Politecnico di Milano
School of Industrial and Information Engineering

Polo territoriale di Como
Master of Science in Sound and Music Engineering

Development of an interface for an HRTF-adaptation of users of interactive applications

Supervisor: Prof. Marco Tagliasacchi

Assistant Supervisor: Alejandro Gasull Ruiz

Master Graduation Thesis by: Simone Erli
Student id. 783486

Academic Year 2014-2015

Abstract

In the latest years, the field of binaural reproduction has gained remarkable importance among the different spatial sound reproduction techniques. As an example, there are some interactive video games on the market which require a particular recognition of sound features through the headphone.

One of the challenges that the binaural technique presents is the difficulty to satisfy the physical hearing cues of many users only with a single HRTF dataset. Therefore, it will be required to have a configured system so that the externalization of spatial sound will be improved for different users.

The goal of this work is to develop a web interface that combines one or more experiments, supporting a possible video gamer on training his or her hearing for the video-game platform or scenario. Moreover, there will be an analysis of the performances of two HRTF datasets, and a comparison from sound scenarios on which one is better at localizing sounds.

After a review of the current literature, an approach to the problem will be proposed, followed by an analysis of the results and conclusions.

Sommario

Negli ultimi anni, il campo della riproduzione binaurale ha acquistato notevole importanza tra le diverse tecniche di riproduzione spaziale del suono. Ad esempio, ci sono alcuni videogiochi sul mercato che richiedono un riconoscimento particolare delle caratteristiche sonore tramite cuffia.

Una delle sfide che la tecnica binaurale presenta è la difficoltà di soddisfare gli stimoli dell'ascolto fisico di molti utenti con un solo set di dati HRTF. È necessario quindi avere un sistema configurato in modo tale che l'esternalizzazione del suono spazializzato venga migliorata per diversi utenti.

Lo scopo di questo lavoro è sviluppare un'interfaccia web che combina uno o più esperimenti che supporteranno un possibile videogiocatore nell'allenamento del suo udito per la piattaforma o lo scenario del videogioco. Inoltre, ci sarà un'analisi delle prestazioni di due set di dati HRTF e un confronto tra scenari sonori su quale sia il migliore nel localizzare i suoni.

Dopo un esame della letteratura attuale, verrà proposto un approccio al problema, seguito da un'analisi dei risultati e dalle conclusioni.

Acknowledgements

Foremost, I would like to express my greatest thankfulness to my thesis advisor Prof. Marco Tagliasacchi for his continuous support and assistance regarding this work. My experience in Germany would not have been possible without his contacts with the Fraunhofer IDMT team.

Secondly, my great and heart-felt acknowledgments go to Alejandro Gasull Ruiz for his patience, willingness, and close collaboration with me throughout all the steps of my dissertation. His suggestions gave the most important mark to this work and its outcomes.

A sincere gratitude goes to all the people of the Fraunhofer Institute for Digital Media Technology for their kind and always friendly support. In particular, I want to acknowledge Christoph Sladeczek for having provided me the opportunity to work on the topic of acoustics.

I wanted also to thank all my friends in Ilmenau and in Italy. Every moment I have lived with them was really valuable, amazing, and unforgettable. Special thanks go to Luca Cuccovillo, whose experience inspired me to reach my way.

This work is particularly dedicated to Kirsten, who has always believed in me and to whom I owe every single moment, to my mother Rosanna, whose motivation was crucial for this work, to my brother Alessio, and to all the members of my entire family.

The last word and biggest devotion goes however to my father. Though he is no longer with us, his presence is always felt and somehow he bequeathed me something in order to carry me through my most formidable obstacles. I will be grateful to him all my life for this.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Every challenge in the history of sound technique development has taken its own unique direction. In particular, the researches on the integration of high-quality sound for interactive purposes have resulted in an extensive effort that must definitely not be left aside.

There are numerous key points that must be taken into account in the production of a possibly attractive application, and surely two of them draw the attention of designers. The first aspect is the production cost, and regarding this point it can be said that it is the most priority for a company. The second aspect is the product quality, a crucial part in terms of appearance and final overall rating of the output as well as persuasion of potential clients.

Another fact that has always to be taken seriously into consideration while developing an interactive application is that human sense of sight has always been dominant in orientation and perception compared with the sense of hearing. This implies that the focusing on auditory aspect is crucial for the creation of an immersive interactive application.

Sound-based interactive video games and consoles were not conceived initially to include a high-resolution sound system. This is due to many reasons, of which the most important is that a reasonably good signal processing system needs a lot more power and effort to be built up and to work, compared

to a lower-quality one. Nonetheless, some technologies have developed in the recent years so that the listener can perceive sound stimuli as they were almost real.

Having said this, binaural reproduction is a technique that plays a big role in this challenge and that gained a remarkable importance among the various existent spatial sound reproduction techniques. With the help of 3D audio rendering technique it is possible to produce an audio application with immersive sound.

1.1 What does "Binaural Reproduction" mean?

First of all, the term *binaural* means that the sound is coming to both left and right ear. The two incoming input signals are elaborated by the auditory system and the brain in such a way as to localize the origin of the sound properly and to segregate sound sources [1]. It can be said in addition that binaural technique enhances the experience of a possible user, because it makes use of different techniques in order to try to recreate the sound as if it would be without headphones. In this sense it differs from stereophony, which implies the mere splitting of sound into the two left and right channels in order to recreate directionality. Stereophony does not encompass the binaural recording of sounds though, which is taken at the ears in order to give back the sound as if it would be heard in a natural way.

Possible types of interactive application that make use of binaural reproduction technology are mainly interactive video games and movies. The fact that there could exist the possibility that a video gamer or a film watcher hears the sound from a specific direction has always been taken into account in the most finely created interactive applications. Binaural technology is also important for virtual reality systems, and everyday more for portable devices such as tablets or smartphones.

Binaural reproduction plays definitely a main role in the production of a suitable and at the same time reasonably priced audio system that can help

towards the perception of a rich quality sound. Moreover, its level of sound fidelity can compete with the quality of effects of parallel technologies used for image rendering, leading to a great contribution for the construction of a totally immersive virtual reality.

Binaural hearing provides the basis for perception of spatial sound events. More specifically, its technology reduces the sound fields in acoustic spaces into the characterization of the input signals of the human hearing system, i.e. the sound pressure at the eardrums. Its goal is to create a virtual scene which should not differ from the real scene in terms of acoustic impression [2].

1.2 Statement of the problem

In commercial video games and interactive applications the reproduction of the sound is usually channel-based, be the system stereo or multichannel. The work here reported aims to find features to include in a sound system of an interactive application to let the audio experience be enhanced to the player. This has to be implemented and rendered with an efficient but possibly cheap approach, namely an adaptive way that makes use of very few data but also consents to dynamically satisfy hearing cues and externalize spatial sound for many users. Since high-definition sound rendering systems require expensive devices, a suitable arrangement is provided to headphone users. The end user of such a binaural application has to be supported through the insertion of a short hearing training session that can improve the audio perception in its whole.

1.2.1 Hypothesis of findings

The sound directional perception for the user will be enhanced thanks to the technology used, therefore the assessment will be done through a listening test. Such cognitional ability is estimated to be increased especially for horizontal incoming directions. Considering the results, an improvement of around 10%

on average is expected. With the proposed method some elements and factors will be analyzed and thus the most performing combination of parameters will be taken into account as the most suitable. See more on this in chapter 3.

1.3 Contribution and significance of the study

The work here described aims to develop the study of binaural technology and to make it more accessible and convincing to the market by means of accessible technology. A parallel purpose may be to contribute to an enhancement of the sound perception in suitable platforms, bringing also a more immersive reality and possibly a higher sense of "satisfaction". New doors may be open to researches or further studies on the topic of binaural audio integration with interactive applications given the results of this work.

1.4 Scope and delimitation

The work will be basically centered on the development of a web application for a listening test scenario which will communicate with a sound rendering machine to carry out the binaural reproduction on headphones. Although loudspeakers can also be used along with binaural technology, the use of headphones is preferred basically because of the absence of crosstalk effects. The use of loudspeakers related to binaural impairs indeed the sound because, while having a better externalization, without any crosstalk compensation they mix up the coming-out sound. Therefore this effect is often not desired in binaural technology.

Headphones are not only cheaper but it can also be said that they can give out a comparable rendering as the loudspeakers and, according to [2], they can be correlated to real free field listening. In the experiment carried out by Oberem et al. at least 50% of the subjects could not distinguish if the sound they listened to was coming from loudspeakers or from the headphone.

The extent of the final evaluation will be limited to the recognition of

directions in the horizontal plane, or the so-called azimuthal angles. It does not thus involve the vertical directions belonging to the median plane, also referred to as elevation angles, for the reason that the latter are complicated to estimate due to psychoacoustic complexities.

1.5 Document structure

The document is organized as follows:

- In chapter 2 a summary of the developments and main features regarding the technology used in this work is described.
- In chapter 3 the deployment of the conceived solution is shown.
- The final results and analyses are reported in chapter 4.
- Eventually, chapter 5 contains a final summary of the work and suggests further developments.

Chapter 2

Fundamentals

2.1 Sound

Sound signals are defined as vibrations propagating through mediums, like air or water. Our body is sensitive to them in the sense that the pressure, caused by the propagating waves of the vibration, reaches our ear and these signals are then converted into electrical signals to be interpreted by the brain. When describing sound, it is often depicted with sinusoidal plane waves, whose main properties are:

- Pressure: it is defined as the force of a sound wave on a surface perpendicular to the direction of sound. Another definition for it can be the local pressure deviation from the ambient atmospheric pressure. It is usually measured with microphones. Its measurement unit is the Pascal ([Pa] = [N] / [m^2]). Its relative value is better expressed with respect to measurements in a logarithmic scale, relative to a reference value, namely the sound pressure level or SPL, defined by

$$SPL = 20 \log_{10} \left(\frac{\tilde{p}}{p_0} \right) \ dB$$

with $\tilde{p}$ as the root mean square pressure, and precisely [3, p. 18]

$$\tilde{p} = \sqrt{\frac{1}{T}\int_0^T p^2(t)dt}$$

and p_0 as the reference sound pressure, usually set to 20 µPa which is the approximate value of the human hearing threshold in midfrequencies.

- Intensity: it is a measurement of sound energy flow and is defined as the sound power per unit area. Its mathematical formulation relates the pressure with the particle velocity, i.e.

$$\mathbf{I} = p\mathbf{v}$$

with $\mathbf{I}$ and $\mathbf{v}$ being the vectors of sound intensity and particle velocity respectively.

- Frequency: it describes the number of wave cycles happening in one second. Namely,

$$f = \frac{1}{T}$$

with T defined as the wavelength period, which is the duration of a cycle. Frequency is the parameter that mostly contributes to the *pitch*, defined as a psychoacoustical feature to let a melodic sound be described as high or low. Concerning the human ear, the range of audible frequencies goes from 20 Hz to 20 kHz.

- Wavelength: it is the spatial distance between two consecutive peaks or valleys of a wave. Its mathematical formulation with respect to the sound is

$$\lambda = \frac{c}{f}$$

being c the sound speed, which is about 343.9 m/s in the air, and f the frequency of the wave.

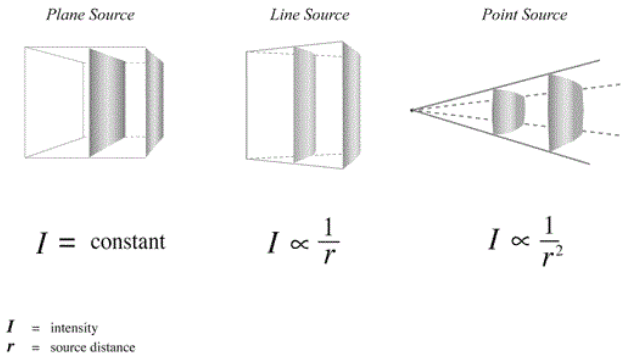


Figure 2.1: Plane, line, and point source

2.2 Sound sources

In order to characterize the sound, it is normally represented as emitted by ideal structures called sound sources. The direction of sound propagation from these is actually perpendicular to every point of their ideal surface if the propagation takes place in a homogeneous medium. Sound sources are of different types and have different properties depending on their shape.

The three main types of sound sources generally taken into consideration for sound models are, as depicted also in figure 2.1:

- **Point source**: it propagates the sound in radial directions from itself. The sound energy is the same in all directions. An ideal point source can be thought as a sphere having infinitesimal radius. According to the inverse square law, the sound intensity from a point source falls off as the inverse square of the distance. Being the energy equally distributed over the surface, which is in turn inversely proportional to the square radius of the sphere, the proportion between the SPL and the source radius r is

  $$SPL \propto -10\log(4\pi r^2) \propto -20\log(r) \approx -6.02r$$

  that is, the SPL decreases of about 6.02 dB when the distance is doubled.

- **Linear source**: as the point source, the propagation has the same radial pattern in all lateral directions. However, since an ideal linear source can be represented as a cylinder, or a line containing infinite sound sources, with infinitesimal radius r and infinite length l, the SPL is lowered by 3.01 dB for each doubling of distance from the source, because

$$SPL \propto -10\log(2\pi rl) \propto -10\log(r) \approx -3.01r$$

- **Plane source**: it is depicted as a flat surface with almost null thickness and infinite length and width. It radiates sound in the same way from its sides. Since it produces a plane wave, the sound intensity have no attenuation in an ideal case if the sound distance increases from the source, therefore its SPL is theoretically constant.

These descriptions of ideal sources are made assuming free-field conditions, i.e. in a situation where there are no other sources than the one playing and there are no reverberation effects due to obstacles encountered by the waves. In most of the cases, a sound recording or reproduction scene includes other elements that interfere with the sound waves. The main effects that influence the total sound pressure in a sound field are:

- Reflection: it is defined as the change of the direction of a sound wave according to the material of which the reflecting surface is made and the angle of incidence with it. If the surface has a finite impedance, the sound waves are subject to absorption or refraction phenomena, depending on the surface composition.

- Scattering: it is an effect similar to reflection but the wave bounces back in a random direction. This is primarily due to non-uniformities in the medium which it encounters. It reduces the energy of the sound and, in general, helps in perceiving a clearer sound. The related phenomena can also be called *diffuse reflections* whereas *specular reflections* refer to the normal reflection phenomenon.

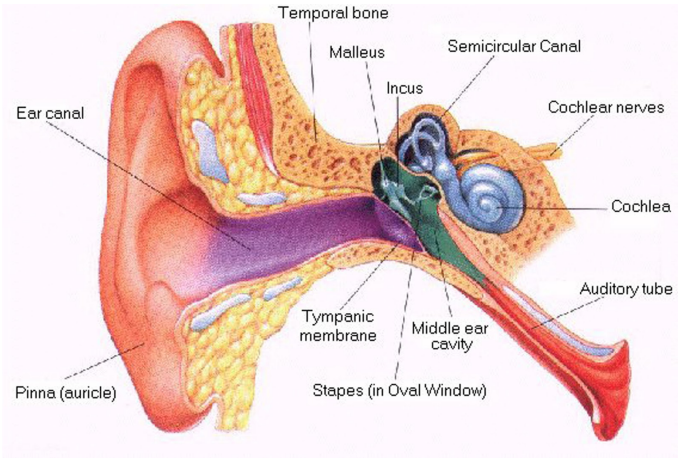


Figure 2.2: The auditory system

- Diffraction: it occurs when a sound wave encounters objects with free edges, corners or edges in a room, or boundaries between materials with two different impedances. If the object is small compared with the wavelength, the incident wave remains unaffected. If not, a shadow region results from a total cancellation of the incident wave by the diffraction wave [3].

2.3 Auditory system

Modeling the hearing system and focusing on the binaural effects to the left and right ear, leads to the understanding some phenomena such interaural or binaural cues. As in figure 2.2, the peripheral hearing system is composed by [3]:

- Outer ear: it is formed by:
  - Ear canal: it can be seen as a tube of constant width, whose walls have high acoustic impedance. It can be thus modeled as a simple one-dimensional resonator with a 3 kHz resonance frequency. Our sensitivity has its maximum in this frequency range. The ear canal

is delimited by the eardrum, which is a thin membrane that converts the pressure arriving at its surface into mechanical signals recepted by the middle ear.

  - Pinna: it is the visible part of the ear and its role is to amplify and modify the sound especially in the high frequency range. Thanks to its anatomy, it acts as a filter, bringing spectral notches in some frequency regions.

- Middle ear: it consists of the bones malleus, incus, and stapes, which transmit the sound mechanically. The impedance match provided by the middle ear improves the efficiency of sound transmission significantly. It terminates with the oval window, which connects the stapes and the inner ear.

- Inner ear: it consists of the cochlea and the semicircular canals which contain the human organ of equilibrium. The cochlea is a spiral-shaped organ with two parts divided by several membranes. The basilar membrane plays a central role in the transduction of vibroacoustic stimuli into neural electric signals, given by the cochlear fluid. The sound wave in the fluid excites the basilar membrane to transversal waves. The vibration amplitude on the basilar membrane varies with frequency along its length, and the displacements on the basilar membrane are detected by sensory cells which transmit electrical pulses to the local nerves.

Also the head influences the sound filtered and processed by our ear. Given that a sound source around the listener is currently active, the head is an obstacle to the incoming sound. This gives rise to two important phenomena that affect the sound perception: head diffraction and head shadowing. Both effects originate two interaural cues respectively, which help in localizing the sound source:

- *ITD* or *Interaural Time Difference* (figure 2.3): supposing we have a non-centered sound source (i.e. away from the median plane), it is defined as the time distance that occurs between the instant in which the

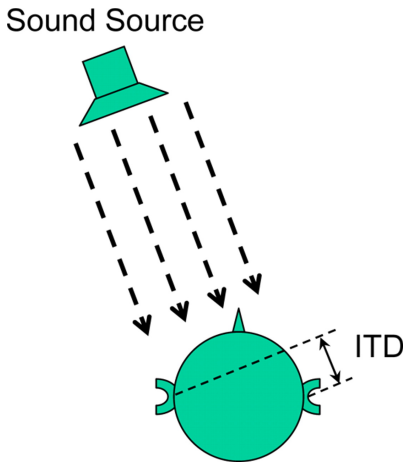


Figure 2.3: Interaural time difference

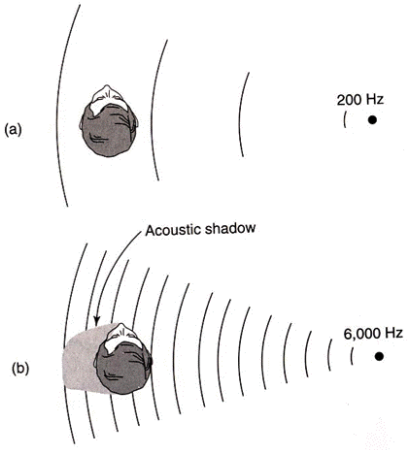


Figure 2.4: Interaural level difference at low and high frequencies

sound reaches the ear located on the same side of the source and the instant in which it reaches the opposite ear. It is extremely important for localization, moreover it is a key point belonging to the Duplex Theory (Rayleigh, 1907) according to which the ITD causes a phase difference between the ears and plays a fundamental role at low-frequency recognitions, approximately below 1 kHz, of sources not located on the median plane. As reported by Wightman and Kistler [4], experiments can show that it is the dominant cue for wide-band signal recognitions.

- *ILD* or *Interaural Level Difference* (figure 2.4): it is defined as the difference of the intensity level between the perceived sounds coming at both ears. Classified as a secondary cue in comparison to the ITD, it consents us to localize a source in a more refined way, thanks to the fact that this cue is present mostly at high frequencies. Since the head shadows the sound waves coming to the opposite side, the ear located on the same side of the source can use this cue to resolve confusions along with spectral cues mostly by means of the pinna [4]. Also Olsen and Carhart [5], who conducted a study about the benefit brought by this cue to the hearing impaired, stated that the ratio between the signal level at the eardrum and that at the entrance of the ear canal is particularly greater at a frequency interval between 2 and 5 kHz.

2.4 Spatial audio

2.4.1 Brief history of stereo

According to Sunier [6] , the history of stereophony draws back to 1881 when Clément Ader implemented a telephone system at the Paris Opera to let theater stage sound be heard. He placed two groups of transmitters on the left and on the right of the stage itself, and two receivers, likewise left and right, arriving to a telephone apparatus.

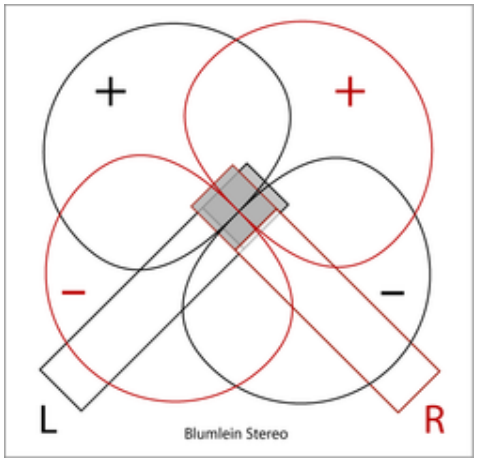


Figure 2.5: The Blumlein pair

Alan Blumlein created in 1930 a two-channel stereophonic recording system with two microphones placed as close together as possible but angled out toward left and right sides. This configuration, shown in figure 2.5, is nowadays known as *Blumlein Pair* and it was drawn so as to eliminate time and phase differences and enhance spectral features like amplitude and energy.

A stereophonic telephone transmission of an orchestra live performance was carried out in 1933 by Bell Laboratories. Three microphones were places in front of the orchestra and linked to a respective loudspeaker in a remote room. To recreate directionality, each loudspeaker must have been in the same place of the corresponding microphone, whereas the two halls had to be of the same size and shape.

After a period of lack of interests in stereophonic techniques after World War II, the revival took place with the commercialization of standards for home stereo tape reproductions. This let stereophony be accessed by a huge public and also record companies began to produce stereo tapes.

Based on what is reported in [7], Dolby Stereo was launched in the mid-1970s. The peculiarity in its inner technology was that it used optical sound prints and there were two sound tracks together in the same space occupied by the traditional mono track. Then matrixing techniques were employed to

incorporate also the center and the surround channel. Dolby Stereo was then introduced as a new standard and the movie industry experienced a kind of renaissance thanks to the new technology.

Dolby Surround introduced in 1982 was meant for home playing and included thus in its devices the possibility to play out the surround channel, whereas the related Pro Logic technique further released had also the possibility for the center channel to be decoded.

As the digitalization process in the audio technology industry went along, Dolby Laboratories decided to implement a digital system for the cinema screenings. To wit, a separate new digital optical track was included to satisfy the new *5.1* configuration described below. Along with Dolby Digital, another format was introduced and launched in those years by the company Digital Theater Systems Inc. (DTS), whose technology was based on using lossy compression at a higher bit rate compared to that of Dolby Digital. In the end DTS systems can fully compete with Dolby surround-sound technologies and to some users it can even seem that DTS performs a slightly little clearer and more natural.

2.4.2 Two-channel reproduction

The definition of a two-channel sound reproduction system, or in common expressions, a stereo system, is an arrangement that makes use of two loudspeakers. Stereo systems can be *true*, which means that a signal from a live sound is recorded by a pair of microphones and is played back through loudspeakers. This lets the sound be played out with the possibly closest fidelity.

Another classification would be that of *artificial* stereo systems, comprising a set of recorded mono sounds to be controlled in amplitude at the two playback channels by means of pan-pot mixing. Thus an artificial direction for each of the sounds can be displayed and a full simulated sound field can be created.

The ideal placement of the two loudspeakers from the listener position would be made so that almost an equilateral triangle can be drawn between the loudspeakers and the listener, as in figure 2.6. Here with “listener” is meant

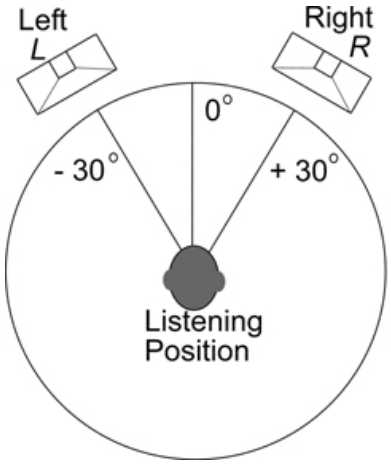


Figure 2.6: Position of the loudspeakers in a two-channel reproduction system

the playback area intended to be covered by the sound field, also known as *sweet spot*. This is done in order to maintain certain features like directional cues and evenness of some frequency ranges like basses.

At the recording stage, two microphones are used and placed in strategically chosen locations relatively to the sound source. The two recorded sounds will have similar properties but tend to differ in phase (time of arrival) and SPL. The listener will then use those differences to locate the recorded sounds.

Stereo systems, compared to mono, give a spatial impression to the listener thanks to level and phase information. This difference is revealed in the *phantom image*, an apparent sound source which is "brought to life" and made hovering in the space between the two loudspeakers. While the monaural version comes from an equal signal sent to the speakers, thus it is sounding at the center of the scene, the stereo version is created from two signals which are slightly different. The latter is therefore more realistic and detectable than its mono version[1].

[1] http://www.moultonlabs.com/more/principles_of_multitrack_mixing_the_phantom_image

Two-channel playback systems are suitable for a small sweet spot. If these systems are used in large areas such as theaters or big rooms, there will be a lack from uniform coverage of the entire listening area. It will be necessary for that to have a loudspeaker system for each channel but the cost of this implementation is very high.

Two channel systems are then a poor choice for music reinforcement for example, since many listeners in the area will hear a completely different sound mix compared with other listeners[2]. Moreover, a two-channel system has no feature regarding front to back movement, so it is far away to be considered a surround system.

2.4.3 5.1 surround sound system

The *5.1 Surround Sound* system was introduced by Dolby Laboratories in 1982. The configuration is called in this way because it makes use of five speakers, which are left, center, right, left surround, and right surround respectively, and an added channel standing for low-frequency effects (LFE), whose speaker is commonly known as *subwoofer*.

The center channel carries mainly the sources that are wanted to be located at the center of the sound scene, like human voice in movies or vocals in music. It also helps to centrally stabilize the phantom image, in order to anchor the sound field. The surround channels, instead, carry the low-level ambient and diffuse sounds.

The placement of the speakers is described in figure 2.7. According to the ITU standard ITU-R BS.775-3[3], the center channel is placed at 0° with respect to the front, the left and right channels are located at ±30°, and the surround channels can be set in a position between ±100° and ±120°. These

[2] http://www.mcsquared.com/mono-stereo.htm

[3] https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.775-3-201208-I!!PDF-E.pdf

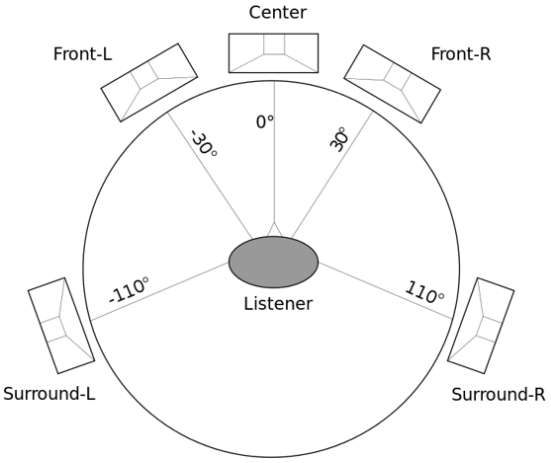


Figure 2.7: Position of the loudspeakers in a 5.1 reproduction system

ones are placed above the ears of the listeners and possibly along walls, so as to recreate an extremely diffuse sound. Finally, the subwoofer can be basically placed anywhere. If it is located near walls, the bass section will be enhanced, even though a less controlled bass quality can result from this.

Between stereo and 5.1 systems, the latter configuration is more suitable for reproducing audio in movies, because of the greater diffuse sound and sense of presence offered by it. Instead, the stereo configuration fits better in case of audio reproduction of music and TV programs. Certainly, there is often no need in this case to have an enhanced surround effect, since the audio scene is not centered or surrounded. Apart of that, the surround systems present no feature for sound reproduction in terms of heights, hence there is no deep sensation of a complete acoustic space.

Dolby Digital 5.1 was followed by the 6.1 and later by the Dolby Surround 7.1 configuration for cinemas in 2010, which add to the former playback pattern one and two back surround channels respectively. Later on, Dolby Atmos technology was announced. It allows up to 128 audio tracks to be properly set in theaters for dynamic rendering based on the local capabilities. As a consequence, it can be configured for additional loudspeakers like new front, surround, or ceiling-mounted height channels to achieve precise panning of

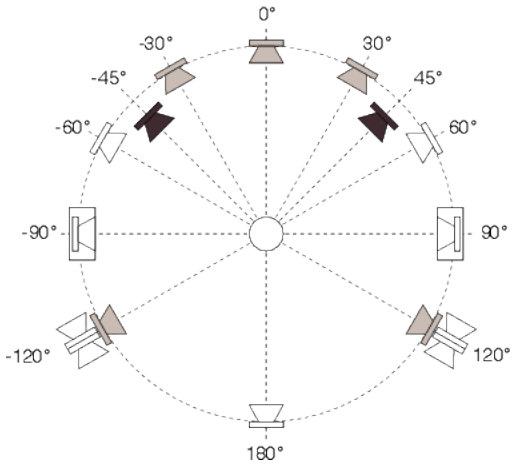


Figure 2.8: Position of the loudspeakers in a 10.2 reproduction system

certain sounds [8].

2.4.4 10.2 and further

The 5.1 standard was even not sufficient for some people. Based on it, a system dubbed *10.2* and developed by Tolminson Holman from THX was introduced. As seen in figure 2.8, it has five loudspeakers matching the 5.1 requirement and positioned at respectively 0°, ±30° and ±120° with respect to the front (those at ±120° can actually be equipped with two orthogonal loudspeakers at their rear) plus other two speakers at ±60°, one at 180°, two at ±45°—those being also 45° higher, in order to give an impression of height or an emphasized envelopment—and two subwoofers placed at ±90°, these with the same signal arriving at them but with inverted polarity, so as to possibly cancel the direct sound and give the listener a more diffuse sound [9].

The 10.2 system can also be improved with an addition of two point surround channels acting as diffuse radiators. With these, the sound can be further reflected off walls before arriving at the listener. This led to the 12.2 configuration. Another format, developed by the Japanese NHK, is the *22.2* system. This can allow up to 22 speakers (and 2 subwoofers) to be distributed

over three layers in height, and is compatible with all the previous audio formats. The higher the number of channels of surround systems, the greater the possibility to involve larger spaces or a bigger number of listeners, in order to recreate a very fine sense of audio spatiality.

2.4.5 Wave field synthesis

The principle of wave field synthesis, whose first studies were carried out in 1988 at Delft University in Netherlands, says that, having a source that emits a spherical wave, it is possible to reconstruct the wave field in a volume as long as the wave front on a continuous surrounding surface is known, and the respective wave front can be reconstructed by means of sampling methods. This conclusion is possible because of the existence of the Huygens principle which states that every point of a wave front can be taken as the source of an elementary wave that interacts with the surrounding environment. The Kirchhoff-Helmholtz integral, which is its mathematical formulation, is the core of wave field synthesis. Placing an equally arranged microphone array in the original sound field and a reproducing the recorded signals with an equally arranged speaker array—both arrays having n components—it is possible to obtain the synthesized wave front in the area covered by the speakers.

On the one hand, the binaural signals synthesized with this technique arise in a more natural way compared to a binaural reconstruction of the field itself. On the other hand, wave field synthesis gives rise to a certain number of practical constraints [10]:

- **Discreteness of the array (spatial aliasing)**: due to the discretization of the secondary source distribution, spatial aliasing, given by the erroneous reproduction of a wave field above the aliasing frequency f_{alias}—which depends on the spacing of the loudspeakers and on the geometry between the source and the listener—produces spatial and spectral errors and artifacts like coloration and localization.

- **Room reflections**: possible reflected waves interfere in the spatial per-

ception and therefore a wave field synthesis array cannot be rendered properly. Room compensation algorithms help in the minimization of this effect.

- **Restriction to the horizontal plane**: wave field synthesis is practically restricted to the horizontal plane despite the fact that theoretically it is not. The first consequence of this is the impairment of perception of distance, depth, spatial impression and envelopment, while the second aspect is the rise of cylindrical waves, which bring errors affecting the level roll-off.

- **Limitation of array dimensions**: the finite length of the loudspeaker array gives rise to after- and pre-echoes and contributes to additional coloration.

- **Effects on perception**: some effects are still not known in detail, although psychoacoustics can come in handy in order to understand them.

2.4.6 Ambisonics

Ambisonics is a reproduction technique invented by Michael Gerzon in the early 1970s at the Mathematical Institute of Oxford, UK. Its deployment involves a number of playback channels—limited only by the number of transmit channels defined by the ambisonic order—while it allows a reproduction of full 3D acoustic spaces with several moving virtual sources.

It is based on the holographic theory, which states that any sound field can be expressed as a superposition of plane waves. Thanks to the Kirchhoff-Helmholtz integral, it is possible to reproduce the original sound field by an infinite number of loudspeakers arranged on a closed contour. The ideal reproduction would be in a spherical volume surrounded by a finite number of loudspeaker, and the respective area is then called *sweet spot* [11].

The main advantages of Ambisonics are the following:

- Decoding and encoding phase in Ambisonics are decoupled so that the loudspeaker placement is completely flexible.
- The recording method can be set up in a simple way with just the need of multiple microphones at the center of the scene.
- All speakers contribute to the sounds involving any direction, thus giving greater localization also to the sides and the rear [12].
- Ambisonic approach gives in the overall great benefit in computational efficiency for binaural sound reproduction.

The main disadvantages are instead the following:

- Sources can only produce planar wave fronts, thus the rendered wave field must be a combination of planar wave fronts; this problem affects small environments like cars, where the speakers must be sufficiently far from the listener.
- Is it not widely known and its concept is difficult for people to grasp.
- Due to the fact that the sources are played out by several speakers with strong correlation, Ambisonics is prone to phasing artifacts when the listener moves or turns.

2.5 Binaural technology

2.5.1 Definition

Binaural hearing is a discipline of psychoacoustics and leads to directional localization and distance perception. Furthermore, it helps in understanding speech or nearby sounds in noisy environments thanks to its consistent noise suppression, described as the "cocktail party effect" [3]. According to its theory, the sound must be reproduced in a way as separate and accurate as possible.

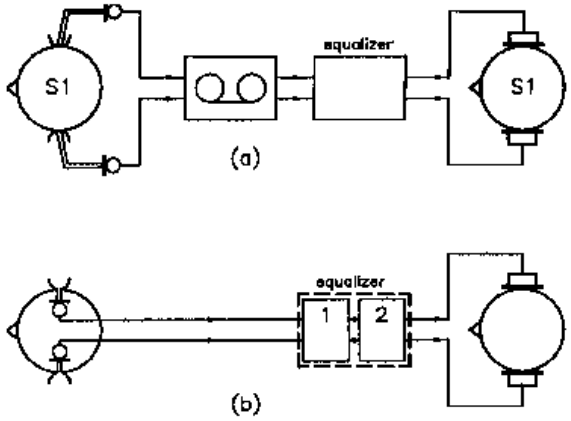


Figure 2.9: Examples of binaural recording and reproduction with a real listener (a) or with a dummy head (b). In both cases, equalization between the two stages is needed. See also section 2.5.3

Therefore, to give out an auditory impression that can be very close to the hearing of real 3D sound, headphones are the most suitable equipment to combine with binaural, though also loudspeakers are capable to yield affordable performances when fed with binaural signals. While the sound is influenced by direction-dependent linear distortions, mainly reflections and diffractions caused by head and torso, the dependence comes also from the distance in case of spherical waves. The flow of sound under these conditions is resumed in a scattering pattern called *Head-Related Transfer Function* (HRTF) and described in section 2.5.3.

Binaural recording can take place with either real listeners (see figure 2.9a) or properly made dummy heads (see figure 2.9b) as subjects. The reproduction of binaural audio with headphones requires also a well-defined 3D audio scene to be played back with the aid of sound rendering techniques. These must include spatialization of the different sound sources and the construction of a full response depending on angular directions.

2.5.2 Brief history of binaural

The contents of this section are based on the work of Paul [13]. It can be said that Ader's experiment in 1881 was the first try ever to comply with the definition of binaural technology (see section 2.4.1, line 1). The very first binaural implementation was presented in 1932 when Bell Laboratories introduced *Oscar*, a dummy having a pair of sensitive microphones in front of his ears (see figure 2.10a). After World War II, stereophonic recording and reproduction equipment were improved thanks to the introduction of artificial heads.

From the 1960s on, more sophisticated human head models begin to burst on the binaural scene. An article written by Nordlund and Lidén in 1963 [14] reports the first experiments with a dummy head including the simulation of the ear canal, inasmuch as the microphones were placed approximately at the eardrum. Later on, in 1967, Bauer et al. and Kürer et al. created manikins with approximations of both pinna and ear canal [15, 16].

There was still a challenge to overcome, namely the introduction of a standard dummy head representing an average adult. *KEMAR* (Knowles Electronic Manikin for Acoustic Research) was presented in 1972. Its ear simulator could reproduce the behavior of eardrums, ear canals, and pinnae of different sizes (see figure 2.10c). These features led KEMAR manikin to become a reference for *in-situ* measurements.

In the 1980s, a significant contribution was given by Neumann GmbH with their KU81 artificial head, which had smaller microphones in order to reduce the roll-off of the response at high frequencies. Its pinnae were also derived from mean parameters of different subjects. Another noteworthy work of those times was that of Genuit, who developed a system for technical sound reproduction that featured free-field equalization for headphone reproduction [17]. The later *HMS II*, whose version number 5 is represented in figure 2.10b, featured head and shoulder simulation and simplified elements to the purpose of obtaining a representative manikin.

From the beginning of the 1990s audio manufacturers implemented new

(a) Oscar (b) HMS II.5 (c) KEMAR

Figure 2.10: Pictures of some well-known binaural dummies

technologies in their products such as digital signal processing. A German producer, Cortex Instruments, offered dummies like their *MK 1* model, which featured a human-like head and torso geometry and pinnae adapted to standards. New solutions for dummy heads were sought and analyzed, like the use of the best localization performance criterion instead of means of values taken from individuals.

Later on, a great discussion on whether binaural recording equipment must be standardized or individualized arose then as further developments went on and this debate is actually valid up to now. In particular, standardization is favored industrially in order to compare data coming from different dummy heads in a better way, whereas individualization is kept by some researchers as a possible solution for localization problems.

2.5.3 HRTF

Certainly many techniques have been deeply studied through various researches done in the 3D audio field. One of the most important and challenging one is how to model the sound that comes to our ears. This has been taken into account since the first studies and the respective model is called HRTF (Head-Related Transfer Function), namely the function that models the transformations that a sound coming from a source—positioned at a certain point

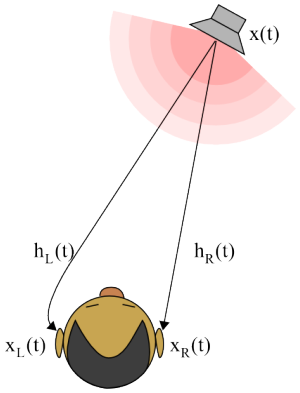


Figure 2.11: HRTF principle. The impulse responses h_L and h_R are also depending on azimuth and elevation

in the space surrounding the listener—undergoes until it reaches the eardrum. The synthesis of an HRTF is usually made so that its model will be eventually cost-effective, that is to say, in a reasonable level of complexity while keeping suitable rendering performances.

The first usage of this function in literature was made by Jens Blauert under the term of Free-Field Transfer Function (FFTF) and dates back to 1974. According to Vorländer, the details of the HRTF were discovered by Shaw [18] and it is defined by the sound pressure measured at the eardrum or at the ear canal entrance divided by the sound pressure measured with a microphone at the center of the head but with the head absent [3].

Sound filtered by means of HRTFs preserves spatial hearing cues, in such a way that the auditory system links spatial attributes to auditory events, like in natural free-field listening [19, 20]. The mathematical principle of the HRTF goes as follows. Let $x(t)$ be the signal coming from a source in the horizontal plane at azimuth θ and elevation ϕ, where with azimuth and elevation it is intended the angles on the horizontal plane and on the median plane respectively. The left and right headphone signals, $y_L(t)$ and $y_R(t)$, undergo

an HRTF filtering and can be expressed as:

$$y_L(t) = h_L(\theta, \phi, t) * x(t), \qquad y_R(t) = h_R(\theta, \phi, t) * x(t),$$

respectively, where $*$ represents the convolution operation, and $h_L(\theta, \phi, t)$ and $h_R(\theta, \phi, t)$ are the left and right *head-related impulse responses* (HRIRs) corresponding to the direction (θ, ϕ) (see also figure 2.11). In the frequency domain, this can be seen as a signal multiplication, hence

$$Y_L(f) = HRTF_L(\theta, \phi, f)X(f), \qquad Y_R(f) = HRTF_R(\theta, \phi, f)X(f),$$

where $HRTF_L$ and $HRTF_R$ account for the left and right components of the HRTF, respectively.

The measurement of an HRTF takes place with in-ear microphones and speakers surrounding the subject, the latter ones being placed in a circle every angular step. Measurement signals are produced from the speakers to measure the impulse response of each source direction to each ear, therefore the whole process is quite tedious and elaborating. Indeed, in order to achieve a good reproduction of binaural sounds, recordings must be made for every angular step and thus the process can last long.

The HRTF can be individualized or non-individualized. In the first case, the recordings are taken singularly for human listeners. In the latter case, they take place using a dummy head. It must be pointed out that, when the subject of the recording measurements is a human, he is not allowed to move his or her head during the process, otherwise the results will be marred by errors. Supports for head blocking come in handy in this case, but this requires a lot of patience to the listener. When using a dummy head, the process is of course simpler, but then a careful choice about the type of dummy head must be made, in such a way that it would represent a good amount of the subjects.

Many elements actually influence the HRTF, from the body to the surrounding environment: the most influent are the body parts, that is, ears, head, shoulders and torso, in order of importance respectively. A scheme that

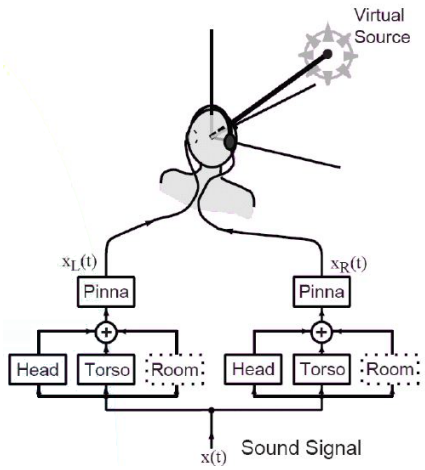


Figure 2.12: Schema including the main elements influencing the HRTF

can be suitable for an HRTF representation is shown in figure 2.12, where also room effects can be taken into account if they are considered in the scenario.

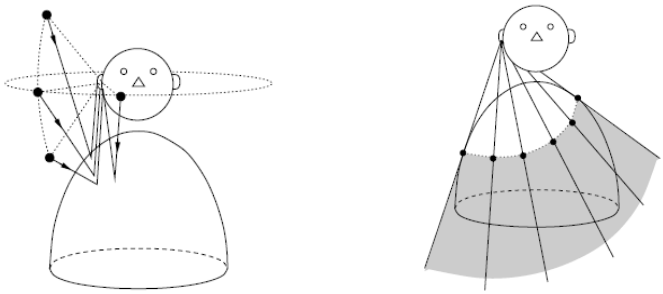
There is also another key point that affect the composition of the system, and that is the subjective component. This part is actually very complex to study due to its subjective nature, since every listener has a different background in social and physical terms. This can also be seen for example in the structure of the pinnae where this diversity comes out to be even greater among humans. Nonetheless, researches in psychoacoustics have always come along to help giving out interesting results about the study of the relation between subjective and objective factors.

Head A good head approximation to consider for the HRTF is a sphere having ears located at the same height at opposite sides. The main sound effect carried along with the head is diffraction [21]. While both ITD and ILD (described in section 2.3) let binaural audio through HRTFs sound convincing for an improvement in localizing sources, some still very known problems that lead to errors can result from the listening. These problems are related to the localization of a source and can arise as a consequence of listening to a

binaural recording with an HRTF pattern that does not match the features of one's own ear system. The most relevant are known as:

- **Front-back or elevation discrimination**: this problem occurs especially with the use of a non-individualized HRTF. The areas in which this effect happens are called *cones of confusion* and are the front-back part on the horizontal plane (whose effect in particular is called *front-back confusions*) and the up-down part of the elevation localization area (whose effect in this case is called *up-down confusions*). Since ITD and ILD are small in the median plane, it becomes harder to have a clear localization of possible sound sources as they get closer to it [3], although monaural cues and head-tracking techniques can improve this discrimination.
- **In-head localization**: it results from the confusion errors and a weak headphone coupling, the latter depending from the headphone type and its position in relation to the ear. The sound images are localized within the head instead of in the correct location, often tending to come from the rear [22].
- **Localization blur**: it is defined as the minimum audible angle, in other words, the angle of uncertainty while trying to localize one source of sound. Its values vary between 5° and 20° [23].

External ear The external ear models the sound at the end of the HRTF chain, modeling higher frequencies in particular and introducing some resonances. The most influent part that is contributing to the HRTF is the pinna, which gives the greatest contribution to high frequencies and elevation angles. The analyses of features and the consequent model of the corresponding Pinna Related Transfer Function (PRTF) carried out by Satarzadeh et al. [24] report that such a function has a few parameters and can be modeled by two low-order bandpass filters, representing two resonances, and one comb filter, which accounts for one main reflection. The first resonance is uniquely defined by



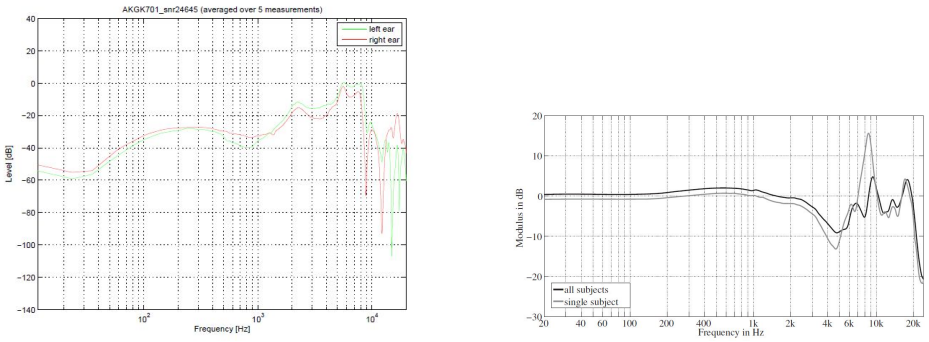
(a) Reflections (b) Shadowing

Figure 2.13: Main effects caused by torso and shoulders

width and depth of a cylindrical approximation of the pinna itself while the second resonance is thought to be correlated with the main reflection's time delay.

The greatest spectral contributions brought to the HRTF by the pinna are shown in terms of spectral peaks and notches: the first are supposed to be in correlation with resonances while the latter are found out by Spagnol et al. [25] to be probably related to three main tracks corresponding to three reflective pinna anatomical surfaces: helix wall, antihelix/concha wall and concha border, found out by the authors with a contour-matching algorithm.

Torso and shoulders Since the role of the torso and shoulder components is not relevant compared to the other parts mentioned before, there is fewer research that has been done on it. Their main effects are in the low-frequency region [21], namely additional reflections to the ear (figure 2.13a) and shadowing of the sounds coming from below the torso, or the so-called "shadow cone" (figure 2.13b). The easiest representation is an ellipsoidal torso located below a spherical head. Reflections are spectrally represented as a comb filter, with notch locations inversely related to the reflection delays. This explains the corresponding behavior with respect to the elevation angles, that is to say, the reflection delay is greater with varying elevations and has its maximum when the source is located above the listener.



(a) (b)

Figure 2.14: (a) Example of headphone transfer function (AKG K701) and (b) equalization filter suggested by Masiero and Fels [26]

2.5.4 Challenges of binaural technology

Influence of headphones The headphone transfer function (HpTF) maps the sound between headphones and eardrums. Recent studies have given out that a proper equalization or compensation of headphones leads to suitable perception of binaural sound and avoids unwanted effects such as spectral coloration. In particular Masiero and Fels suggest a filter (figure 2.14b)that gives valleys instead of peaks—the latter ones more disturbing—in the high-frequency region of the respective HRTF spectrum, which would be otherwise dependent on the headphone fitting and thus on the perception of every individual [26–28].

Furthermore, headphone reproduction is also likely to bring effects resulting from head movements. If the head is in some way turned or shifted, the surrounding virtual reality moves with the listener. This introduces errors leading to misperception of source positions, which can be heard, among other things, inside the head. To solve this problem, devices able to track the position of the head should be integrated with the system. Unfortunately, these systems are not often available at competitive prices, due to the high complexity of their technology.

Auralization Auralization is a technique of re-creating the sound of a certain environment with the aid of computer data. This environment can be of many types: from free fields, in which sound sources are playing without the simulation of any sound obstacle like walls, to diffuse fields, in which the effects given by walls and enclosures are taken into account in such a way as to re-create a simulated sound scene in this kind of environment, though not physically being in it.

It is common to combine originally free-field sounds with the responses of diffuse-fields environment, in order to re-create a diffuse-field sound scene. The HRTF is then the main tool used to re-create such effects in close environments such as rooms, buildings, vehicles or other technical devices. The term "auralization" can be compared to that of "visualization" because in visual illustration of scenes in movie animation and in computer graphics, the process of "making visible" is described as visualization. Analogous effects occur in auralization when primary sound signals are processed into an audible result [3].

Through such method, there is no need of physically building an enclosing environment nor measuring its acoustical properties. It only requires information about the dimensions of walls and obstacles, their materials, and the features of sources and receivers (position, orientation, and directionality). This information is contained in the measured impulse response.

The main drawback of auralization is the time required to simulate the room acoustical behavior and to process the sound through the impulse responses, due to the number of convolutions needed. Torres et al. suggested a simplified model for this technique consisting of a reduced-order representative HRTF for each direction [29].

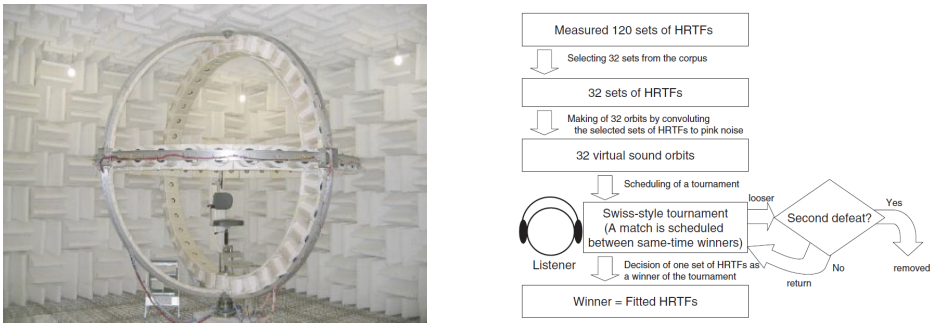
Individualization It is usually referred with the term "individualization" as the process in which an HRTF is measured for a single listener or a group of listeners. Therefore, an individualized HRTF is a function measured with this criterion whereas a non-individualized HRTF is a function taken from

other data rather than those specific of the subjects. Individualization brings a benefit generally in terms of sound rendering and localization, giving thus a better and clarified image of an auditory scene. With an individualization method, the resulting HRTF is the most suitable for that particular subject and if the same transfer function is applied to other subjects, localization errors and confusions may arise.

Listeners adapt to non-individualized HRTFs easily, but better after training and feedback, despite a different auditory experience. In particular, according to the results of Mendonça et al. [30], humans can learn to localize sources with altered spectral stimulation, based on their experience. Wenzel et al. state also that there are a considerable quantity of similarities between the audio emitted by free-field sources and virtual sounds filtered by non-individualized HRTFs [23].

Despite the fact that individualization gives improvements to the ability of localize sound sources during binaural reproduction, it is still a problem to implement a system to individualize HRTFs in a commercial environment. The main factors are both the costs of time and money. To measure a suitable individualized HRTFs, a reasonable method, a fine quantization step—usually around 2° on the median plane—and quality ear microphones and loudspeakers are needed. Moreover, the listener has to sit in the same position for several minutes due to the time length of the process, and this becomes the greater obstacle, because it suffices just a movement of the head and the results are compromised. Methods to block the head can be incorporated, but this requires a lot of patience to both the subjects and the experimenters anyway.

A lot of works were done so far in the field of individualization. The first works were primarily focused on creating the individual HRTFs, namely, a sample of persons were taken voluntarily and subjected to individual measurements of Head-Related Impulse Responses (HRIRs), that is to say, the corresponding function in time of the HRTF according to the Fourier Transform. These HRIRs were taken with a set of loudspeakers surrounding the listener in a circle or in a dome, then processed with signal processing tools (like Matlab)



(a) Loudspeakers (b) Procedure

Figure 2.15: Material and algorithm used in the DOMISO test by Iwaya [32]

subsequently[4]. Other works were instead concentrated on building up HRTF datasets from single measurements, specifically dummy heads, see for example the MIT KEMAR database[5] in which a dataset was constructed from a KEMAR dummy head. Similar works were also done at the Fraunhofer IDMT institute in Ilmenau, Germany, especially with Cortex and KEMAR dummy heads.

Further investigations were based on individualization of HRTFs upon already existent databases. Seeber and Fastl developed a work in which the selection of the individual HRTF had to follow some evaluation criteria, like externalization, minimization of front-back confusions, match of the presented and perceived directions, focused virtual auditory image. They made a subsequent experiment in which a two-step procedure was carried out such that the final HRTF, singled out of five sets selected in the first step according to some other criteria, would be selected as the best-matching one [31].

The method introduced by Iwaya, called DOMISO (Determination method

[4]See for example the CIPIC database (http://interface.cipic.ucdavis.edu/sound/hrtf.html) or the IRCAM database (http://recherche.ircam.fr/equipes/salles/listen/), just to mention a few.

[5]http://sound.media.mit.edu/resources/KEMAR.html

of OptimuM Impulse-response by Sound Orientation), calls for the following steps [32] (see also figure 2.15b):

- The system selects 32 HRTFs randomly from the whole set of already measured HRTFs with the loudspeaker set in figure 2.15a.
- An orbit of 13 virtual sound images, created with a one-second long pink-noise sound, for every 30 degrees counterclockwise from the front of the listener in the horizontal plane is prepared.
- The orbit is then shown to the listener as a preamble to the listening test.
- A tournament-like style of presentation of the HRTFs to the listener is set up.
- The listener selects the HRTF that resembles the orbit at the best from every match.
- The winning HRTF is eventually select as the fitted set.

The technique is said by the author to be:

- easy, due to less physical restrictions to the listener.
- fast, because the task requires just 15 minutes to be completed against 2 hours of a complete HRTF measurement.
- computationally simple, because no signal compensation for the headphone is required, since DOMISO uses the VAD (Virtual Auditory Display) system to be individualized.

The author also claims after some variance analysis of the front-back confusions that the HRTF produced in this way achieves the best performances regarding localization on the horizontal plane with respect to the listener's own HRTF.

Wenzel et al. used a HRTF from a representative subject of one experiment made previously by some of the authors. Their trial consisted in the judgment of a free-field location of noise stimuli processed with this HRTF, either through speakers or headphones. The aim of the experiment was to compare the localization between the free-field case and the headphone reproduction—namely, a virtual free field—and the subjects were not told anything about their performances as a feedback. The data were presented in an elevation range between -36° and +54° with intervals of 18° and at random azimuth values with a step of 15° (24 azimuth positions in total). Wenzel et al. were the first to note the presence of up-down confusions and regarded the calculation of confusion rates as a separate statistic. Their judgment of the average location was based on a centroid vector computation. The results showed in general that interaural difference cues were maintained with headphones while, although the spectral details responsible for elevation and front-back discrimination were distorted by the synthesis with non-individualized HRTFs, 14 out of 16 subjects showed good accuracies in both free-field and virtual-source conditions [23].

Mendonça et al. did instead a work concerning a possible adaptation to non-individualized HRTFs. Since they state that there is the possibility for humans to improve their localization with altered spectral simulation based on single experiences and thus perceptual learning processes—brain plasticity is actually the governing process in this case—they made up a set of trials to assess the quality of localization improvements and the long-term effects of their method. They did two experiments, one for the judgment of azimuthal angles, in which the white noise stimuli were played in the front-right quadrant of the horizontal plane (from the front to the right), and one for the elevation angles, in which the same signals were given in the front-upper quadrant of the median plane (from the front to the top of the head). The interval between one location and the next one was of 10° and the HRTF database with which the signals were convolved was the CIPIC database. The steps of the approach were the following [30]:

- **Pre-test**: the sounds were presented randomly and repeated 10 times. The subject had to click the estimation point on a proper touch screen. Each signal had the duration of 3 seconds with an interval of 2 seconds between the stimuli.

- **Training**: in this part, the subjects trained on points located in the aforementioned quadrants with an angular interval of 30° between each other. The two sub-experiments that the participants had to follow were:

  - **Active learning**: the subjects could listen to any of the presented stimulus to their liking. This part lasted five minutes.

  - **Passive feedback**: the training sounds were displayed to the subjects and after a given point on the touch screen, the correct answer for each try was given. This lasted until participants could answer correctly in 80 percent of the tries in the azimuth experiment, or 70 percent of the tries in the elevation experiment. Each sound had here a duration of 3 seconds, with an inter-stimulus pause of 4 seconds.

- **Post-test**: this experiment was exactly the same as the pre-test and done in order to draw out and compare the results of the training sessions. Moreover, the same test was then done five times, so as to judge the progresses in the long term, specifically after: the first training, one hour, one day, one week and one month.

All the four subjects had an improvement from this test. Two of them were already experienced and probably thanks to their previous background, their improvement were higher that the other two, who were inexperts. After one month, the mean decrease in localization error was of 3.48° on the horizontal plane and of 9.88° on the median plane compared to the initial training day.

Chapter 3

Proposed approach

3.1 Research design

The test conducted at Fraunhofer Institute for Digital Media Technology in Ilmenau, Germany, was created in order to find out an approach to let a user improve his or her localization ability before using an interactive application. This application can be a video game or a movie with a high level of sound displacement in the various scenes. The idea is to find an appropriate solution by means of a short experiment that consents the listener to achieve a higher sense of locating the direction of a sound, and at the same time trying not to annoy or stress him or her too much with a long-duration test, thus finding out a cheap, quick and effective way to deal with it.

Two HRTFs were chosen for the evaluation, which were the Cortex and the KEMAR HRTFs, already measured before in the anechoic room of the Fraunhofer Institute with the respective Cortex and KEMAR dummy heads. For the sake of simplicity, the sets with the measurements conducted setting a distance from the center of 1 m were chosen. Unfortunately, there was no more information regarding the methods used for the HRTF measurements. Half of the test population performed the test using KEMAR HRTF, the other half used the Cortex one. The sounds of the experiment were played entirely

through headphones.

Only the horizontal plane, thus azimuthal angles, was taken into account in order to simplify the judgment of the results. In order to have a thorough evaluation of the ability to discriminate between sound directions, 32 possible source positions were displayed homogeneously in a circumference on the horizontal plane at a distance of 1 m from the center. Not more than one position would be played at a time, in order to analyze the performances of just one direction and not that of sources playing simultaneously, which would confuse the user. The virtual sources are also positioned right respect to the localization blur criterion: since this value on the horizontal plane is for humans 1° in the frontal direction, 10° on the right-left axis, and 5° in the back direction [3], setting 32 sources means having a localization blur of 11.25°. Hence the users would theoretically suffer less from localization problems.

Partly according to the methodology used by Mendonça et al. [30], the test should have included an initial part, in order to analyze the localization performances of the subject at the beginning, an intermediate step, composed of a set of experiments designed to train the user to better recognize the sound directions within a short time, and a final part, done to assess the possibly modified perception ability of the experimentee. The experiments at issue were:

- *Select one source*: the user could select one position at a time and listen to it, in order to train himself or herself about that specific position.
- *Trial with feedback*: a source was displayed in one different position at a time and the listener had to guess its position. The answer was given to the source as a feedback in order to have a cognition of errors. The vicinity was displayed also following the criteria for vicinity described below.
- *Source going around*: a source playing while going in a circle clockwise at successive discrete steps starting from the front and the user had just to hear and focus on the sound direction. The source remained in a position

on the circle for 500 ms and then went on the next position while keeping playing the same sound file. This method was investigated because of its quick duration.

Such tests were chosen for their simplicity and short duration, following the initial idea of having a quick but effective training.

The data of 36 persons who took part in the experiment was collected in the end. In order to compare the results, the candidates did two out of the three training experiments proposed. Therefore there were 6 trials in which the subjects were divided, namely:

- Cortex 1+2: it included the "Select one source" and "Trial with feedback" experiments and the whole test was made synthesizing the Cortex HRTF with the played sound
- Cortex 1+3: it comprised the "Select one source" and "Source going around" experiments and the whole test was made by means the Cortex HRTF
- Cortex 2+3: it was composed of the "Trial with feedback" and "Source going around" experiments and the whole test was made using the Cortex HRTF
- KEMAR 1+2: same as the Cortex 1+2 but the KEMAR HRTF was employed
- KEMAR 1+3: same as the Cortex 1+3 with KEMAR HRTF
- KEMAR 2+3: same as the Cortex 2+3 using the KEMAR HRTF

In any of the cases, the number of participating subjects was then 6.

Impulse-like or short-duration set of sounds suit best for a listening test, because they help significantly in lateralization of signals [33, 34]. Furthermore, many sounds used in virtual reality environment like video games is of impulsive nature—for instance, sounds of a game in which steps, enemy voices,

shots, helicopters, and so on, appear. The choice of the sound to be played fell thus on the "Castanets" WAV file from the EBU archive[1]. The file was subsequently converted from stereo to mono and with a sampling frequency of 48 kHz instead of 44.1 kHz, in order to be adapted to the rendering system described in section 3.4.

The test was then carried out following the steps here described:

- The person received an oral explanation of the entire method of the experiment, in order to let understand what he or she was going to do and possibly avoid any misunderstanding or mistake.
- The subjects were then asked to sit down and wear the headphones. After a brief written repetition of the procedure, they were given an initial test, in which a primary localization of the virtual sources had to be done without a previous training.
- Next, the training session began. Depending on the case, the experimentee was told to select the respective tests to do and there were some recommendations to follow according to the choice: that is to say, in case of the "Select one source" experiment, the recommendation was to repeat the listening trying at least 10 different sources, whereas if the experiment "Source going around" was selected, the user should let the source play at least for two consecutive laps.
- Ultimately, the user had to do the final test, which was designed exactly in the same way as the initial test.

In the initial test, the "Trial with feedback" experiment, and the final test, the number of trials for the sources were 32. As a matter of fact, each source was played once in a random order, though the listener was just aware of

[1] https://tech.ebu.ch/publications/sqamcd

the number of the trials for the experiments and not of this fact. The decision of playing the sources once for every direction was taken for the sake of comparison of the final results.

The total duration of the test was estimated on average to be about 15 minutes, in any case no more than 20. In an ideal implementation in an application, this time would be even less, approximately less than 10 minutes, because the initial and the final test would not be needed anymore. After all the experiments, an analysis of the most performing experiment case among the six reported above was done. Then, based on the work of Mendonça et al. [30], a set of informal listening tests was set up regarding the effects of the training in relation to time. To wit, one subject repeated the whole test with the most efficient configuration of training steps in different instants of time and the evaluation of the progress with respect to the starting point was performed after that.

3.2 Population of interest

Normal people with average hearing were sought and invited to be subjects of the experiment, since the entire procedure was designed for general purposes, i.e. not requiring any particular propensity to discriminate sounds. Although around 40 persons took part in the experiment, only the data of 36 of them were taken into account, due to a need of having the same number of participants per experiment case and to some errors that affected the results. Some case studies were made regarding people who had previously taken audio tests, thus experienced subjects, and those who had already had or were having hearing problems, in order to see how the performance of these target groups could be affected. To this purpose, it was asked to the participants at the end of every test if they pertained to each of these two groups.

Figure 3.1: The type of headphone (*AKG K701*) used in the experiment

3.3 Instruments and software

Headphones were preferred to use in this experiment because they lead to a better channel separation and noise suppression. According to what is already written in section 2.5.1, they are of great importance in binaural technology. In addition, the auralization comes out more naturally if the sound is played with headphones, because loudspeakers' natural diffuse field effects such as ambience reflections, diffractions, or wave superpositions, are avoided, thus achieving a condition very close to a free sound field. While speakers give out a more natural auditory environment, having superior externalization properties and thus also creation of frontal sound images, rear sound images are harder to perceive [22].

Furthermore, it is well known that using headphones there are externalization problems resulting in the effects already explained in section 2.5.3. Despite this, sounds reproduced by means of headphones can be easily controlled since there is in principle no need to apply cross-talk cancellation filters—which are instead required for loudspeakers—and there are no sweet spot limitations for headphones. At any case, headphones remain popular especially for multimedia PC applications and virtual acoustic displays [23]. Due to these facts, the decision was to use headphones for reproduction, and eventually, a headphone

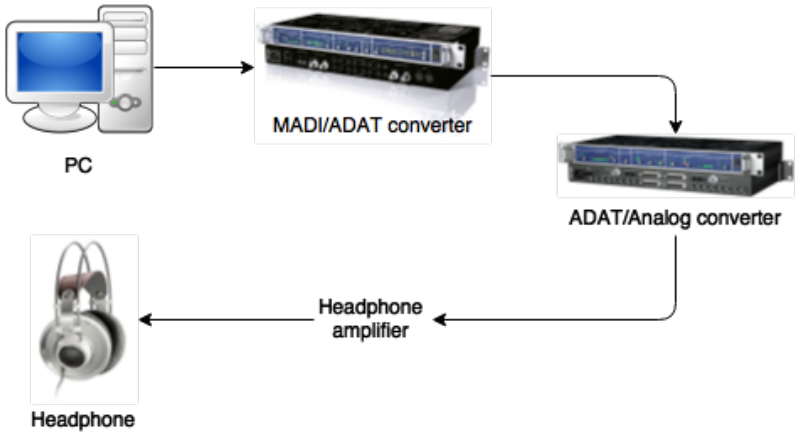


Figure 3.2: Schema of the hardware setup for the experiment

of the type *AKG K701* was utilized (see figure 3.1).

The experiment ran on a Linux machine with *CentOS 6* as operating system. This computer was suitable for running the experiment software since it was intended to be left exclusively for performance applications and also because it had a configuration compatible with the audio rendering tool. The sound renderer was already implemented and it sufficed to know how to configure it.

With the aid of an *XML* configuration file containing the definition of all the block elements that allowed the sound to be processed and directed to the audio interface and could also communicate with the visual interface, the output of a suitable binaural audio was given to the headphones. Particularly in the machine, the sound card was provided with an interface following the *MADI (Multichannel Audio Digital Interface)* protocol standardized by the Audio Engineering Society (AES)[2]. In figure 3.2 is shown the complete hardware

[2]See http://www.aes.org/tmpFiles/aessc/20150630/aes10-2008-r2014-i.pdf.

setup[3,4].

The visual interface with the rendering section was actually embedded in a software framework, developed in the Fraunhofer institute, called *UI.FM*. It includes many other applications that represents manifold aspects of sound and spatial audio production, besides the fact that it is also designed for mobile devices and for a simultaneous use of more than one designer. It is rendered for applications to be loaded with Google Chrome browser. One of the particular aspects of this framework is to build up a customized audio scene by means of a model made up of, among other things, planes, sound sources, listeners, and speakers.

3.4 Implementation

When a mono source is convolved with a pair of head-related impulse responses, namely when, being $s(t)$ a mono signal source and having two different HRTFs that account for the left and the right ear respectively, the principle explained in section 2.5.3 holds, with the HRTFs belonging to a predetermined database of head-related transfer functions, and therefore any direction of sound incidence can be simulated [3, p. 142-144]. In the approach here presented, the convolution is done by means of filter blocks provided by the renderer.

The sound mapping and processing and the communication with the interfaces are all defined in the XML configuration file provided to the renderer and are depicted in figure 3.3. The FIR filter had a length of 512 samples and it had a set of as many filters as the number of HRTF files, whose path was described in the configuration. All the blocks were described in the XML file

[3]Specifications of the MADI/ADAT converter: http://www.rme-audio.de/en/products/adi_648.php

[4]Specifications of the ADAT/Analog converter: http://www.rme-audio.de/en/products/adi_8_ds.php

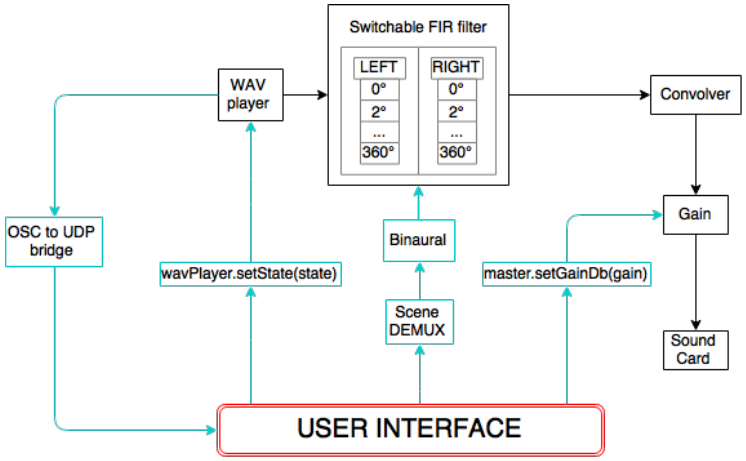


Figure 3.3: Diagram of the blocks and connections of the sound renderer, as given in its basic XML configuration file

of the renderer, which recalled their C++ implementation. The red-contoured block represents the user interface, the black-contoured blocks are the rendering blocks including the filters and the interface with the sound output, and the blue-contoured blocks stand for the messaging section.

Regarding the connections, the black arrows between the rendering components mean the passing of audio data, while the blue ones denote the exchanging of OSC (Open Sound Control) messages, carried using the UDP protocol, between the blocks. The channels going out from the main filter are certainly two, each one accounting for the left or right channel to bring to the external output. Being two HRTFs to be used, there was one rendering file with the paths of the Cortex data and another one with the same structure but with the paths of the KEMAR data.

The user interface programming framework is built up as a web application container and therefore is based on *JavaScript* programming. Its main server has to be started with the *Node.js* environment, so as to load particular modules that aid in the construction of an application. The MVC (Model-View-Controller) paradigm is employed in UI.FM and thus, while building an application, a programmer must carefully take this aspect into account. There

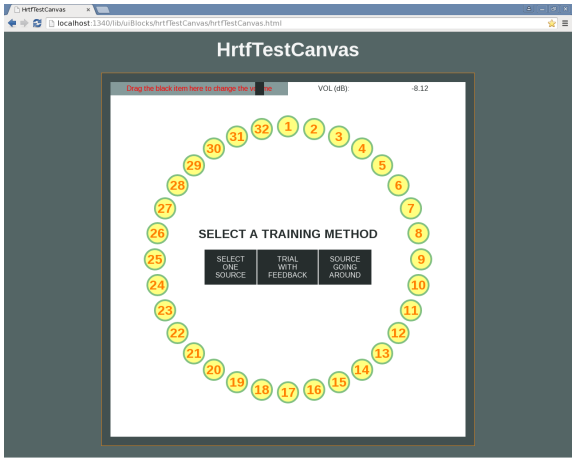


Figure 3.4: Screenshot of the main window of the test

are predefined as well as customizable JavaScript classes for the construction of each of the components. The application made for the listening test included a single source, taken from the model, who was placed around the plane by modifying its original coordinates and switched on or off according to the case. The source position values were passed automatically to a module communicating with the renderer. Buttons and instructions for the experiment were located at the center of the scene as in figure 3.4 and there was also a tool to regulate the volume located to the upper left of the screen at the user's disposal.

3.5 Distance and score evaluation

While the listeners made the test, some method of evaluation of the distance between the clicked position and the right position must have been done in order to assess the performance and the errors made by the person concerned. An easy approach was to take into account an array of positions that included the right circle, its three consecutive neighbor positions on its left, and those on its right. The selection contributed to a final score which was the general assess of the listener performance. This method was also used in the experiment

"Trial with feedback" to show the user the relative position of its choice for the audio source after each try. The following held:

- If the clicked position was also that of the virtual source, it would count 1 point to the total score. In the feedback experiment, the sign "Right" would appear to the user and the corresponding circle would light up in green.
- If the clicked position was one position far away from the virtual source, left or right, it would add 0.75 points to the total score. In the feedback experiment, the sign "Very close" would appear to the user and the corresponding circle would color in light orange.
- If the clicked position was two positions far away from the virtual source, it would bring 0.5 points to the total score. In the feedback experiment, the sign "Close" would appear to the user and the corresponding circle would become orange.
- If the clicked position was three positions far away from the virtual source, then 0.25 points would be added to the total score. In the feedback experiment, the sign "Roughly close" would appear to the user and the corresponding circle would change its color to light red.
- If instead the clicked position was none of the above, no points would be given. In the feedback experiment, the sign "Wrong" would appear to the user and the corresponding circle would be lit up in red.

The maximum score one could reach was of course 32. The same scoring system was then used in the initial and final tests, without the answers being shown. The total scores of these tests were at last compared in order to select the most well-performing experiment case out of the six listed in section 3.1 at line 56.

3.6 Data gathering and processing

The data were gathered by means of the JavaScript file of the test application, with which the results of the initial and final tests were produced. Since the file managing in JavaScript requires external modules, thus an additional work to integrate new modules in Node.js, the output of the scores was written to the HTML code of the produced page. To the first 21 participants, these results were shown as a report at the end of each initial and final test. During the course of the experimentation, from the 22nd subject on, they were decided not to be shown anymore, thinking about a possible source of compromising the results due to a probable more commitment by the user in doing the final phase. The data were then retrieved from the returned HTML code by means of the Google Chrome web console, copied in text files, and analyzed within a spreadsheet.

Means were made out for the total scores of each trial, divided by test case, and also an analysis was carried out between the Cortex and KEMAR cases and between those of the experiment sequences—i.e., having "Select one source" as experiment 1, "Trial with feedback" as experiment 2, and "Source going around" as experiment 3, the sequences were, respectively, 1+2, 1+3, and 2+3. In particular, the key values were differences between initial and final test scores, who stood for the effective improvements. As an additional analysis, the test results in the following cases were assessed:

- Experienced subjects, that is, subjects who already took part in other audio tests;
- Subjects having former or current hearing problems;
- Subjects to which their results were shown, at least after the first test.

Subsequently, the files were fed to a Python script in order to give out dispersion plots recalled in section 4.1.

Chapter 4

Analysis of findings

4.1 Presentation of data

The results reported in table 4.1 were calculated based on the method described in section 3.5. The left column of the table represents the different case studies. First, the general performance gain is given, i.e. the average improvement comprising all cases. Secondly, the enhancement calculated taking into account the cases of usage of Cortex and KEMAR HRTFs singularly is shown. Next, the improvement regarding the three general cases of training experiment sequences (to recall: number 1 corresponds to the "Select one source" experiment, number 2 represents the "Trial with feedback" test, and number 3 stands for the "Source going around" experiment) is listed. After that, the same values taken from the cases listed in section 3.1 at line 56, which are derived anyway from the combination of the Cortex and KEMAR with the "1+2", "1+3", and "2+3" cases, are printed out.

At last, there are particular scenarios which were decided to be taken into account and therefore analyzed. The first case is the scenario in which data from "experienced" persons, that is to say, subjects who already went under an audio test before, is calculated and shown. The second case represents the mean gain in experiments with participants that said to have, or have

Scenario	*Amelioration* (%)
General	9.14
Cortex	11.02
KEMAR	7.25
$1 + 2$	7.49
$1 + 3$	11.00
$2 + 3$	8.92
Cortex $1 + 2$	7.55
Cortex $1 + 3$	16.93
Cortex $2 + 3$	8.59
KEMAR $1 + 2$	7.42
KEMAR $1 + 3$	5.08
KEMAR $2 + 3$	9.24
Experienced	9.31
Hearing problems	5.58
Shown results	10.80

Table 4.1: Table showing the mean values of improvements for each single case

Scenario	*Initial performance* (%)	*Final performance* (%)
General	39.50	48.63
Cortex	40.80	51.82
KEMAR	38.19	45.44
Experienced	41.13	50.44
Hearing problems	40.63	46.21

Table 4.2: Table showing the initial and final mean results in the cases of interest

had, hearing problems. The third case, considered in the course of the whole experiment session, is that of people who were shown their performance after the initial and the final test. In table 4.2, instead, the mean pre-test and

post-test results of the general, Cortex, KEMAR—these ones taken for all subjects—experienced, and hearing problems scenarios are listed. These data are shown because they can be relevant, especially the initial part, in order to be compared with the general cases regarding performances.

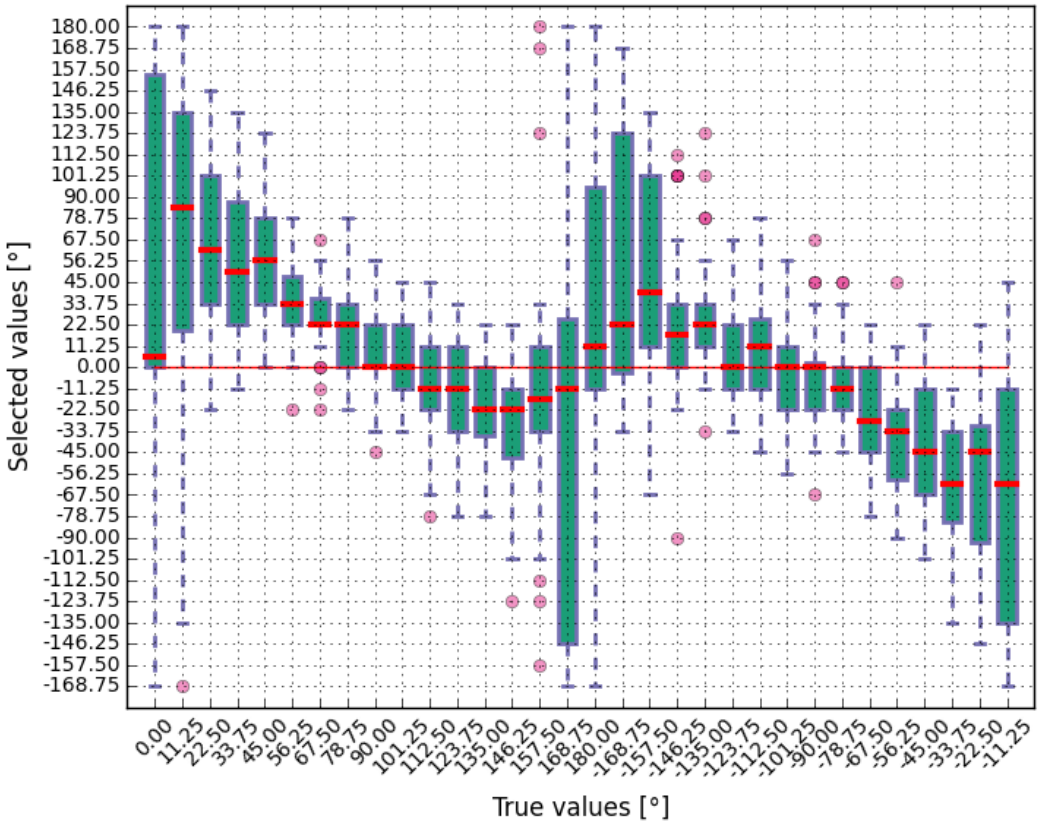


Figure 4.1: Data after the initial tests

Box plots with statistics of the angular distance from the angles selected by the users with respect to the given source positions (x axis) are here shown. In the following graphs, the x axis contains the various angles in which the sources or the positions were given, namely, the 0° corresponds to the front of the listener, the right angles are in clockwise steps with respect to the front until 180° (back), and the left angles go from -180° to 0° (front). On the

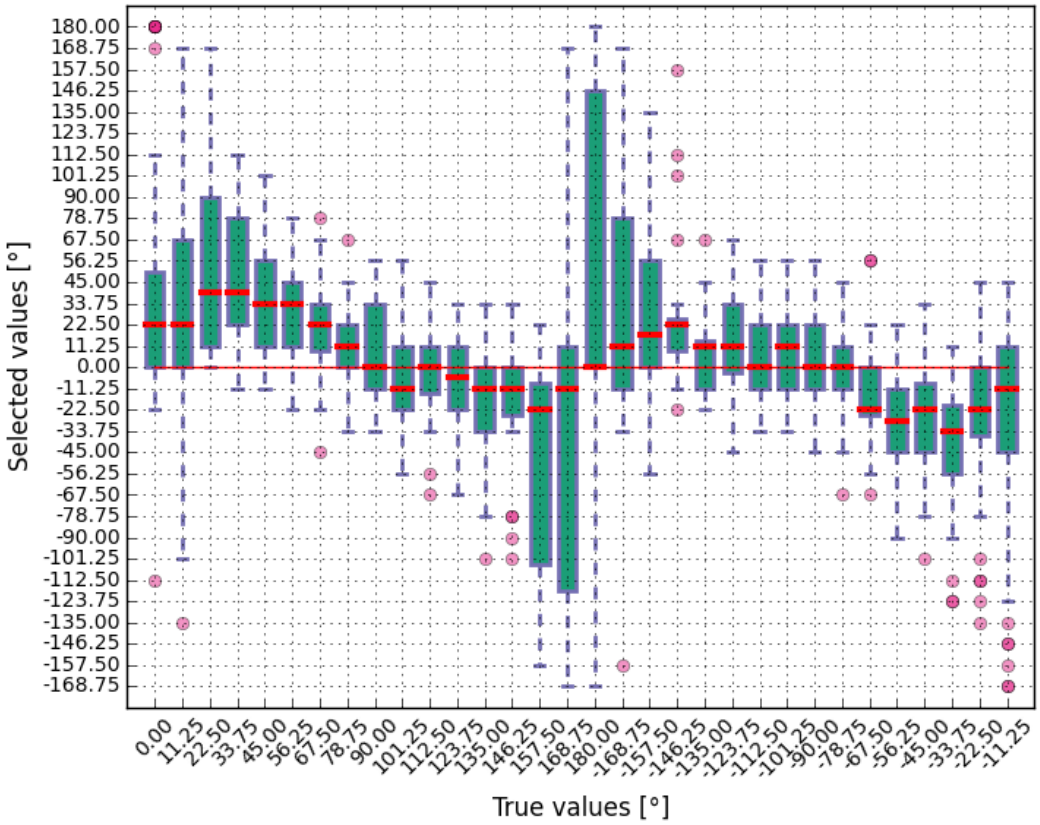


Figure 4.2: Data after the final tests

y axis instead, the difference in angles from the right position varies from -168.75° to 0° (left hemisphere relative to the playing source) and from 0° to 180° (right hemisphere). The red central horizontal line on the 0° value is the line representing the right-guessed values.

4.2 Quantitative analysis

According to table 4.1, it comes out that there is a general average gain in performance of 9.14% between the location of guesses in the initial tests and those in the final tests, hence the initial hypothesis is satisfied since this value

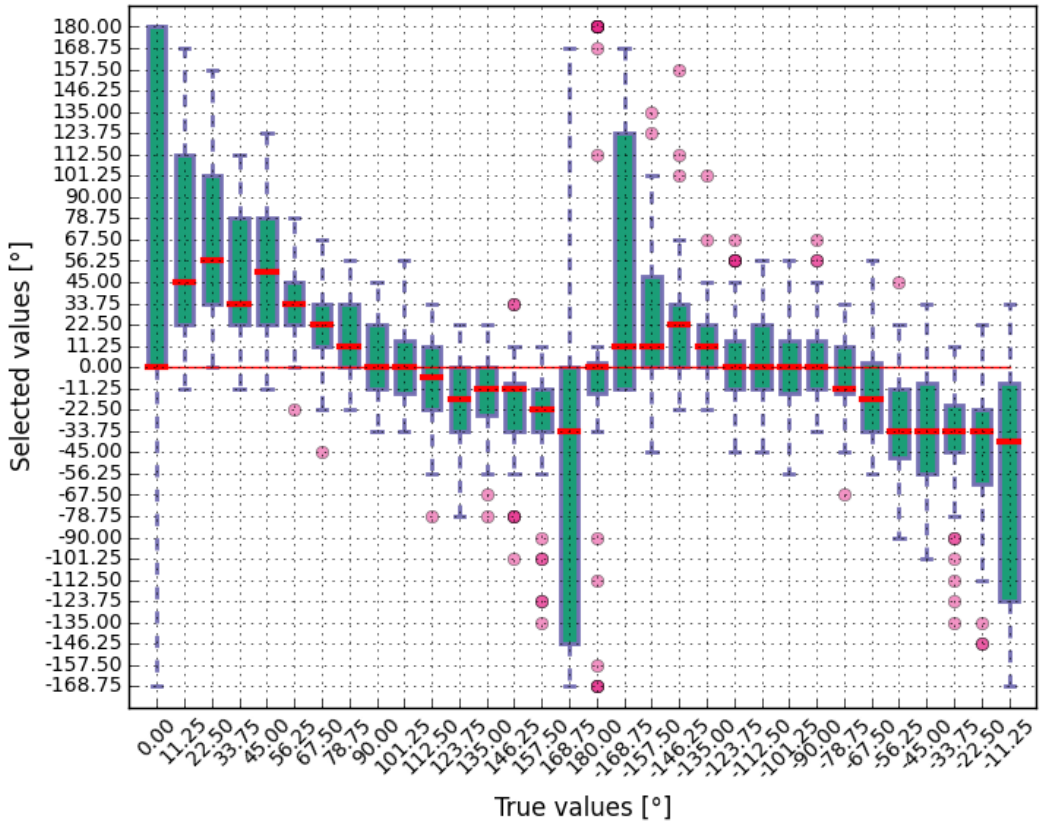


Figure 4.3: Data for the Cortex tests

is close to the foreseen 10%. With respect to the other cases, the Cortex scenario yields a higher improvement compared to the KEMAR case and the winning training configuration is the "1+3", but regarding the latter a remark must be done. The difference between the "Cortex 1+3" and the "KEMAR 1+3" scenarios is to notice, in particular, it seems that the KEMAR HRTF performs bad in combination with the experiments "Select one source" and "Source going around". There are no relevant differences between the Cortex and the KEMAR situation with respect to the "1+2" and the "2+3" training sequences.

Experienced users' performance does not seem to differ much from the av-

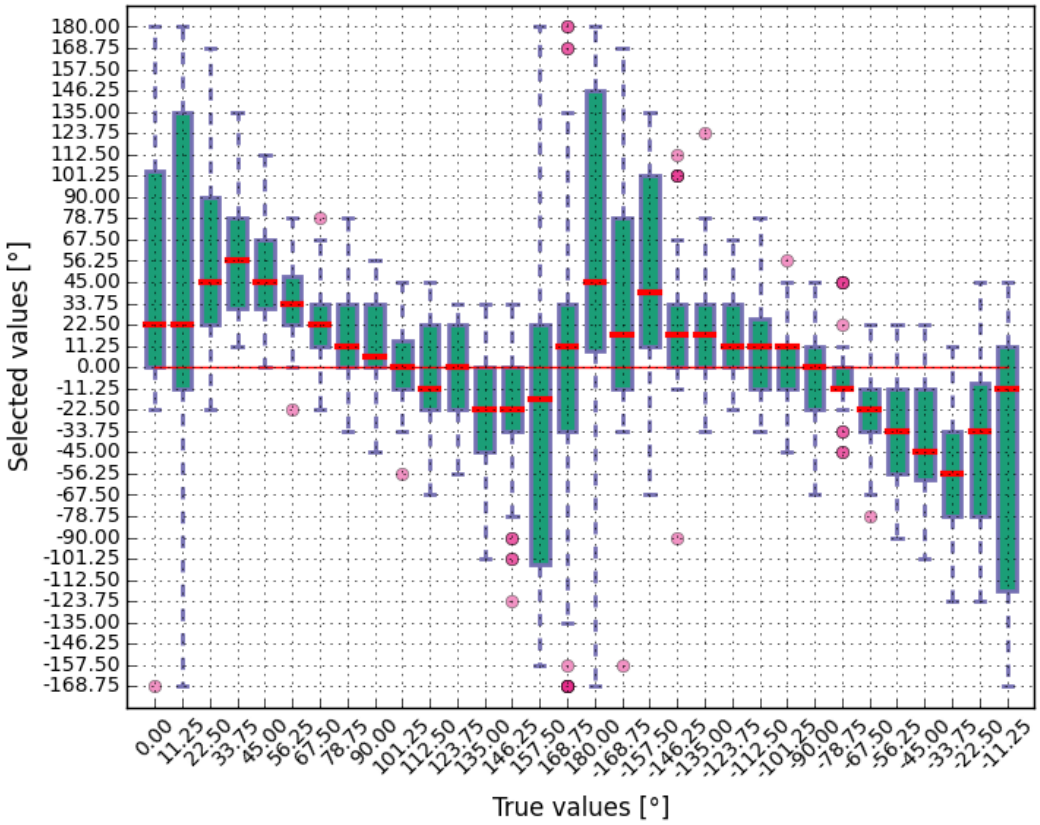


Figure 4.4: Data for the KEMAR tests

erage value while the achievement of the subjects to which the initial results were shown presents a slightly higher deviation. This value is actually affected by the fact that most of the trials in this case fell in the Cortex scenario. Persons affected or having been affected by hearing problems fall instead behind, as expected. From table 4.2 it comes to light that Cortex listeners and experienced listeners have almost the same performances. About the experiments in which the results were shown after the first test, the average gain seems to be relatively high but this included in most cases the use of Cortex, which is the most probable reason for the production of this value.

Comparing the plots of the performances of the initial test in figure 4.1 and

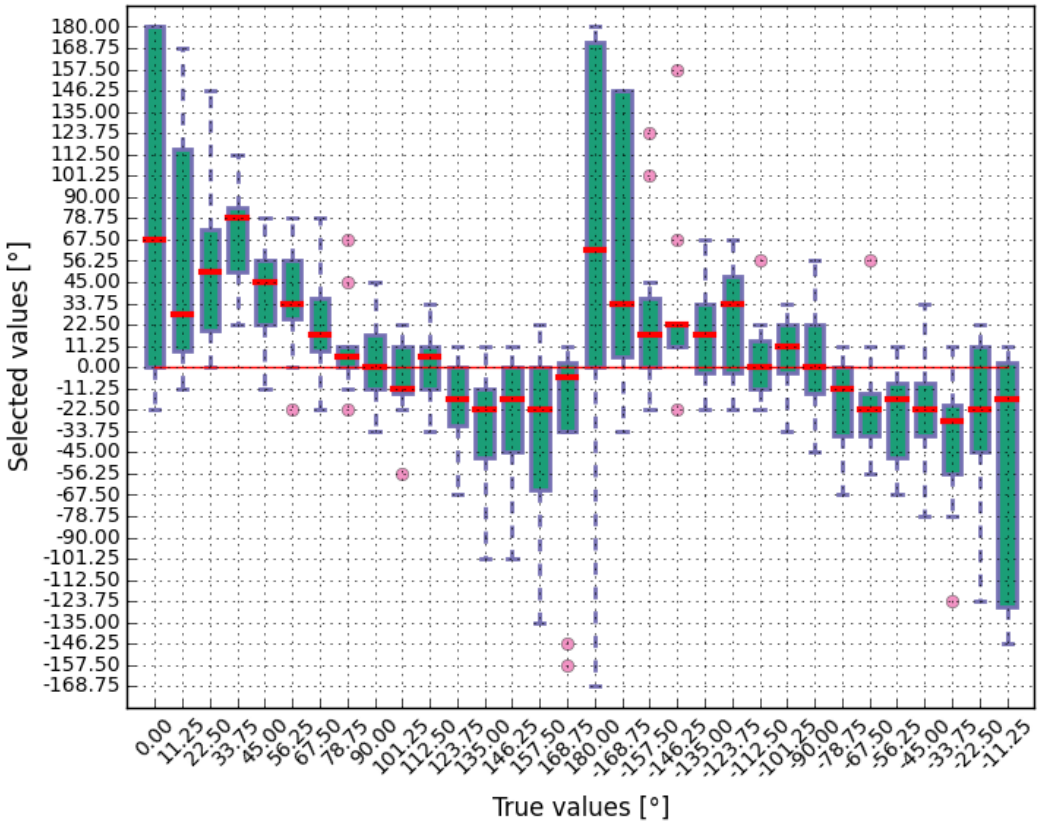


Figure 4.5: Data after the tests of the sequence "1+2"

those after the final test in figure 4.2, it is evinced that there is an increasing in the number of estimated sources close to the original one. From all the plots instead (from figure 4.1 to 4.9), it can be inferred that there is a strong tendency to select the source at the very left or at the very right—namely those at ±90° relative to the front—in case the incoming sound is played from the nearby sources, especially if they are three or four positions away (until about 40° degrees of circular distance). One of the participants of the experiment reported that sometimes he felt as the same source was repeated, although each source was actually heard once.

The fact that the Cortex HRTF gives the best performances is furthermore

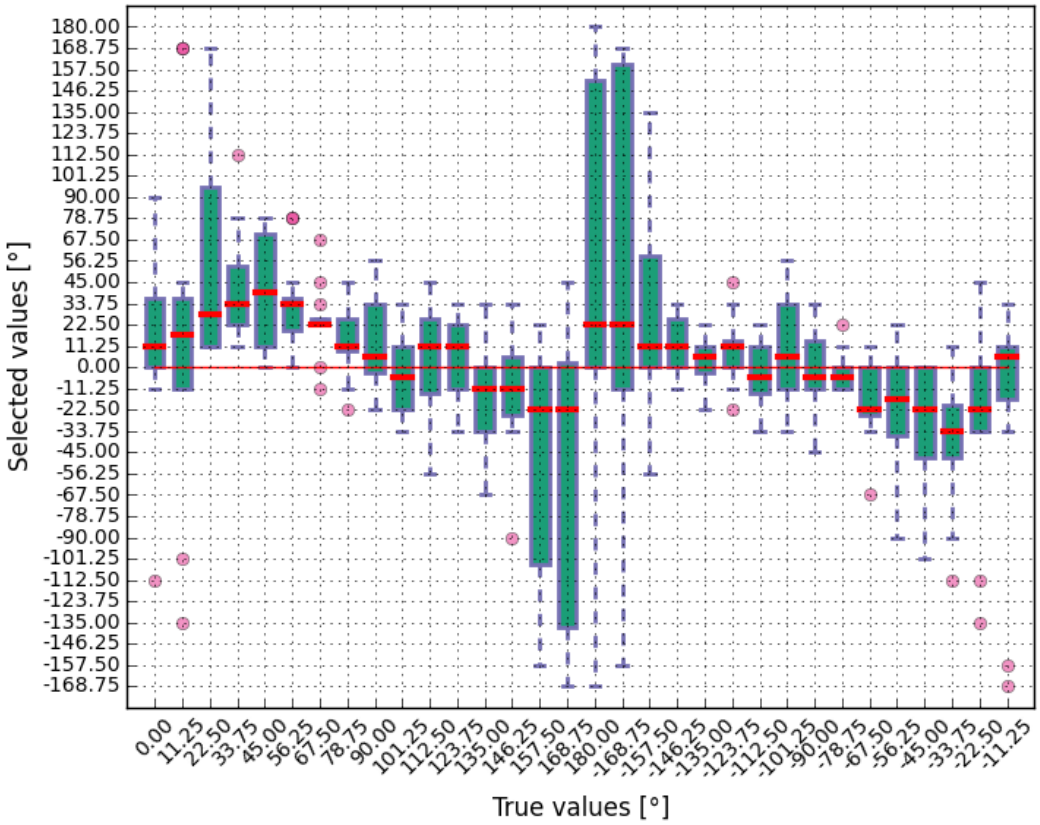


Figure 4.6: Data after the tests of the sequence "1+3"

analyzed in figure 4.3. From the plot, it emerges that this function has a finer resolution and helps better in solving front-back confusions compared to the KEMAR case in figure 4.4, where the choices in the central and edge columns are more scattered. It can also be seen that the displacement between the direction of the played sources and that of the selected sources is reduced with the aid of the Cortex HRTF.

About the training experiment results, the sequence of tests "1+2" performs worse, though with a little difference, in comparison to the other cases, and has the most scattered values on the plot (see figure 4.5). "1+3" is the configuration which gives the best performance in recognizing the lateral sources (see figure

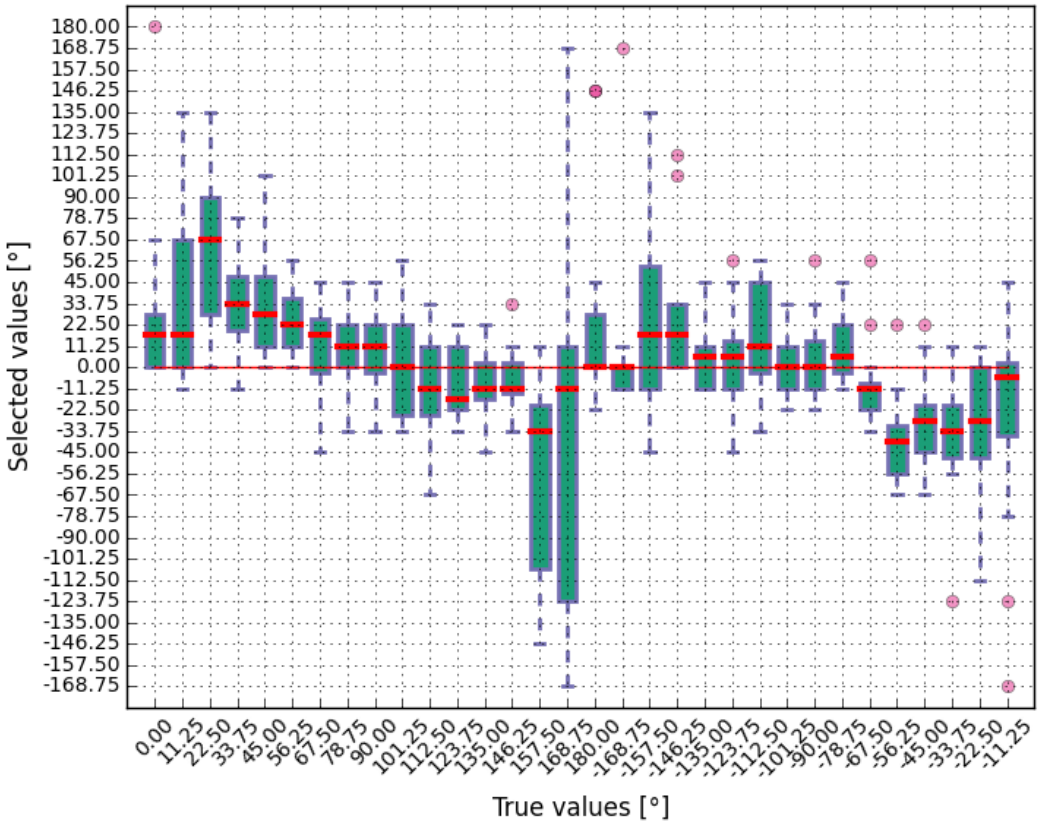


Figure 4.7: Data after the tests of the sequence "2+3"

4.6) whereas the sequence 2+3 attains the best outcome in the front-back confusion cone (see figure 4.7). The "Cortex 1+3" and the "KEMAR 1+3" cases' performances can be seen in details in figures 4.8 and 4.9. The scenario including KEMAR falls behind, and particularly for front-back confusions, as it can be seen that the dispersion is stronger in the plot. It can be surmised that the statistics of the sequence "1+3" depend highly on the choice of the corresponding HRTF.

Nonetheless, the "Cortex 1+3" scenario was discovered to be the most well-performing and was therefore used in the informal post-experiments. The results from these are reported in figure 4.10. As it can be seen from the

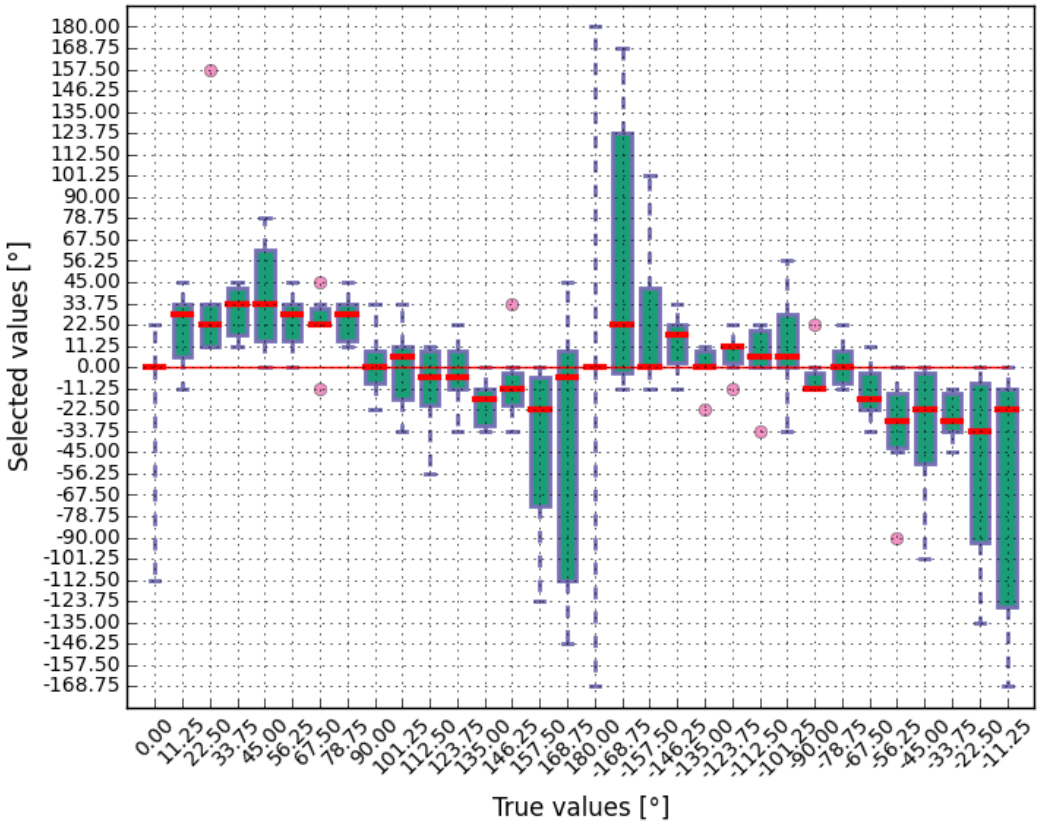


Figure 4.8: Data regarding the sequence “Cortex 1+3”

graph, the tests were done in different, though not continuous, instants of time, extending until about 10 days from the first event. The improvement is actually variable and does absolutely not show a steady pattern of its evolution over time, therefore it does not depend on it. Instead, the values collected after the final informal tests seem to converge within a relative narrow range, thus it can be said that the effect of the test lasts with the progress of time.

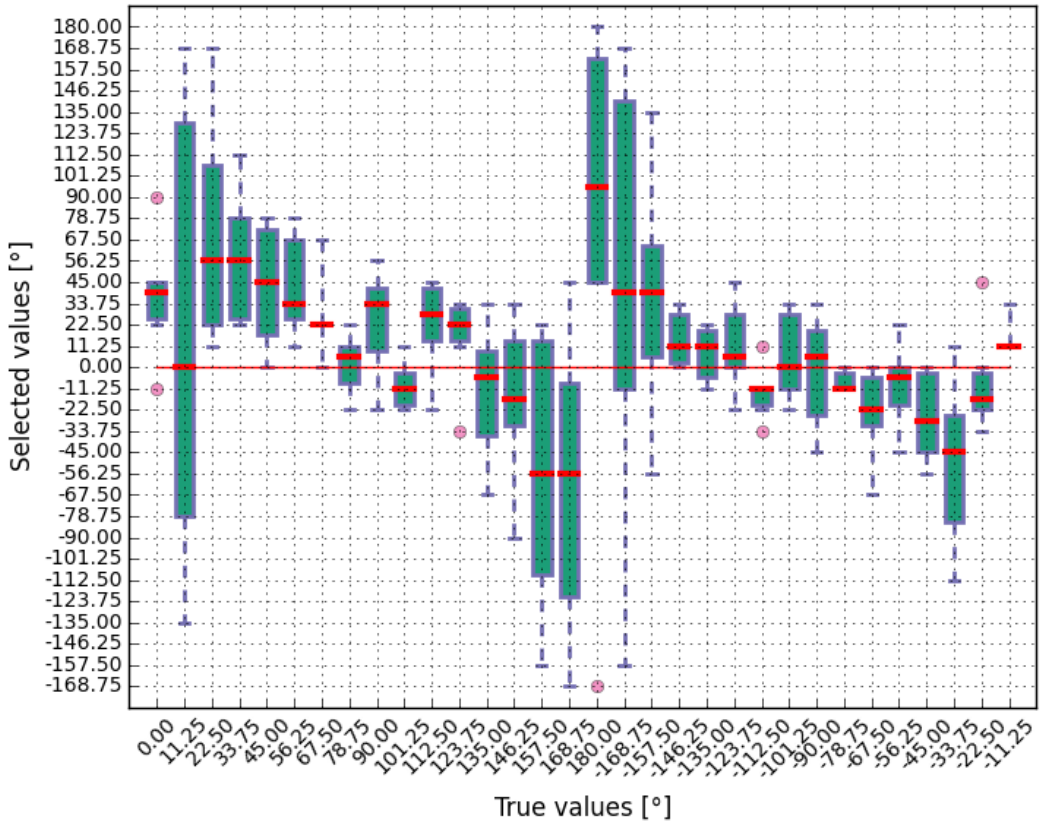


Figure 4.9: Data regarding the sequence "KEMAR 1+3"

4.3 Qualitative analysis

Five of the participants stated that the directions in the horizontal plane were perceived as being above, namely on the upper part of the frontal plane (also known as coronal plane, the one dividing the body in front and back). It can be also seen from the phenomenon resulting also in the graphs that the choice in the lateral angles went for the directions located exactly at $\pm 90°$ and this can be a consequence of the aforementioned effect, which can be also due to incorrect measurements of the HRTF in these directions. As already said in section 3.1 at line 14 there was no more information to be found with respect

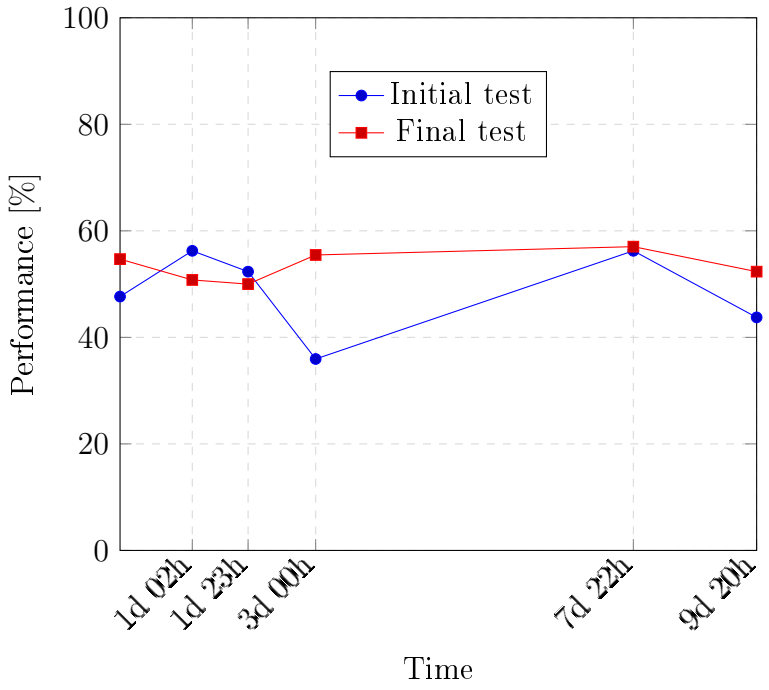


Figure 4.10: Results of the informal tests

to the HRTF measurements.

According to the sound produced by the system, the resolution of the front-back confusions was helped by the fact that the sources heard in the rear lacked of some high-frequency content, thus the sound played appeared less brilliant in comparison to the sources heard in the front. Although this complies also with the findings of Blauert [35, p. 107-116] and one of the main causes of this phenomenon is the shadowing of the body, it was actually a very subtle difference to be clearly distinguished.

The test in its whole was reported by some participants to have a certain degree of difficulty. This fact is due perhaps to a consistent number of sources which were put on the scene of the test. In fact, the number of sources could have been reduced, primarily because of the high level of mistakes at lateral angles. One subject reported also that there were some possible confusions be-

tween 45° and 135° and between -45° and -135° directions, whereas another one told that he felt that the audio was being "clipped" for some source positions, especially at the front.

4.4 Interpretation

The statistics concerning the experiment "Source going around" say that this test can help to improve the localization performances of a listener. Moreover, since it is also quick and requires a more dense concentration in following the moving audio source in a short period of time, it would not stress a user of an interactive audio application and is therefore recommended to be included as a training experiment in combination with one of the other two tests.

Chapter 5

Conclusions

5.1 Problem description

In this work, the goal was to create a suitable binaural environment for auditory training in interactive applications which include 3D audio content. Such an environment shall be made up, before starting the application, with a series of training experiments that would help the listener to improve his or her sense of location of items emitting sound, hence audio sources. The tests shall not last long, since the user would get tired otherwise and probably unwilling to play or watch.

In order to have a greater focus on plain sound, thus avoiding effects as room simulation or diffuse sound field, binaural technology was used in the audio reproduction through headphones. The position of the sound varied within a circle around the listener and the test comprises three parts, specifically:

- Initial test, to assess the localization ability before training;
- Training part, in which users did some experiments to improve this skill;
- Final test, identical to the initial, to check the performance after training.

The scenarios for each test subject were divided in six groups, according to:

- the Head-Related Transfer Function (HRTF) used—two HRTFs were taken under analysis;
- the types of training sequences—the performance of three experiments, gathered in sequences of two, were estimated.

After that, the statistics regarding improvements brought by different scenarios, depending on the HRTF or the training methods used, were evaluated. Also cases of experienced subjects and persons with hearing problems were further analyzed. At last, a study concerning the effects of such experiment over time was carried out.

5.2 Salient findings

The most salient findings are listed in table 4.1. Some values stand particularly out, for instance, the improvement values of:

- the Cortex scenario in comparison to the KEMAR;
- the "1+3" sequence of experiments compared to "1+2" and "2+3", though its performance varies depending on the HRTF used;
- the "Cortex 1+3" scenario against the other six of the same group.

Although the best-performing case is "Cortex 1+3", an eventual use of one of the scenarios including the "2+3" sequence would help more in discriminating the front-back confusions, that is, sources presented to the front while heard from the back and vice versa.

5.3 Recommendations and further studies

The scenario "Cortex 1+3", precisely that having HRTF data from Cortex dummy head and the training experiments "Select one source" and "Source

going around", is recommended to be implemented in an ideal 3D audio interactive application. The use of headphones against loudspeakers is advisable for the features of decoupling and isolation of sound channels.

To improve the results, the study of other cases, including experiments that would differ from those used in this work, scenarios with different audio files to be displayed, in order to analyze the people's reaction to new sounds, as well as scenarios with a lower number of sources, so as to reduce the number of errors while having a simpler scenario, are suggested for the future.

Another possible study would be to include the control of head movements. Head movements help improving front-back discriminations and their supervision can be done with the support of head-tracking equipments, in order to track the position of the head and to allow certain types of movements along a given trajectory. Head tracking has already been introduced in some interactive devices to avoid possible mismatches between the actual and the perceived localization in the headphone. But since these devices require a strong effort and design to be built, their price is not accessible sometimes. As always, the higher the price, the better the quality, so this can be an obstacle in terms of costs.

References

[1] Janina Fels. "Trends in Binaural Technology". In: *Proceedings of AIA-DAGA Conference on Acoustics*. 2013.

[2] J. Oberem, B. Masiero, and J. Fels. "Authenticity and naturalness of binaural reproduction via headphones regarding different equalization methods". In: *Proceedings of AIA-DAGA Conference on Acoustics*. 2013.

[3] Michael Vorländer. *Auralization*. Ed. by Springer. 1st ed. Springer-Verlag Berlin Heidelberg, 2008.

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

[5] Wayne Olsen and Raymond Carhart. "Head Diffraction Effects on Ear-Level Hearing Aids". In: *Audiology* 14 (1975), pp. 244–258.

[6] John Sunier. *The story of stereo, 1881-*. Gernsback Library, 1960. Chap. 2, pp. 27–47.

[7] Joseph Hull (Marketing Communications Manager at Dolby Laboratories Inc.) "Surround sound: Past, Present, and Future. A history of multichannel audio from mag stripe to Dolby Digital". In: *Dolby brochure*. Dolby Laboratories Inc., 1997.

[8] Kris Sangani. "The vastness of sound [Production Cinema]". In: *Engineering and Technology Magazine* 8.6 (July 2013), pp. 78–79.

[9] Geoff Martin. *Introduction to Sound Recording*. URL: http://www.tonmeister.ca/main/textbook/.

[10] Günther Theile and Helmut Wittek. "Wave field synthesis: A promising spatial audio rendering concept". In: *Proceedings of the 7th Int. Conference on Digital Audio Effects (DAFx'04), Naples, Italy*. Oct. 2004.

[11] Markus Noisternig, Alois Sontacchi, Thomas Musil, and Robert Holdrich. "A 3D Ambisonic Based Binaural Sound Reproduction System". In: *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. June 2003.

[12] Michael Gerzon. "Don't say quad—say psychoacoustics". In: *New Scientist* 76 (Dec. 1977), pp. 634–636.

[13] Stephan Paul. "Binaural Recording Technology: A Historical Review and Possible Future Developments". In: *Acta Acustica united with Acustica* 95.5 (Sept. 2009), pp. 767–788.

[14] Bertil Nordlund and Gunnar Lidén. "An Artificial Head". In: *Acta Otolaryngologica*. Vol. 56. 2-6. 1963, pp. 493–499.

[15] Benjamin B. Bauer, Allan J. Rosenheck, and Louis A. Abbagnaro. "External-ear replica for acoustical testing". In: *The Journal of the Acoustical Society of America*. Vol. 42. 1. July 1967, pp. 204–207.

[16] Ralf Kürer, Georg Plenge, and Henning Wilkens. "Correct Spatial Sound Perception Rendered by a Special 2-Channel Recording Method". In: *Audio Engineering Society Convention 37*. Oct. 1969.

[17] K. Genuit. "Ein breitbandiger rauscharmer Kunstkopf mit hoher Dynamik und der Eigenschaft der originalgetreuen Übertragung von Hörereignissen". EP0126783 A1. EP Patent App. EP19,830,105,141. Dec. 5, 1984.

[18] E. A. G. Shaw. "External Ear Response and Sound Localization". In: *Localization of Sound: Theory and Applications*. Ed. by R. W. Gatehouse. Amphora Press, 1982.

[19] F. L. Wightman and D. J. Kistler. "Headphone simulation of free-field listening. I: Stimulus synthesis". In: *The Journal of the Acoustical Society of America* 85.2 (Feb. 1989), p. 858.

[20] A. Meshram, R. Mehra, Hongsheng Yang, E. Dunn, J.-M. Franm, and D. Manocha. "P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound". In: *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. Sept. 2014, pp. 53–61.

[21] V. Ralph Algazi, Carlos Avendano, and Richard O. Duda. "Elevation localization and head-related transfer function analysis at low frequencies". In: *The Journal of the Acoustical Society of America* 109.3 (Mar. 2001), pp. 1110–1122.

[22] Chong-Jin Tan and Woon-Seng Gan. "Concha Excitation as a Means of Introducing Individual Cues for Spatial Hearing". In: *Audio Engineering Society Convention 106*. May 1999.

[23] E.M. Wenzel, M. Arruda, D.J. Kistler, and F.L. Wightman. "Localization using nonindividualized head-related transfer functions". In: *The Journal of the Acoustical Society of America* 94.1 (July 1993).

[24] Patrick Satarzadeh, V. Ralph Algazi, and Richard O. Duda. "Physical and Filter Pinna Models Based on Anthropometry". In: *Audio Engineering Society Convention 122*. May 2007.

[25] S. Spagnol, M. Geronazzo, and F. Avanzini. "On the Relation Between Pinna Reflection Patterns and Head-Related Transfer Function Features". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.3 (Mar. 2013), pp. 508–519.

[26] Bruno Masiero and Janina Fels. "Perceptually Robust Headphone Equalization for Binaural Reproduction". In: *Proceedings of the AES 130th Convention*. Audio Engineering Society. London, UK, May 13-16, 2011.

[27] R. Bucklein. "The audibility of frequency response irregularities". In: *Journal of the Audio Engineering Society* 29.3 (1981), pp. 126–131.

[28] S. Schmidt. "Finite Element Simulation of External Ear Sound Fields for the Optimization of Eardrum-Related Measurements". PhD. Ruhr-Universität Bochum, 2009.

[29] J.C.B. Torres, M.R. Petraglia, and R.A. Tenenbaum. "HRTF modeling for efficient auralization". In: *2003 IEEE International Symposium on Industrial Electronics (ISIE)*. Vol. 2. June 2003, pp. 919–923.

[30] Catarina Mendonça, Jorge A. Santos, Guilherme Campos, Paulo Dias, and José Vieira. "On the Adaptation to Non-Individualized HRTF Auralizations: A Longitudinal Study". In: *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*. Mar. 2012.

[31] Bernhard U. Seeber and Hugo Fastl. "Subjective selection of non-individual head-related transfer functions". In: *Proceedings of the 2003 International Conference on Auditory Display* (July 6-9, 2003).

[32] Yukio Iwaya. "Indvidualization of head-related transfer functions with tournament-style listening test: Listening with other's ears". In: *Acoustical Science and Technology*. Vol. 27. 6. 2006, pp. 340–343.

[33] Piotr Majdak and Bernhard Laback. "Effects of interaural time differences in fine structure and envelope on lateral discrimination in electric hearing". In: *The Journal of the Acoustical Society of America*. Vol. 120. 4. Oct. 2006, pp. 2190–2201.

[34] Richard J. M. van Hoesel and Richard S. Tyler. "Speech perception, localization, and lateralization with bilateral cochlear implants". In: *The Journal of the Acoustical Society of America*. Vol. 113. 3. Mar. 2003, pp. 1617–1630.

[35] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1983.