the fourth stimulus, as for the case of the symmetrical with respect to the median plane stimulus, i.e the first stimulus, the situation is the same. With a certain confidence the user perceived the sound arrived at 92 degrees and the narrow confidence interval show that this angle of arrival is not well rendered.

Also a sound at 65° is perceived wrongly, arriving from 94° for the users, even if in this case the confidence interval is slightly wider, but the users were quite confident about their answers. In other words, the user perceived a sound rendered at 65° almost in front of the right ear.

The results of the sixth stimulus are coherent with the observations made so far; indeed, a sound source synthesized at -20° is perceived as arriving from -85°. Moreover, a narrow interval confidence and a high mean confidence value of response indicate that the users when responding were sure about the response given.

In the last stimulus, the correct value is just outside the confidence interval. The sound source was at 135° but the mean response value is 112°. Also in this case like in the third stimulus, the users perceived that the sound was located in the back, but they were not able to indicate the exact angle of arrival.

We notice from an analysis of the results that the listeners are able to understand if the virtual source was located in the left or right hemisphere. The results show that the users didn't perceive intermediate angles between 0° and 90° (the same hold for the specular with respect the median plane i.e between 0° and -90°): indeed a source located at 45° or at 65° is likely to be perceived around 90°.

In the second configuration tested, the sound sources were positioned at the same locations as in the anechoic room. By adding reflections the system should improve the sense of realism but we expect a degradation of the capability of the users to localize a sound source. Therefore, we expect greater localization errors.

The results are shown in Figure 6.3: in the following we comment on the individual results.

The first observation is that, as expected, the general mean confidence values are reduced: this indicates that the users were less sure about the given responses. Following this trend, also the confidence intervals are wider.

As in the anechoic condition, the first stimulus arriving from -45°, is perceived as arriving from -90°. The user were able to localize the sound as arriving at the left ear, but not to indicate the correct angle of arrival.

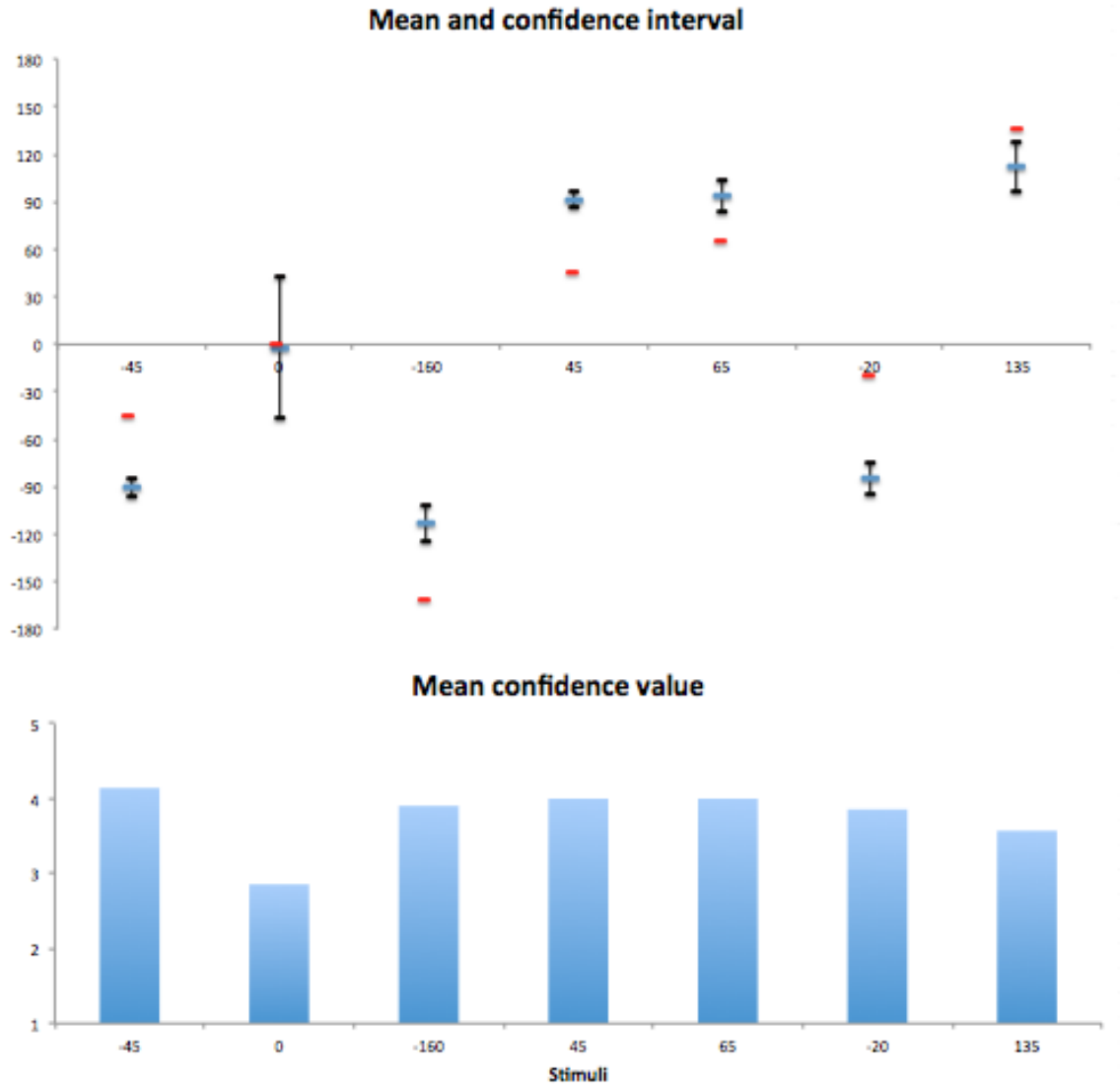

Figure 6.2. Representation of the results for the first test in an anechoic condition, in terms of mean values (blue dashes) and confidence intervals (black boundaries). For each stimulus is also shown the mean confidence value of the given response.

The results of the second stimulus present a worsening with respect to the previous results. The confidence interval has become wider and even if the mean value is located inside this interval, the sound source synthesized at 0° is perceived as arriving from -34°. The reflections deceived the users.

The results of the third stimulus are affected by the front-back confusion: the users perceived a sound arriving from -62° while the source is located at -160°. In addition, the mean confidence value for the responses of this stimulus is very low with respect to the other ones, and the confidence interval is wide. The results suggest that the users were only able to correctly perceive the sound in the left side, but they were wrong in indicating the frontal hemisphere.

In the fourth stimulus there is a slight improvement with respect to the anechoic

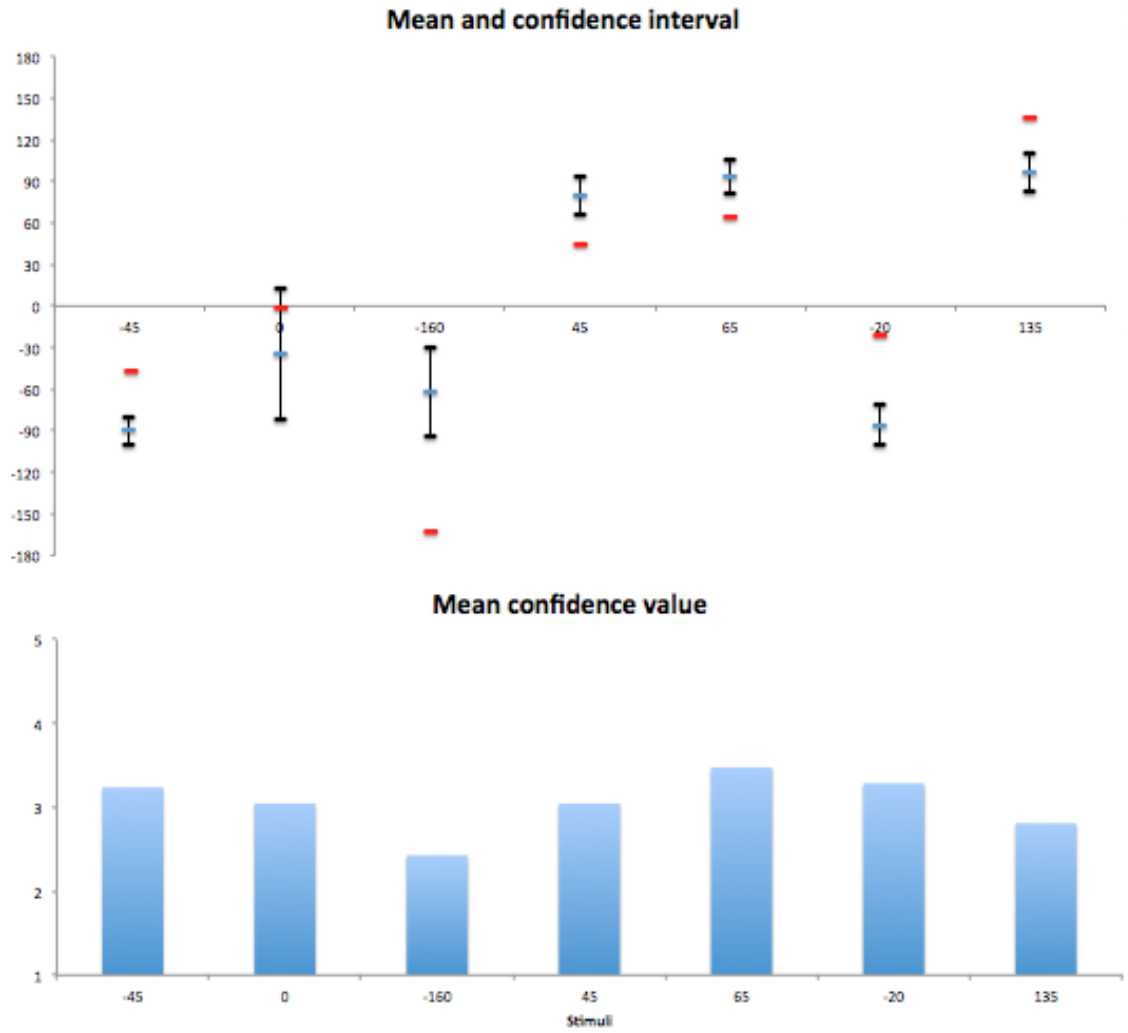

Figure 6.3. Representation of the results for the first test in a reverberant room, in terms of mean values (blue dashes) and confidence intervals (black boundaries). For each stimulus is also shown the mean confidence value of the given response.

condition even if also in this case the users perceive the sound arriving from a different angle. The mean value of the responses is 80°, while the sound arrived from an angle of 45°.

In the fifth stimulus a sound source at 65° was rendered, and the users responded with a mean confidence of 3.4 that the source was located at 93°. The confidence interval is not too wide, and it does not include the correct value. As in the anechoic condition, a sound at 65° is perceived around 90°.

The results of the sixth stimulus shows that there is a bit of uncertainty around the mean value of responses - 86°, but the correct value of - 20° is far. Also in this case, the users perceived the sound in the left hemisphere as arriving more or less from - 90°.

The results of the last stimulus are worse than those in anechoic conditions: indeed, the results show that the users were not able to perceive well that the sound

was in the back hemisphere. The sound source was located at 135°, but the mean value of the responses is 97°, with an upper limit of the confidence interval that is 111°.

We can state that also in reverberant conditions the users were not able to perceive intermediate angles between 0° and 90° and its symmetrical with respect to the median plane. In addition, in a non anechoic room the front-back confusion is more evident as the users were not more able to precisely separate the sound sources located at the anterior or at the posterior angles.

6.1.2 Trajectories localization results

The second listening test has been performed presenting at the users a sound stimulus (a male speech extracted from the European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) CD [114]) along with a visual reference, i.e a representation of three different graphical trajectories describing the paths followed by a virtual listener inside the environment. The task of the user was to indicate which of them represents the stimulus heard. The listening environment used is the concert hall model described in Chapter 4 since its reverberant characteristics at the same time provide good distance cues but do not degradate so much the localization capability of the user.

We presented four different stimuli, and the corresponding trajectories are depicted in Figure 6.4 where the arrow indicates the orientation of the head and the direction of the movement while the filled black circle represents the starting point and the red circle the virtual sound source.

For the first stimulus, at the users were presented a paper on which there are represented three different possibilities, depicted in Figure 6.5, where at each trajectory is assigned an identification number. The user is asked to indicate a number between 1,2 and 3, that indicates the choice done. The trajectory marked as 1 corresponds to a virtual listener that starts far away from the virtual sound source and approaches it, and that a given time rotates to the right. The trajectory marked as 2 illustrates a moving listener away from the source that starts from the left side of the room and goes toward the opposite side. In the third trajectory is represented a listener starting away from the virtual sound source and approaching it.

For the second stimulus, a new paper with three different graphical trajectories was presented to the users. The paper for the second stimulus is depicted in Figure 6.6. The first trajectory represents a virtual listener that starts moving from the left side toward to the right side, then rotating of 90° to the right and moves away from the source. At a given time it rotates again of 90° to the right and moves toward the left side. The trajectory marked as 2 represents a virtual listener moving away from the sound source. The trajectory number 3 corresponds to a virtual listener that moves from the right side to the opposite side.

For the third stimulus at the users were presented the combination of graphical trajectories depicted in Figure 6.7. The trajectory marked as 1 represents a virtual listener that at the same time moves from the left side to the right side crossing the room and moving away from the sound source. The trajectory number 2 corresponds to a virtual listener that moves from the right side to the opposite side. Finally, the third trajectory illustrates a virtual listener that moves away from the sound source.

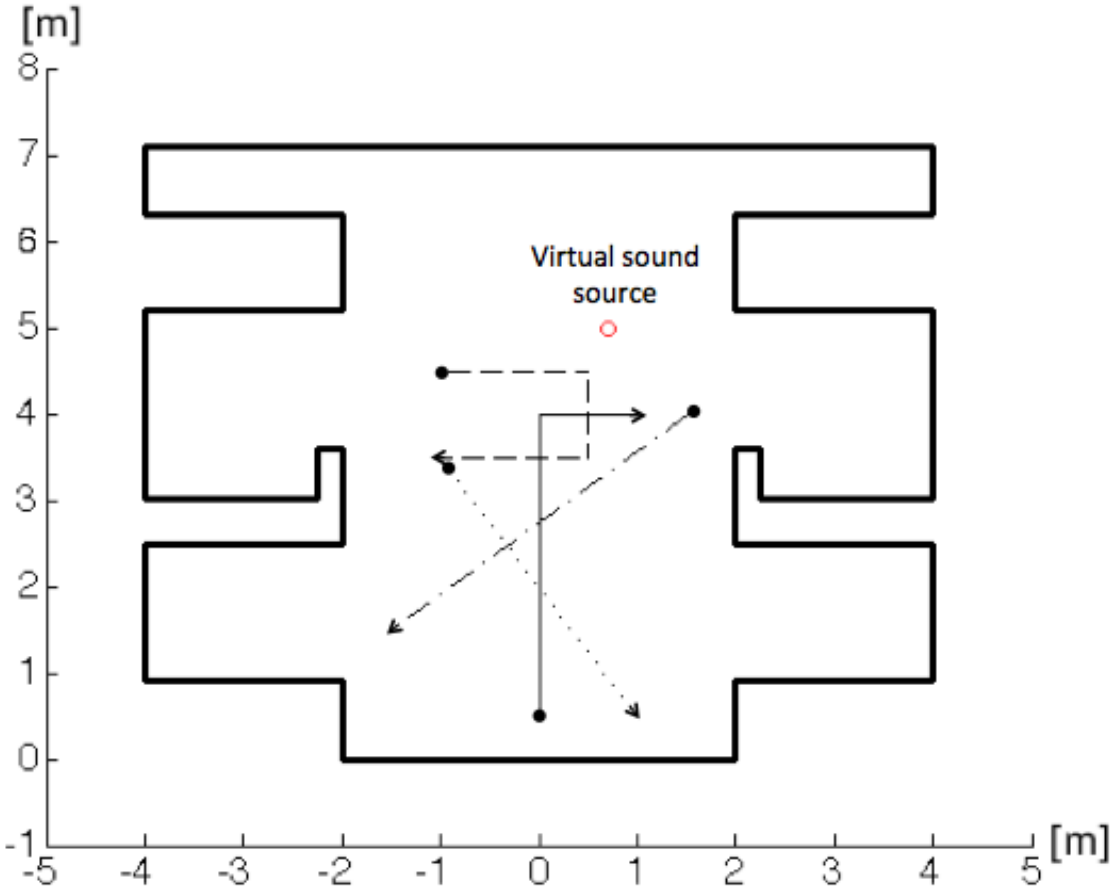

Figure 6.4. The trajectories and the virtual sound source position used in the second perceptual test. The numbers near the end arrows indicate the numerical designation of each trajectory.

For the fourth stimulus at the users were presented the combination of graphical trajectories depicted in Figure 6.8. The trajectory number 1 corresponds to a virtual listener that at the same time moves from the right side to the opposite side crossing the room and moving away from the sound source. The trajectory marked as 2 illustrates a virtual listener that moves away from the source and at a given time rotating of 90° to the left and then continuing to move. In the third trajectory is represented a listener that starts away from the virtual sound source and approaches it.

For all the stimuli, the correct trajectory is the first one. The results of this perceptual test are depicted in Table 6.1 in a matrix form: on the rows it is possible to observe the percentage of the responses for each stimulus. It is possible to note that for three stimuli out of four, the most of the users identificated the correct trajectories. Observing this table we can note that for the first stimulus 16 users over 21 identified the correct response, for the second stimulus 18 users over 21, for the third stimulus 12 over 21 and for the fourth stimulus 7 over 21.

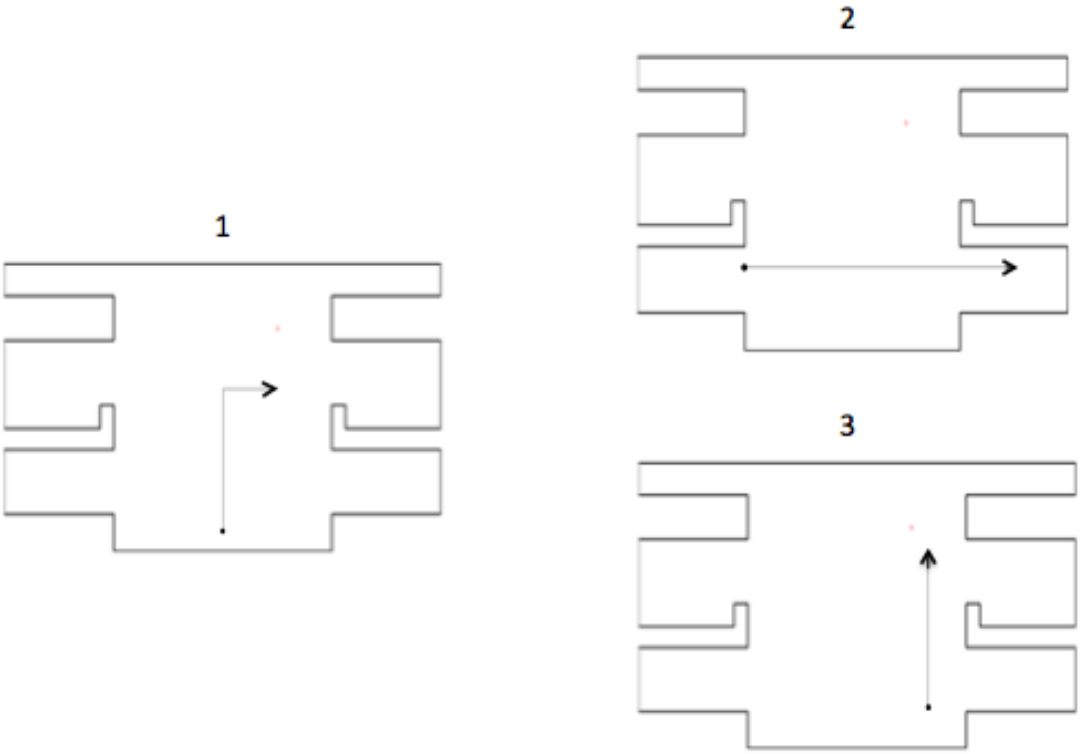

Figure 6.5. Combination of three trajectories presented at the user for the first sound stimulus.

Observing the first paper and the results for the first stimulus, we can note that the most of the users were able to identify the correct response with respect two wrong linear trajectories: in fact we can conclude that the change of direction in the correct trajectory of the virtual listener, permitted to the listener to discard the second and third trajectories, that are linear. In addition, the correct trajectory provides an initial approach toward the virtual source, that is used by the users to discard the second choice.

Evaluating the second stimulus, we can note that also in this case the correct trajectory presents a change in direction, while the wrong ones are straight. Also in this case the most of the users were able to identify the correct response (Table 6.1). In particular a double change in direction corresponding to a change in the interaural differences helped the users in their choice: the users first heard the stimulus over the left ear, and then, after a double change in direction, over the right ear. The other two possibilities always expected a sound over the right ear. Indeed, this stimulus is the most correctly identified.

The third stimulus has been correctly identified by the most of the users but we can note (Table 6.1) that the percentage of correct responses is lowered with respect to the first two stimuli. Almost all the listeners were able to discard the third trajectory that doesn't present a departure from the virtual sound source, but not all were able to perceive the direction of the sound: probably many of the users

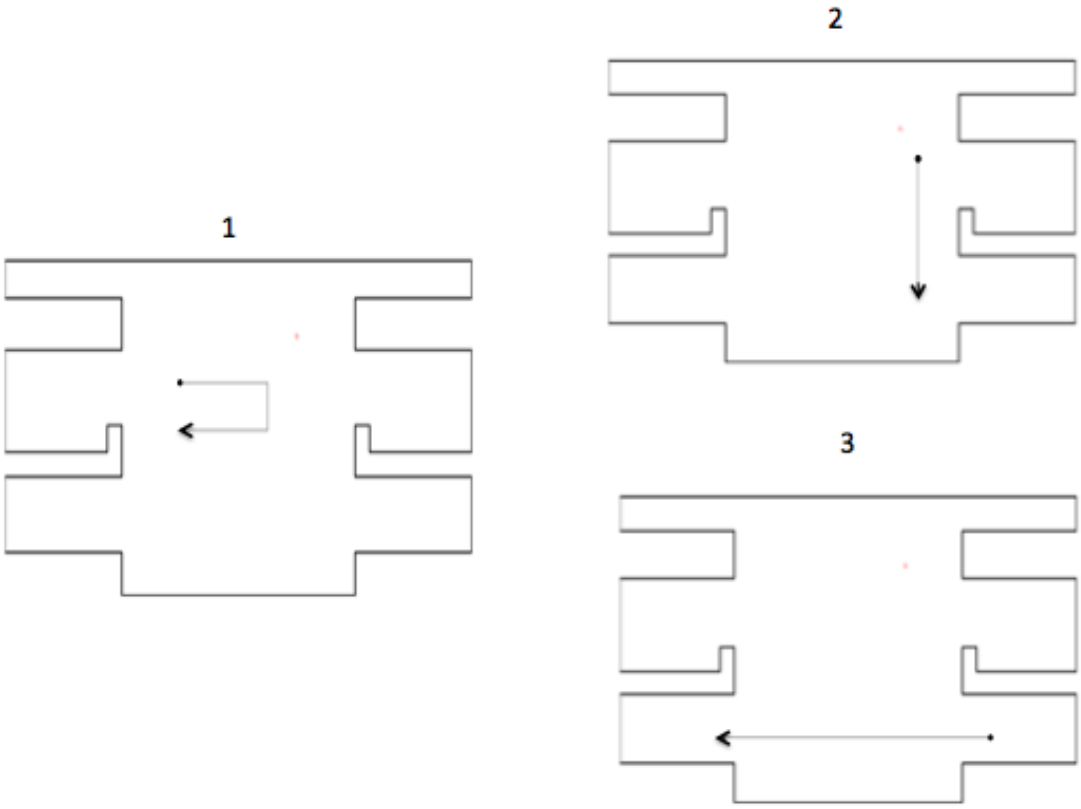

Figure 6.6. Combination of three trajectories presented at the user for the second sound stimulus.

based only on the change of the stimulus intensity and on the direct-to-reverberant ratio. Indeed, in the first trajectory the user mainly perceived the sound with the left ear, while in the second the converse is true. In conclusion, in this stimulus where a change in direction is not present, the users presented some difficulties in making the right choice.

In the last stimulus the most of the users, in particular almost two out of three, were not able to identify the correct response. As in the first and second stimuli, one of the trajectories presents a change in direction. However, this does not correspond to the correct one, which represent a moving listener that starts close to the virtual sound source and moves away toward the opposite side of the room: thus the user should perceive a sound that decreases in intensity and that is perceived mostly on the right ear until it is perceived more or less on the back. Because of this increasing in distance almost all of the users were able to discard the third choice, that corresponds to a moving listener that approaches the sound source.

We can conclude that the rotation of the virtual listener's head is an important factor for the user in order to identify the correct trajectory, as well the distance cues like the variations of intensity of the sound and the direct-to-reverberant ratio. In general better results are obtained for non linear trajectory of the moving virtual listener, and only for one stimulus over three there were severe identification

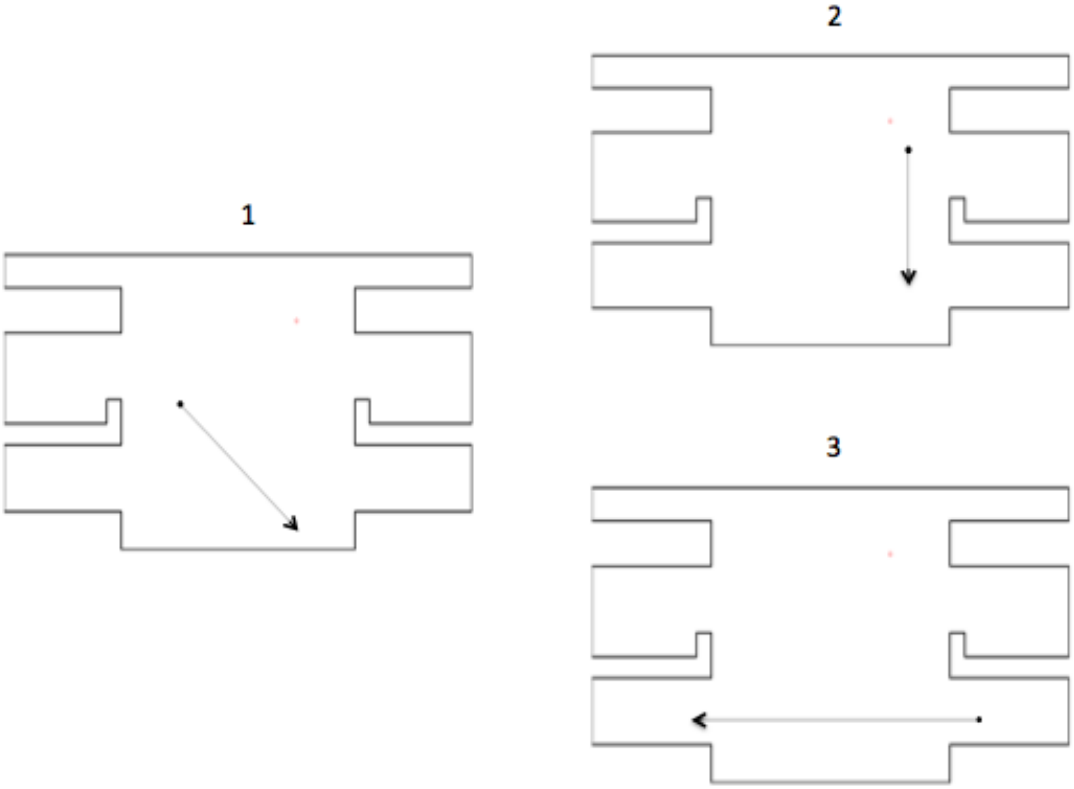

Figure 6.7. Combination of three trajectories presented at the user for the third sound stimulus.

problems.

6.1.3 Rendering system results

Appearent source width and listener envelopment evaluation

A listening test has been conducted to investigate the capability of our implemented system to render an acoustic scene and provide immersivity to the user. In order to evaluate the influence of the room in the perception of a realistic sound and to evaluate the immersivity given by our system, we use two descriptors introduced in Chapter 5.

ASW is a descriptor for the perceived width of a sound scene or a sound source within an acoustical scene. It has been found, that the perception of ASW strongly relates to the early part of the sound field, i.e. within a time frame of up to 50-80 ms arriving after the direct sound [129]. Especially the lateral reflections have a major influence on the perception of ASW. LEV is described as the feeling of being enveloped by sound or the effect of feeling "inside". It is mainly dependent on the late reverberating part of the sound field [129]. It is important to note that ASW and LEV influence each other in a perceptual way. When the LEV perception increases, this translates in an increase of the the perceived ASW [130].

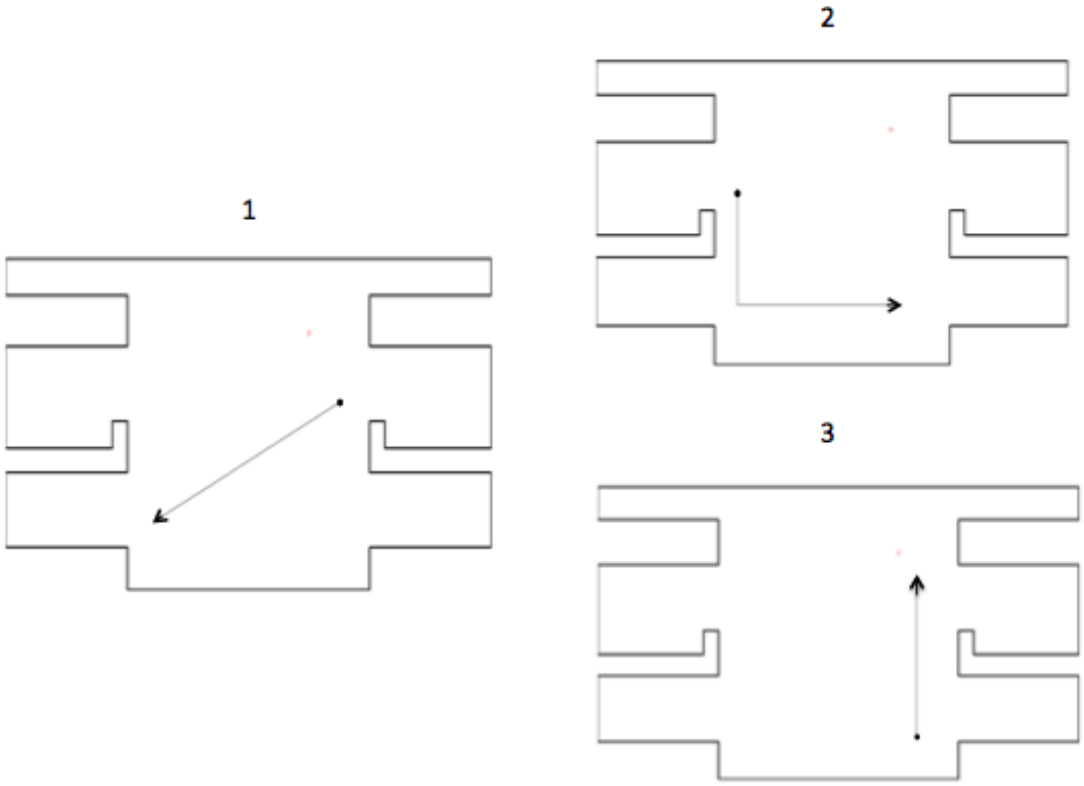

Figure 6.8. Combination of three trajectories presented at the user for the fourth sound stimulus.

A concert-hall model (Section 4.1) was employed to model early and late room reflections. An ensemble of two virtual sound sources was simulated and the virtual listener was placed in the sweet-spot. The listeners were asked to evaluate the perception of ASW and LEV for a file audio (a musical sample, in particular an excerpt from Suzanne's Vega *Tom's Diner*) in a given room configuration with respect to a reference stimulus heard before. The compared room configurations are shown in Table 6.2.

The listening test results are depicted in Figure 6.9 with 95% confidence for ASW and LEV, respectively. The results for all configurations indicate, that the rating of ASW correlates with the perception of LEV. For tests configurations 1-3, the different room configurations, namely *early room*, *late room* and *full room*, are compared to the dry scene.

The result for the first configuration shows that adding only early room reflections to the scene slightly increases the impression of a broader virtual sound source width. Also the feeling of LEV is slightly increased. Presenting only the late reverberation part of the room significantly enhances this effects (configuration 2). A room with both, early and late reverberation is rated similarly (configuration 3).

Stimuli	Response 1 (correct)	Response 2 (wrong)	Response 3 (wrong)
First stimulus	76.19%	0%	23.81%
Second stimulus	85.71%	9.52%	4.76%
Third stimulus	57.14%	38.10%	4.76%
Fourth stimulus	33.33%	61.90%	4.76%

Table 6.1. Representation of the frequency of responses as a percentage in the second test.

Configuration	Reference stimulus	Test stimulus
1	no room	early room
2	no room	late room
3	no room	full room
4	early room	late room
5	late room	full room
6	early room	full room

Table 6.2. Test stimuli for the evaluation of test condition 1.

In configuration 4 the early and late part of the sound field are compared to each other. It can clearly be seen, that the late reverberation has a larger influence on ASW perception than early reflections, so the perceived width is significantly rated wider. This also holds for the perception of LEV.

Configuration 5 deals with the influence of early reflections in a late reverberating environment, showing that they only lead to no or at most to a very small increase in ASW and LEV perception. The small confidence interval indicates the test subjects have been quite certain about this.

The influence of late reverberation is tested by using configuration 6: comparing a full room with the early part of the room, we may notice that ASW and LEV are rated similar to configurations 4. This reinforces what has been observed in configuration 5, i.e. the main perceived part of the sound is the late reverberation.

After the comparison of different room configurations, also the influence of different listening positions was evaluated. We compared a configuration with only the late reverberation with respect to a configuration with only the early reflections as a reference: what changed from stimulus to stimulus was the listening position of the virtual listener, that moved away on the same vertical axis. The compared configurations are shown in Table 6.3, where it is shown the distance of the listener from the virtual sources.

In configuration 7 the results show that near the virtual sound sources, a listening environment modeled only with late reverberation gives the user an impression of a broader virtual sound source width with respect to a listening environment modeled only with early reflections; in addition, it enhances the sense of immersivity.

When the position of the virtual listener has moved away from the source, but it is not so far, as in configuration 8, the virtual sound source width results even more broader to the users, and also the feeling of envelopment increases.

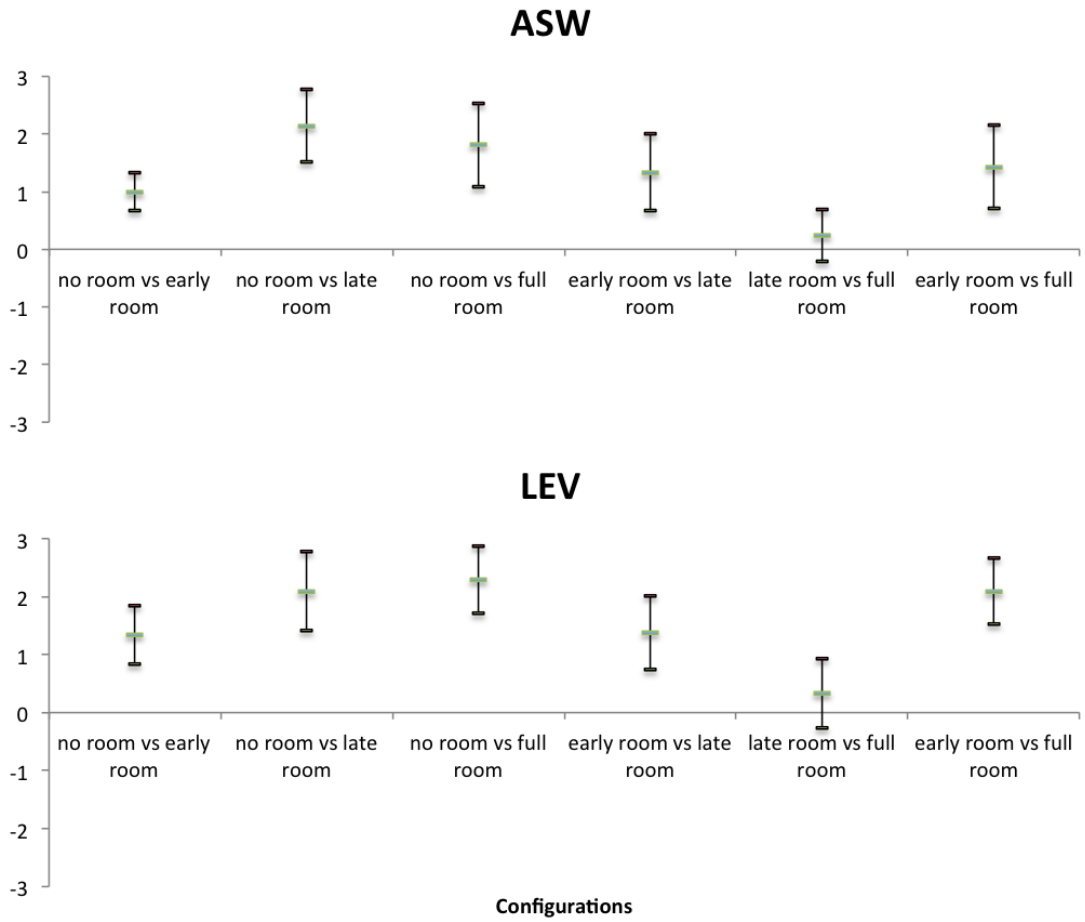

Figure 6.9. Influence of different room configurations on ASW and LEV perception. The listening position is held constant.

When the early reflections are added to a dry sound, a slight increase of ASW perception is noted: in addition LEV is rated a bit higher. On the other hand adding the late reverberating part of the sound field strongly increases the impression of LEV and ASW, and when the late reverb is considered, early reflections seem to have no significant influence on the perception of ASW and LEV. Finally, for a listening position away from the source the results are similar to the previous configuration: there is however an increasing of ASW and LEV with respect to a configuration with only early reflections.

We must note, however, that all the results present a wide confidence interval, i.e., they show a bit of uncertainty. Nevertheless, we can say that increasing the distance from the source, the listener perceives a wider sound source and an higher sense of immersivity. In particular, the listener should perceive a more reverberant sound that means an increase of the LEV descriptor. This increase translates also into the perception of a wider sound source, i.e in an increase of the ASW.

Configuration	Reference stimulus	Test stimulus	Distance
7	early room	late room	0.75 m
8	early room	late room	2.75 m
9	early room	late room	4.15 m

Table 6.3. Test stimuli for the evaluation of test condition 2.

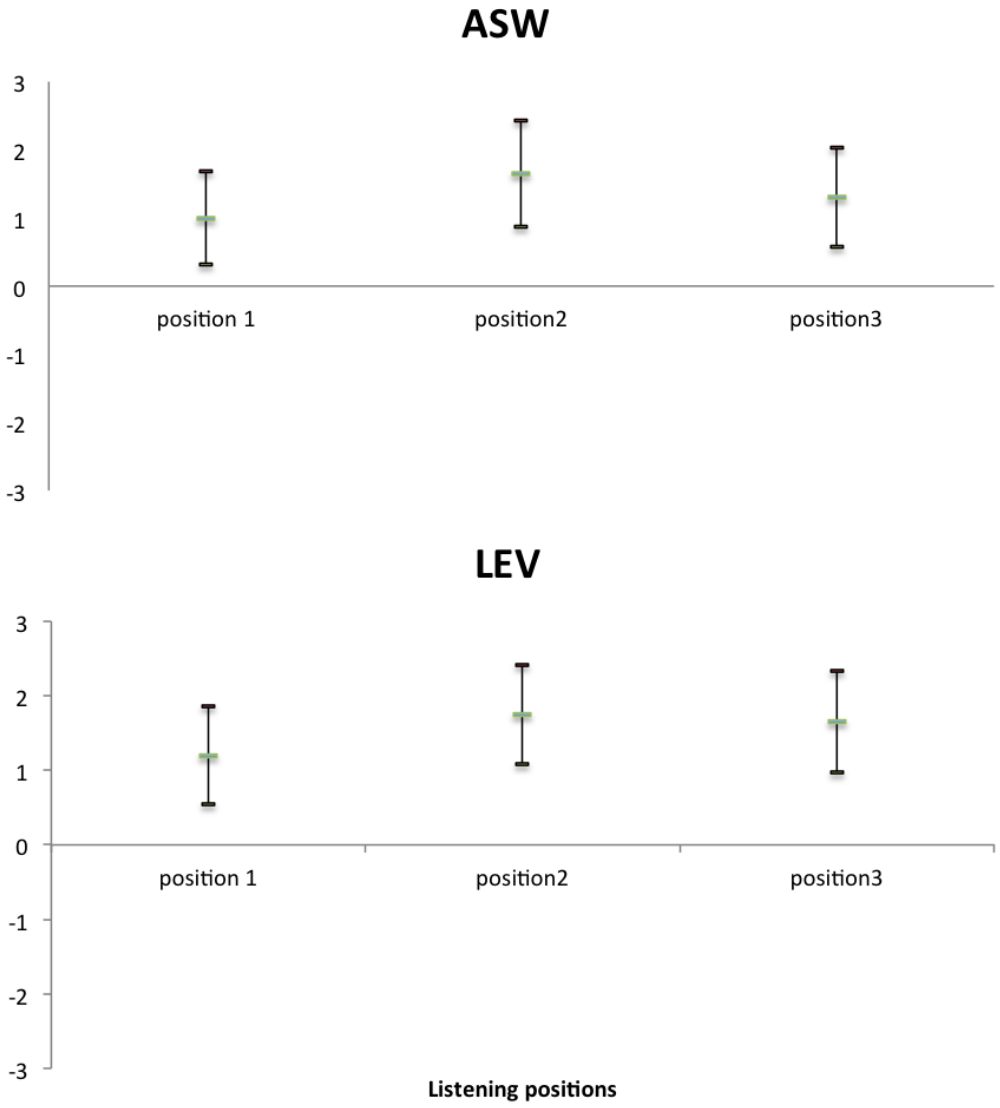

Figure 6.10. The trajectories and virtual source position used for the trajectory test. The numbers near the end arrows indicate the numerical designation of each trajectory.

Qualitative evaluation of the system

Finally, we were interested in experiences from the first-person point of view, in which subjective reports were used for the assessments. The overall quality evaluation

criteria used in this test are described in Section 5.4.2.

In Figure 6.11 we show the results of the subjective test, in form of the mean judgments expressed by all the listeners.

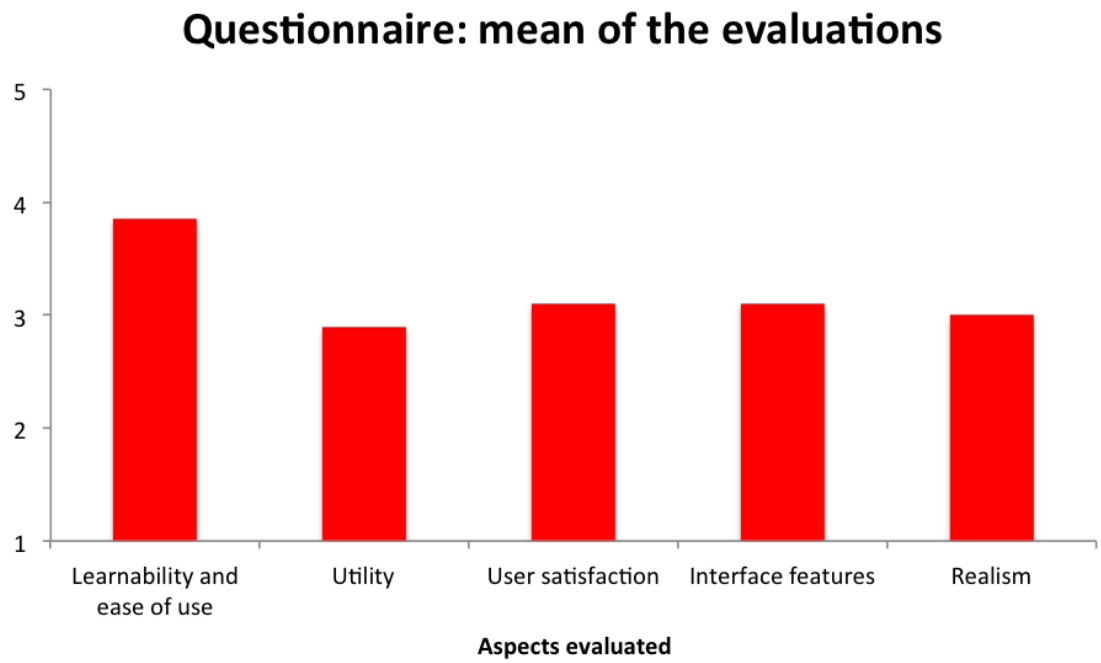

Figure 6.11. Histogram representing the mean of the user responses in the questionnaire.

In terms of learnability and ease of use, the participants ranked the system highly (3.85 average on a scale with a maximum value 5) and stated that they were able to use the tool without difficulties; in conclusion, the users have found the system features easy to understand.

In terms of utility, the participants ranked the system with a mean value of 2.9 over a maximum of 5. In particular, all the users stated that it would be nice to be able to choose between the traditional listening experience through headphones and this one, but some of them stated they have some difficulties in thinking about a use of this implementation in their everyday life. We can note that the evaluation of this criteria was greatly influenced by the use of a MIDI controller instead of a head-tracker: indeed the users thought about a use in real life of the system with the same configuration.

In terms of user satisfaction, we can state that in general the users like the tool. One user stated a preference for faster response time; some other users observed that the system was able to provide a new way of listening sounds through headphones that they had never tried. Overall, the considered aspect was evaluated with a mean value of 3.1 over 5.

In terms of interface features, all the users agreed about the utility of the head motion cue stating that helped in a better comprehension of the situation. Some of the users stated that the interface was a bit too basic, but at the same time that it offered all the information needed for the listening experience. This criteria obtained

a mean value of 3.1/5.

Finally, the degree of realism was ranked with a mean value of 3 over 5. Some users stated that the system was able to provide a sense of immersivity in the sound and that they enjoyed the listening. In particular all the users agreed that the rotation of the virtual listener's head corresponded to a coherent change in sound perceived.

6.2 Conclusions

The first test confirms what was already known in the literature, i.e a system that uses non-individualized HRTF presents difficulties in render static sound sources. As the presented results have shown, some users perceive very different from ohters: maybe a post-screening of the results could be helpful. A post-screening operation is a rejection technique performed after having gathered all the results, in order to reject results from unreliable subjects. This operation is primarily used to eliminate subjects who cannot make appropriate identification or discriminations. The application of a post-screening method may clarify the tendencies in a test result. In addition, the analysis of the results were complicated, by the front-back confusion.

In general, the results show that the users were able to localize a sound in the left or in the right, and also in certain conditions, to separate a sound in the front from a sound in the back. On the other hand, the users perceived the sound as arriving from four quadrants and were not able to identificate the correct angles of arrival. The sources located in the front hemisphere on the left, were all identified as arriving more or less from -90°. By adding the reverberation, there was a reduced capability in front-back separation and the responses given by the users ranged over larger intervals. The localization errors however increased along with a decrease of the mean confidence value of the responses.

The experiments related to the moving virtual listeners confirmed that motion cues, and especially the head rotation, play a fundamental role in sound localization. Indeed, the results of such tests are positive, and, in particular, the users perceived very well the direction changes.

We can state that the presence of a reverberation tail significantly improves the appearent source width and also the listener immersivity. A configuration with only the early reflections is perceived more immersive than a dry sound. The presence of a listening environment enhances the sense of realism and immersivity, the presence of late reverberation significanlty increases the perception of a spatial listening.

A change in listening position, moving away the virtual listener from the virtual sound sources the perception is not so much affected.

The questionnaire reported the users satisfaction. In particular, the system was not able to position very well sound in space in static condition, but when the virtual listener is moving, and in particular the head, the users reported the perception of a surround sound that rotates around the head. The users experienced the perception of a sound in motion and a realistic listening. The results show that the implemented system is easy to use and that a change in the controller for the rotation of the head

can improve the overall perception, leading to higher judgement scores.

Chapter 7

Conclusions and Future Works

In this thesis we have proposed an implementation of a rendering sound system through headphones. In particular the features of our system allow the user to explore the virtual listening environment, navigate inside it and change the orientation of the virtual listener's head. The proposed system can find application, for instance, as an interactive music player.

In the literature we find two big classes of approaches aimed at implementing a spatialized system through headphones. On one hand, there are approaches based on the measurement of the HRTFs, that results in customized and expensive systems. The measurement procedure produces accurate HRTFs leading to very realistic sound reproduction. The main drawbacks are the prohibitive measurement time and the fact that the measurements are highly individualized. As a consequence, the implementation of such system in consumer products turns to be difficult. On the other hand there are approaches based on structural HRTF models, in which no measurements are required. These approaches are based on a combination of single localization cues models, that could be customizable or not. In this thesis we followed this second approach, which allowed us to obtain a computationally efficient and non-individualized system.

The requirements definition phase was much important. The work started with a research in the literature about existing guidelines and theory about sytems based on non-individualized HRTF. We were interested in understanding the system requirements for render a sound in space, around the listener's head. We collected information about the results and the test obtained from various researchers in order to identify the main important factors that influence the perception of a sound in space, in particular in a dynamic context. This phase led us to discover that the head motion overcomes some limitation derived from the use of a non customized system.

It is then followed a development phase started implementing the single cues involved in the listening. After we had a complete structural HRTF model implemented, we needed to think about the listening environments and in particular to the navigation inside it. Finally the integration of the head rotation was taken into account. The last step about the implementation regarded the user interface: indeed, the user has to be able to choose a song to be heard and we wanted reinforce the experience with a visual indication of the orientation of the head.

Finally, we conducted a set of perceptual tests aimed at evaluating both the

static and dynamic features of the systems.

The work done reported some interesting results and it allowed to have an implementation to work on and to improve. In particular we noticed that the results follow what reported in literature: it is very difficult to render a spatialized sound in static conditions, but the users reported satisfaction about the realistic motion of a virtual listener. Indeed, the listeners stated that the system was able to well simulate the experience of a sound source in motion around the head.

Future works

The implemented system can be used in different contexts: it is possible to integrate in videogaming, or in multimedia presentation, in order to introduce some visual cues in the case of dynamic situation.

It is possible to extend the subjective validation of the sound localization test through an experimental setup, in order to obtain measurements from a real sound source in a real listening room. This can be done by a comparison between the results of the implemented system in this thesis and the results in a real situation.

The context of this thesis should be considered like a prototype of a more complex and efficient application. The current implementation is developed in Matlab, and some features are not optimized to work in real time. A successive step is the implementation of this system over mobile systems. Finally, we can envision the possibility, for such a system, to be at the base of innovative applications in the field of augmented reality. For instance, one could imagine a spatial music browser, where the user is free to explore a virtual environment in which the audio contents are located in different space positions.

Bibliography

[1] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc*, 45(6):456–466, 1997.

[2] Thomas Sporer. Wave field synthesis - generation and reproduction of natural sound environments. In *Proceedings of the seventh International Conference on Digital Audio Effects*, 2004.

[3] K. Brandenburg, S. Brix, and T. Sporer. Wave field synthesis. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, pages 1–4, May 2009.

[4] http://interface.cipic.ucdavis.edu/sound/tutorial/psych.html.

[5] S. Spagnol, M. Geronazzo, and F. Avanzini. On the relation between pinna reflection patterns and head-related transfer function features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(3):508–519, March 2013.

[6] http://www.cns.nyu.edu/ david/courses/perception/lecturenotes/localization/localization.html.

[7] C.P. Brown and R.O. Duda. A structural model for binaural sound synthesis. *Speech and Audio Processing, IEEE Transactions on*, 6(5):476–488, Sep 1998.

[8] http://interface.cipic.ucdavis.edu/sound/tutorial/psych.html.

[9] V.R. Algazi and R.O. Duda. Immersive spatial sound for mobile multimedia. In *Multimedia, Seventh IEEE International Symposium on*, pages 8 pp.–, Dec 2005.

[10] D.R. Begault. *3-D Sound for Virtual Reality and Multimedia*. AP Professional, 1994.

[11] Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West. A beam tracing approach to acoustic modeling for interactive virtual environments, 1998.

[12] Navarun Gupta, A Barreto, and C. Ordonez. Improving sound spatialization by modifying head-related transfer functions to emulate protruding pinnae. In *SoutheastCon, 2002. Proceedings IEEE*, pages 446–450, 2002.

[13] V. R Algazi, R. O Duda, Ramani Duraiswami, Nail 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*, 112, 2002.

[14] Richard O. Duda, V. Ralph Algazi, and Dennis M. Thompson. The use of head-and-torso models for improved spatial sound synthesis. In *Audio Engineering Society Convention 113*, Oct 2002.

[15] C.P. Brown and R.O. Duda. An efficient hrtf model for 3-d sound. In *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, Oct 1997.

[16] V. Ralph Algazi, Eric J. Angel, and Richard O. Duda. On the design of canonical sound localization environments. In *Audio Engineering Society Convention 113*, Oct 2002.

[17] http://www.akaipro.com/product/lpd8.

[18] http://interface.cipic.ucdavis.edu/sound/hrtf.html.

[19] V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano. The cipic hrtf database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 99–102, 2001.

[20] Jonathan Sterne. *The Audible Past: Cultural Origins of Sound Reproduction.* Duke University Press, 2003.

[21] Paul Collins. Theatrophone: the 19th-century ipod. *New Scientist*, 197(2638):44 – 45, 2008.

[22] Frederic L. Wightman and Doris J. Kistler. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 1999.

[23] Masaharu Kato, Hisashi Uematsu, Makio Kashino, and Tatsuya Hirahara. The effect of head motion on the accuracy of sound localization. *Acoustical Science and Technology.*

[24] T. Lokki, L. Savioja, R. Vaananen, Jyri Huopaniemi, and T. Takala. Creating interactive virtual auditory environments. *Computer Graphics and Applications, IEEE*, 2002.

[25] Alan D. Blumlein. Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. *J. Audio Eng. Soc*, 6(2):91–98, 130, 1958.

[26] A. D. Blumlein. Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems, 1933.

[27] Ross H. Snyder. History and development of stereophonic sound recording. *J. Audio Eng. Soc*, 1(2):176–179, 1953.

[28] Mark F. Davis. History of spatial coding. *J. Audio Eng. Soc*, 51(6):554–569, 2003.

[29] J. C. Steinberg and W. B. Snow. Auditory perspective -physical factors. *American Institute of Electrical Engineers, Transactions of the*, 53(1):12–17, Jan 1934.

[30] V. Pulkki. *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. Helsinki University of Technology, 2001.

[31] Chris Kyriakakis. Fundamental and technological limitations of immersive audio systems. In *Proceedings of the IEEE*, pages 941–951, 1998.

[32] Darren B. Ward and G.W. Elko. Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation. *Signal Processing Letters, IEEE*, 6, 1999.

[33] http://www.dolby.com/us/en/technologies/dolby-digital.html.

[34] Dmitry N. Zotkin, Ramani Duraiswami, and Larry S. Davis. Rendering localized spatial audio in a virtual auditory space, 2002.

[35] Jens Ahrens, Rudolph Rabenstein, and Sascha Spors. The theory of wave field synthesis revisited. In *Audio Engineering Society Convention 124*, May 2008.

[36] W. Snow. Basic principles of stereophonic sound. *Audio, IRE Transactions on*, AU-3(2):42–53, March 1955.

[37] Yih Hsing Pao and Vasundara Varatharajulu. Huygens principle, radiation conditions, and integral formulas for the scattering of elastic waves. *The Journal of the Acoustical Society of America*, 59(6):1361–1371, 1976.

[38] Diemer de Vries and Peter Vogel. Experience with a sound enhancement system based on wavefront synthesis. In *Audio Engineering Society Convention 95*, Oct 1993.

[39] Mark A. Poletti. Three-dimensional surround sound systems based on spherical harmonics. *J. Audio Eng. Soc*, 53(11):1004–1025, 2005.

[40] Malham and A Myatt. 3-d sound spatialization using ambisonic techniques. *Computer music journal*, 19, 1995.

[41] Michael A. Gerzon. Periphony: With-height sound reproduction. *J. Audio Eng. Soc*, 21, 1973.

[42] Jerome Daniel, Sebastien Moreau, and Rozenn Nicol. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. In *Audio Engineering Society Convention 114*, 2003.

[43] Stéphanie Bertet, Jérôme Daniel, and Sébastien Moreau. 3d sound field recording with higher order ambisonics - objective measurements and validation of spherical microphone. In *Audio Engineering Society Convention 120*, May 2006.

[44] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93, 1993.

[45] Bruce N. Walker, Raymond M. Stanley, Nandini Iyer, Brian D. Simpson, and Douglas S. Brungart. Evaluation of bone-conduction headsets for use in multitalker communication environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(17), 2005.

[46] Raymond M. Stanley and Bruce N. Walker. Lateralization of sounds using bone-conduction headsets. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(16), 2006.

[47] W.G. Gardner. *3-D Audio Using Loudspeakers*. The Springer International Series in Engineering and Computer Science. Springer, 1998.

[48] Gary S Kendall. A 3-d sound primer: directional hearing and stereo reproduction. *Computer music journal*, pages 23–46, 1995.

[49] V.R. Algazi and R.O. Duda. Headphone-based spatial sound. *Signal Processing Magazine, IEEE*, 28(1):33–42, Jan 2011.

[50] V. Ralph Algazi, Jr. Dalton, Robert J., Richard O. Duda, and Dennis M. Thompson. Motion-tracked binaural sound for personal music players. In *Audio Engineering Society Convention 119*, Oct 2005.

[51] Bill Gardner and Keith Martin. Hrtf measurements of a kemar dummy-head microphone. Technical report, MIT Media Lab Perceptual Computing, 1994.

[52] Thibaut Ajdler, Christof Faller, Luciano Sbaiz, and Martin Vetterli. Sound field analysis along a circle and its applications to hrtf interpolation. *J. Audio Eng. Soc*, 56(3):156–175, 2008.

[53] Wen Zhang, Mengqiu Zhang, R.A Kennedy, and T.D. Abhayapala. On high-resolution head-related transfer function measurements: An efficient sampling scheme. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):575–584, Feb 2012.

[54] M. Pec, M. Bujacz, P. Strumillo, and A Materka. Individual hrtf measurements for accurate obstacle sonification in an electronic travel aid for the blind. In *Signals and Electronic Systems, 2008. ICSES '08. International Conference on*, pages 235–238, Sept 2008.

[55] Peter Balazs, Bernhard Laback, and Piotr Majdak. Multiple exponential sweep method for fast measurement of head related transfer functions. In *Audio Engineering Society Convention 122*, May 2007.

[56] H. Møller. *Fundamentals of Binaural Technology*. Aalborg University, Institute for Electronic Systems, Department of Communication Technology. Aalborg Universitetscenter, Institut for Elektroniske Systemer, Afdeling for Telekommunikation, 1992.

[57] P. Guillon, R. Zolfaghari, N. Epain, A Van Schaik, C. T. Jin, C. Hetherington, J. Thorpe, and A Tew. Creating the sydney york morphological and acoustic recordings of ears database. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, 2012.

[58] http://www.cdc.gov/niosh/topics/anthropometry.

[59] M. Rothbucher, T. Habigt, J. Habigt, T. Riedmaier, and K. Diepold. Measuring anthropometric data for hrtf personalization. In *Signal-Image Technology and Internet-Based Systems (SITIS), 2010 Sixth International Conference on*, 2010.

[60] V. Ralph Algazi, Richard O. Duda, and Patrick Satarzadeh. Physical and filter pinna models based on anthropometry. In *Audio Engineering Society Convention 122*, 2007.

[61] M. D. Burkhard and R. M. Sachs. Anthropometric manikin for acoustic research. *The Journal of the Acoustical Society of America*, 1975.

[62] A. Bertillon. *La photographie judiciaire.* 1890.

[63] A.V. Iannarelli. *Ear Identification.* Paramont Publishing Company, 1989.

[64] Vikas C. Raykar, Ramani 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*, 2005.

[65] Doris J. Kistler and Frederic L Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*, (3), 1992.

[66] Jean-Marc Jot, Scott Wardle, and Veronique Larcher. Approaches to binaural synthesis. In *Audio Engineering Society Convention 105*, Sep 1998.

[67] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1997.

[68] Durand Begault. 3-d sound for virtual reality and multimedia, 2000.

[69] V. Ralph Algazi, Richard O. Duda, and Dennis M. Thompson. Motion-tracked binaural sound. In *Audio Engineering Society Convention 116*, May 2004.

[70] V. Ralph Algazi, Carlos Avendano, and Richard O. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3), 2001.

[71] R. Gilkey and T.R. Anderson. *Binaural and Spatial Hearing in Real and Virtual Environments.* Taylor & Francis, 2014.

[72] Ewan A. Macpherson and John C. Middlebrooks. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, (5), 2002.

[73] W. Epstein and S. Rogers. *Perception of Space and Motion.* Handbook Of Perception And Cognition. Elsevier Science, 1995.

[74] R.O. Duda and W.L. Martens. Range-dependence of the hrtf for a spherical head. In *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, 1997.

[75] Douglas S. Brungart and William M. Rabinowitz. Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479, 1999.

[76] Frederic L. Wightman and Doris J. Kistler. The dominant role of low frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3):1648–1661, 1992.

[77] Cheng-Ta Chang and O.T.-C. Chen. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, 2000.

[78] R.O. Duda, C. Avendano, and V.R. Algazi. An adaptable ellipsoidal head model for the interaural time difference. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 2, pages 965–968 vol.2, Mar 1999.

[79] Simone Spagnol, Michele Geronazzo, Davide Rocchesso, and Federico Avanzini. Extraction of pinna features for customized binaural audio delivery on mobile devices. In *Proceedings of International Conference on Advances in Mobile Computing*. ACM, 2013.

[80] D. W. Batteau. The role of the pinna in human localization. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 1967.

[81] R.H. Gilkey and T.R. Anderson. *Binaural and spatial hearing in real and virtual environments*. Lawrence Erlbaum Associates, 1997.

[82] Carlos Avendano, V.R. Algazi, and R.O. Duda. A head-and-torso model for low-frequency binaural elevation effects. In *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, 1999.

[83] D.S. Brungart and B.D. Simpson. Auditory localization of nearby sources in a virtual audio display. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, 2001.

[84] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. Localization using nonindividualized head related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993.

[85] Hans Wallach. On sound localization. *The Journal of the Acoustical Society of America*, 10(4), 1939.

[86] William M. Hartmann and Andrew Wittenberg. On the externalization of sound images. *The Journal of the Acoustical Society of America*, 1996.

[87] B. Kapralos, M. R. Jenkin, and E. Milios. Virtual audio systems. *Presence: Teleoper. Virtual Environ.*, (6), 2008.

[88] Elizabeth M. Wenzel. What perception implies about implementation of interactive virtual acoustic environments. In *Audio Engineering Society Convention 101*, Nov 1996.

[89] Philip Mackensen, Markus Fruhmann, Mathias Thanner, Günther Theile, Ulrich Horbach, and Attila Karamustafaoglu. Head tracker-based auralization systems: Additional consideration of vertical head movements. In *Audio Engineering Society Convention 108*, Feb 2000.

[90] International Community on Auditory Display Georgia Institute of Technology, editor. *The importance of head movements for binaural room synthesis*, 07 2001.

[91] Gaviriel Salvendy Song Xu, Zhizhong Li. Individualization of head-related transfer function for three-dimensional virtual auditory display: A review. In *Virtual reality*. Springer Berlin Heidelberg, 2007.

[92] *Acoustics and perception of sound in everyday environments*, May 2003.

[93] William M Hartmann. Localization of sound in rooms. *The Journal of the Acoustical Society of America*, 74(5):1380–1391, 1983.

[94] Willard R. Thurlow and Philip S. Runge. Effect of induced head movements on localization of direction of sounds. *The Journal of the Acoustical Society of America*, 1967.

[95] Markus Fruhmann, Philip Mackensen, and Günther Theile. Reduction of dynamic cues in auralized binaural signals.

[96] http://www.acousticintegrity.com/acousticintegrity/Hugo_Zuccarelli.html.

[97] H. Zuccarelli. Process for forming an acoustic monitoring device, 1987. US Patent 4,680,856.

[98] http://dysonics.com/.

[99] http://dysonics.com/rondomotion/.

[100] http://dysonics.com/rondoplayer/.

[101] http://www.qsound.com/technology/overview.htm.

[102] B. Cowieson, J. Arthur, and T. Cashion. Qsound surround synthesis from stereo, 2001. US Patent 6,198,826.

[103] http://www.dolby.com/us/en/technologies/Dolby-Headphone.html.

[104] D. Markovic, A Canclini, E. Antonacci, A Sarti, and S. Tubaro. Visibility-based beam tracing for soundfield rendering. In *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, pages 40–45, Oct 2010.

[105] Olver, Lozier, Boisvert, and Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[106] Markus Guldenschuh, Alois Sontacchi, Franz Zotter, and Robert Höldrich. Hrtf modeling in due consideration variable torso reflections. *The Journal of the Acoustical Society of America*, 2008.

[107] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small room acoustics. *The Journal of the Acoustical Society of America*, 1979.

[108] Paul S. Heckbert and Pat Hanrahan. Beam tracing polygonal objects, 1984.

[109] Brad Rakerd and W. M. Hartmann. Localization of sound in rooms, ii: The effects of a single reflecting surface. *The Journal of the Acoustical Society of America*, 1985.

[110] David Griesinger. The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment. In *Audio Engineering Society Convention 126*, 2009.

[111] V.R. Algazi, R.O. Duda, and D. Thompson. Dynamic binaural sound capture and reproduction, 2008. US Patent 7,333,622.

[112] J.M. Chowning. *The Simulation of Moving Sound Sources.* Audio Engineering Society, 1971.

[113] Francis Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *J. Audio Eng. Soc*, 2002.

[114] Ebu tech. 3253 - sound quality assessment material: recordings for subjective tests.

[115] Beate Klehs and Thomas Sporer. Wave field synthesis in the real world: Part 1 - in the living room. In *Audio Engineering Society Convention 114*, Mar 2003.

[116] ITU-r BS.1116-1, methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. ITU radiocommunication assembly, 1997.

[117] Piotr Majdak, MatthewJ. Goupell, and Bernhard Laback. 3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Attention, Perception, & Psychophysics*, 72:454–469, 2010.

[118] James C. Makous and John C. Middlebrooks. Two dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America*, 87(5):2188–2200, 1990.

[119] Sylvain Choisel and Karin Zimmer. A pointing technique with visual feedback for sound source localization experiments. In *Audio Engineering Society Convention 115*, Oct 2003.

[120] RobertH. Gilkey, MichaelD. Good, MarkA. Ericson, John Brinkman, and JohnM. Stewart. A pointing technique for rapidly collecting localization responses in auditory research. *Behavior Research Methods, Instruments, & Computers*, 27(1):1–11, 1995.

[121] A. W. Mills. On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246, 1958.

[122] Richard L. McKinley, Mark Ericson, David Perrot, Robert Gilkey, Douglas Brungart, and Frederic Wightman. Minimum audible angles for synthesized localization cues presented over headphones. *The Journal of the Acoustical Society of America*, 92:2297–2297, 1992.

[123] György Wersényi. Localization in a hrtf-based minimum-audible-angle listening test for guib application. *Electronic Journal Technical Acoustics*, 1, 2007.

[124] Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye. *Probability & statistics for engineers and scientists*. Pearson Education, Upper Saddle River, 8th edition, 2007.

[125] Jasper van Dorp Schuitman. *Auditory Modelling for Assessing Room Acoustics*. PhD thesis, Delft University of Technology, 2011.

[126] L.B. Meyer. *Emotion and Meaning in Music*. Phoenix books. University of Chicago Press, 1956.

[127] J. Nowak, J. Liebetrau, and T. Sporer. On the perception of apparent source width and listener envelopment in wave field synthesis. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 82–87, July 2013.

[128] ITU-R Recommendation BS.1284-1. General methods for the subjective assessment of sound quality, 2003.

[129] T.J. Schultz. Acoustics of the concert hall. *Spectrum, IEEE*, 2(6):56–67, June 1965.

[130] J. S. Bradley, R. D. Reich, and S. G. Norcross. On the combined effects of early- and late-arriving sound on spatial impression in concert halls. *The Journal of the Acoustical Society of America*, 108(2), 2000.