3-D Sound Enhanced Presence in Virtual Environments

António Sérgio Pádua Videira de Azevedo

Thesis to obtain the Master of Science Degree in Information Systems and Computer Engineering

Examination Committee

Chairperson: Prof. Mário Jorge Costa Gaspar da Silva
Supervisors: Prof. Joaquim Armando Pires Jorge
            Prof. Pedro Filipe Faria Campos
Member of the Committee: Prof. João António Madeiras Pereira

June 2013
Where's your will to be weird?

Jim Morrison
Acknowledgments

During this journey I learned a lot. Not only I made good friends, but also had the possibility to dive into new worlds and knowledge. Many people helped me choosing the right path and I dedicate them this Thesis.

First I would like to thank Ana that supported me unconditionally every day: now we can go to a next level. Second, to my father, my mother and my sister that kept standing at my side in the good and the bad. Also to my second father mother and sister, that were always available to help.

To my advisor and co-advisor, Prof. Joaquim Jorge and Prof. Pedro Campos who guided me so many times. A special thanks do Pedro Lopes who introduced me to the “research world” and Bruno de Araújo who always had time and wisdom to share. A big hug to my friends and kind regards to Tiago Guerreiro, Luis Bruno, Maurício Sousa and others that supported me.

This thesis was framed in project CEDAR and project 6DSPACES. CEDAR was partially supported by the Portuguese Foundation for Science and Technology (FCT): project CEDAR PTDC/EIA-EIA/116070/2009, project 3DORuS PTDC/EIA-EIA/102930/2008 and project PEst OE/EEI/LA0021/2011. 6DSPACES was supported by the Instituto de Desenvolvimento Regional (IDR) under the Program Intervir+ (+Conhecimento) MADFDR-01-0189-FEDER-000001.
Abstract

Virtual Environments are being widely used to improve communications between collocated users. A development of a Virtual Environment is mostly concerned in offering a high quality graphics visualization system to enhance presence. It is known that sound can reduce visual workload and improve awareness of information outside users’ field of view. Impressively, these values are usually forgotten, and the sound system is just used to reproduce some 2-D sound effects.

To encourage the use of a 3-D sound system when developing an Immersive Virtual Environment, we performed three different experiments. Our focus was the benefits of using 3-D sound to improve navigation and enhance presence. We also measured the subjective perception of sound when the loudspeakers were hidden or visible. 3-D sound proved to be essential not only to enhance presence (experiment 1) but also to improve navigation, orientation and way-finding (experiment 2). Experiment 3 proved that a simple, cheap and effortless technique enables greatest positioning errors between the image and the sound and improves users’ confidence and correctness.

Keywords

3-D sound, Sonification, Virtual Environments, Presence, Ventriloquism Effect
Resumo

Os Ambientes Virtuais têm sido vastamente usados para melhorar a comunicação entre pessoas distantes. Quando se desenvolve um Ambiente Virtual, preocupa-se maioria de sistema de alta definição de imagens para tentar aumentar a presença. Sabe-se que o som pode reduzir a poluição visual e melhorar a atenção a objetos fora do campo visual. Impressionantemente, estes saberes são normalmente esquecidos, e o sistema sonoro acaba por apenas reproduzir alguns sons em 2 dimensões.

De forma a encorajarmos o uso de um sistema de som 3-D quando se desenvolvem Ambientes Virtuais Imersivos, efetuámos 3 experiências diferentes. O nosso foco foi o benefício que o som 3-D pode trazer a um Ambiente Virtual, de forma a melhorar a navegação e aumentar a sensação de presença. Também avaliámos quais as diferenças na percepção áudio-espacial, quando o utilizador vê os amplificadores ou quando não os vê.

O sistema de som 3-D que usámos, provou que era essencial não só para aumentar a sensação de presença (experiência 1) como também para melhorar a navegação e a orientação (experiência 2). Na terceira experiência provámos que um método simples e barato (esconder os amplificadores do campo visual do utilizador) pode de facto melhorar a percepção áudio-espacial. Pode também aumentar a confiança e diminuir a percentagem de erro quando um utilizador “julga” de onde vem a fonte sonora.

Palavras Chave

Som 3-D, Sonificação, Ambientes Virtuais, Presença, Efeito de Ventríloquo
# Contents

1 Introduction ................................... 1
   1.1 Motivation ................................... 2
   1.2 Problem Statement .............................. 3
   1.3 Contributions of our work ...................... 3
   1.4 Publications .................................. 5
   1.5 Thesis Outline ................................ 5

2 Background ..................................... 6
   2.1 Fundamentals of Sound ......................... 7
      2.1.1 Propagation of Sound .................. 7
      2.1.2 Complex Waves .......................... 7
      2.1.3 Sound Levels and the Decibel .......... 8
   2.2 Psychoacoustics ................................ 9
      2.2.1 Ear Anatomy ............................. 9
      2.2.2 Loudness Vs Frequency .................. 10
      2.2.3 Area of Audibility ...................... 10
      2.2.4 Loudness Vs Sound-Pressure Level ...... 10
      2.2.5 Audibility of Loudness Changes .......... 10
   2.3 Spatial Audio Perception ..................... 11
      2.3.1 Interaural Time Difference and Interaural Level Difference .... 11
      2.3.2 The Precedence Effect .................. 12
      2.3.3 Head Movement and Source Movement Cues .......... 13
      2.3.4 Head-Related Transfer Function .......... 13
      2.3.5 Distance Perception ..................... 14
      2.3.6 Reverberation ............................ 15
      2.3.7 The Ear as an Analyser .................. 16
      2.3.8 Other Auditory illusions ................. 16
   2.4 3-D Audio Spatialization Challenges ............ 17
   2.5 Spatial Audio Models ......................... 18
      2.5.1 Binaural ................................ 18
      2.5.2 Ambisonics .............................. 18
3 Related Work

3.1 Virtual Environments

3.1.1 Presence and Awareness in Virtual Environments

3.1.2 How to measure Presence in Virtual Environments

3.1.3 Visual-Only Virtual Environments

3.1.4 Auditory-Only Virtual Environments

3.1.5 Audio-Visual Virtual Environments

3.1.6 Multi-Sensory Virtual Environments

3.2 Sonification Techniques

3.2.1 Auditory Icons

3.2.2 Earcons

3.2.3 Data Sonification

3.2.4 Motion Synthesis

3.2.5 Soundscapes

3.2.6 Emotions

3.3 The state of the art

3.3.1 What is being forgotten

4 Assembling the Virtual Environment

4.1 Spatial Audio Model Used

4.2 Simulation Room

5 Experiments

5.1 Experiment 1 - Multi-Sensorial Stimulation and Presence in Outdoor VEs

5.1.1 Motivation

5.1.2 Participants

5.1.3 Virtual Environment Specification

5.1.4 Properties

5.1.5 Procedures

5.1.6 Stimuli Configurations

5.1.7 Experimental Design

5.1.8 Results

5.1.9 Discussion

5.1.10 Summary

5.2 Experiment 2 - Navigation Task

5.2.1 Motivation

5.2.2 Participants
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>CEDAR (Collaborative Engineering Design And Review) Virtual Environment</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>In-phase complex wave</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>Not in-phase complex wave</td>
<td>8</td>
</tr>
<tr>
<td>2.3</td>
<td>Composition of the human auditory system</td>
<td>9</td>
</tr>
<tr>
<td>2.4</td>
<td>Equal-loudness contours of the human ear</td>
<td>10</td>
</tr>
<tr>
<td>2.5</td>
<td>Auditory area of the human ear</td>
<td>10</td>
</tr>
<tr>
<td>2.6</td>
<td>A listener with a perfectly round head</td>
<td>11</td>
</tr>
<tr>
<td>2.7</td>
<td>Difference in localization accuracy for HRTF</td>
<td>16</td>
</tr>
<tr>
<td>3.1</td>
<td>Allosphere</td>
<td>23</td>
</tr>
<tr>
<td>3.2</td>
<td>Inside view of the Allobrain</td>
<td>24</td>
</tr>
<tr>
<td>3.3</td>
<td>User navigating in a CAVE-like VE with auditory cues</td>
<td>24</td>
</tr>
<tr>
<td>3.4</td>
<td>Sighted and visually impaired users collaborating</td>
<td>25</td>
</tr>
<tr>
<td>3.5</td>
<td>World Blind Football Championship 2010</td>
<td>25</td>
</tr>
<tr>
<td>3.6</td>
<td>Sound can improve the perceived quality of objects</td>
<td>26</td>
</tr>
<tr>
<td>3.7</td>
<td>A user with a HMD and sensors</td>
<td>27</td>
</tr>
<tr>
<td>3.8</td>
<td>A user with a HMD, headphones and other mechanisms</td>
<td>29</td>
</tr>
<tr>
<td>3.9</td>
<td>Haptic feedback enhances perceived realism</td>
<td>29</td>
</tr>
<tr>
<td>3.10</td>
<td>File-system hierarchy used by Brewste</td>
<td>31</td>
</tr>
<tr>
<td>3.11</td>
<td>Parameter Mapping Sonification (PMSon)</td>
<td>32</td>
</tr>
<tr>
<td>3.12</td>
<td>Image visually and aurally explored</td>
<td>33</td>
</tr>
<tr>
<td>3.13</td>
<td>Model-based sonification (MBS)</td>
<td>34</td>
</tr>
<tr>
<td>3.14</td>
<td>The Growing Neural Gas (GNG)</td>
<td>34</td>
</tr>
<tr>
<td>4.1</td>
<td>Virtual Environment Room</td>
<td>39</td>
</tr>
<tr>
<td>5.1</td>
<td>Experiment 1 Virtual Environment Room</td>
<td>45</td>
</tr>
<tr>
<td>5.2</td>
<td>Probability distributions</td>
<td>49</td>
</tr>
<tr>
<td>5.3</td>
<td>Probability distributions for selecting a stimulus on each of the five configurations</td>
<td>49</td>
</tr>
<tr>
<td>5.4</td>
<td>Correlation values for engagement/boredom and instantaneous excitement</td>
<td>50</td>
</tr>
<tr>
<td>5.5</td>
<td>Correlation value of the most significant EEG data increases</td>
<td>51</td>
</tr>
<tr>
<td>5.6</td>
<td>Experiment 2 Virtual Environment Scenario</td>
<td>53</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Directions of filtered noise ................................. 14
2.2 3-D Audio Spatialization Challenges and Solutions ................................. 17
2.3 Spatial Audio Models pros and cons ................................. 19

3.1 Comparison between Auditory-Only, Visual-Only and Audio-Visual VEs ................................. 28

5.1 Initial Configurations ........................................... 47
5.2 Number of found collisions in three minutes for each set ........................................... 58
5.3 Search time between collisions in three minutes for each set ................................. 59
5.4 Normalized path length in three minutes for each set ........................................... 60
5.5 Rotation degrees between collisions in three minutes for each set ................................. 61
5.6 ANOVA for three samples of the average of the normalized path length through time ................................. 61
5.7 ANOVA for three samples of the of the normalized rotations degrees through time ................................. 62
Abbreviations

BCI Brain Computer Interface
CEDAR Collaborative Engineering Design and Review
CI Confidence Interval
EEG Electroencephalography
fMRI Functional Magnetic Resonance Imaging
GNG Growing Neural Gas
HMD Head-Mounted Display
HRTF Head-Related Transfer Function
ILD Interaural Level Difference
ITD Interaural Time Difference
IVE Immersive Virtual Environment
MAA Minimum Audible Angle
MAMA Minimal Audible Movement Angle
MBS Model Based Sonification
ORE Outdoor Real Environment
OVE Outdoor Virtual Environment
PI Place Illusion
PMSon Parameter Mapping Sonification
Psi Plausibility Illusion
STD Standard Deviation
VBAP Vector Base Amplitude Panning
VE Virtual Environment
WFS Wave Field Synthesis
Contents

1.1 Motivation .................................................... 2
1.2 Problem Statement ......................................... 3
1.3 Contributions of our work .................................. 3
1.4 Publications .................................................. 5
1.5 Thesis Outline ............................................... 5
While designing Virtual Environments (VEs), researchers focus most of their efforts in providing the user with an high quality graphics visualization system. However, we believe that something is being left behind, namely sound. Sound is essentially used to reproduce the voice of a collocated user or some 2-Dimensional effects. For instance, if the user is facing a large-scale display, he/she can’t be aware of everything that is happening just with the eyes. Outside theirs field of view, might be some valuable information that if missed, will denigrate the sense of immersion. With the right amount of resources, 3-Dimensional sound systems can solve these kind of problems, because they are an essential complement for visualization techniques. If we are used to 3-Dimensional sounds in every day life, why not have 3-D sound systems to enhance presence in VEs?

1.1 Motivation

Virtual Environments can trace their roots to the 1860’s, when the Baldasse Peruzzi’s piece titled Salla delle Prospettive\footnote{http://www.wga.hu/html_m/p/peruzzi/farnesi4.html, accessed on 26/11/2011} was repainted and 360-degree art through panoramic murals began to appear. The first full immersion virtual reality machine was created by Morton L. Heiling titled “Sensorama Machine” (1962)\footnote{http://www.mortonheilig.com/, accessed on 26/11/2011}, it was able to display stereoscopic 3-D images in a wide-angle view, supply stereo-sound, provide body tilting, simulate wind and emanate aromas. This avant-garde invention started a 3-D Virtual Reality era.

Since that time, Immersive Virtual Environments (IVEs) started to appear in several different forms: Head-Mounted Displays (HMDs)\footnote{?}, CAVE-like systems\cite{1,2}, Augmented Reality systems\cite{3} or the Allosphere\cite{4} to name a few. Despite their limitations, all these systems already proved that in the presence of more visual information, users feel more immersed. But they all report some identical reversals: one is that it is much harder to capture users’ awareness for the most important events, and second is that the users’ expectations are much higher, because if presented with a visualization system much closer to the reality, they expect to have other senses (hearing, touch, smell and taste).

Sound systems have also suffered some technological evolution. They have been given less importance than visual systems, but spatial audio models like: Wave Field Synthesis (WFS)\cite{5}, Ambisonics\cite{6,7}, Vector Base Amplitude Panning (VBAP)\cite{8} or Binaural, can reproduce 3-D sounds and therefore create a more “realistic” environment.

In the real world, when we listen to someone speaking, we intrinsically associate the sound source to the movements of the mouth, a physical sound source. Our brain is trained to “connect” what we see to what we hear\cite{9}. The same happens in a VE, but in this case headphones or loudspeakers are responsible to position the sound sources relatively close to the visual sources (virtual sound sources), so that humans perceive sound as coming from the exact same place as the images\cite{10}.
The biggest challenge of a 3-D sound system is to accurately position the virtual sound sources, so that users don’t perceive them as coming from elsewhere other than supposed. When using headphones the listener knows that he/she is “entering” in a virtual space, where the sound is coming from the physical sound sources attached to the ears. This is also true when using loudspeakers. Even though that a virtual sound source can be perceived as coming from elsewhere rather than the loudspeakers, the human will always associate the sound to two sources: the image and the loudspeakers. We, as intelligent beings, connect what we hear to the loudspeakers surrounding us. We unconsciously, don’t believe that the sound source is really coming from the image facing us, because what is virtual is not real.

1.2 Problem Statement

We identified that VEs are not giving enough importance to 3-D sound systems. With the evolution of Virtual Environments the visual information is becoming richer, but since humans have a very limited field of view, a lot of important data is unnoticed. An appropriate 3-D sound system can help solving these kind of issues, it is a complement for visualization techniques. Hearing is an always open channel and independent of the field of view, a VE that can position sounds all around the user can enhance presence and help people navigate [11]. Also 3-D sounds are something that users expect, especially when presented with a very rich visualization.

1.3 Contributions of our work

To encourage the use of a 3-D sound system when developing an IVE, we performed three different experiments. Our focus was the analyses of 3-D sound to improve navigation and enhance presence. We also measured the subjective perception of sound when the loudspeakers were hidden or visible.

To evaluate 3-D sounds and their influence in enhancing presence, we framed our experiment in 6DSpaces Science Center[3]. Our Outdoor Virtual Environment (OVE) was a simulation of a journey through the sea and the forest where vision, smell, haptics and hearing stimuli were activated. We used a conjunction of two methods to measure presence: the one proposed by Slater [12] and the Emotiv EPOC EEG headset[4]. We found that the tactile perception of wind is the most important stimulus in OVEs and that a 3-D sound system is more important to enhance presence than a video-projection. We also discovered that Slater’s and EEG methods used together, help reducing some of the subjectivity inherent to Presence as an attribute.

---

[3]6DSpaces has the goal of improving the sense of immersion in VEs, targeted at museums, science centers and theme parks. The rationale of the approach is to exploit new advances in multimodal user interfaces, scent projectors and haptics in order to advance the state of the art in this field.

The navigation experiment was framed in project CEDAR (Collaborative Engineering Design and Review), an hands-free Avatar-based Teleconference Environment with a large-scale display and a 3-D sound system, used to help collaborative engineering design and review which focus its attention in the oil industry (Fig. 1.1). The scenario used represented a real need for PetroBras. Through a comparison of three setups: video projection-only, video projection and 2-D sound, video projection and 3-D sound, we proved that audio cues make navigation effortless and improve sense of orientation. We found that with 3-D sound, users chose the shortest paths; without audio cues, users lost orientation more often and with 3-D sound users rotate more than with 2-D. We also found that 3-D sound approaches real life behaviours to the Virtual Environment, which off course enhances presence. The worst results were in the video projection-only, which led us say that sounds, especially 3-D ones, are essential for tasks that need a rapid judgement and awareness of what’s surrounding.

To prove that it is worth hiding the loudspeakers to improve the subjective perception of sound in a VE, we performed a very simple task. The user needed to decide if a shifted sound from a talking-head, was or was not in the head’s mouth. Each user did the experiment with the loudspeakers hidden and visible. In the setup where the loudspeakers were hidden, users tolerated greatest distances between audio and visual stimuli (a greatest ventriloquism effect). This simple, cheap and effortless technique is very useful because it enables greatest positioning errors between the image and the sound: it doesn’t exist any sound system that can position virtual sound sources without errors.

Figure 1.1: CEDAR (Collaborative Engineering Design And Review) Virtual Environment.

1.4 Publications

One long paper will be submitted in Mel Slater’s Presence Journal. This paper is focused on Experiment 1 - Multi-Sensorial Stimulation and Presence in Outdoor VEs.

Two more papers will be submitted to conferences related to HCI and Virtual Reality, namely a long paper of experiment 2 (Navigation Task) and a short paper of experiment 3 (Machine-Less Ventriloquism).

1.5 Thesis Outline

The remainder of this dissertation is organized as follows: In section two it is presented a deep analysis on the fundamentals of sound, psychoacoustics, spatial audio perception, 3-D audio spatialization challenges and spatial audio models. In the related work (section three) we present some experiments that evaluated presence and the importance of audio/visual cues or other physical modalities in Virtual Environments. In the end of that section we present the state of the art. On section four we describe the Virtual Environment used for all the three experiments, while in the section five we look at each experiment alone: participants, VE specification, procedures, experimental design, results and discussion. In the end of section five (Experiments) we make a summary of all the three experiments. Finally we draw our conclusions and talk about the future work on sections six and seven respectively.
Contents

2.1 Fundamentals of Sound .................................................. 7
2.2 Psychoacoustics .......................................................... 9
2.3 Spatial Audio Perception ............................................... 11
2.4 3-D Audio Spatialization Challenges ............................... 17
2.5 Spatial Audio Models .................................................... 18
In this section, some fundamentals of sound are provided to better understand the sound waves behaviours. Secondly it is made an introduction to the human psychological and physiological responses to sound (also named psychoacoustics) and are presented several different factors that influence spatial audio perception. Finally some techniques to defeat 3-Dimensional audio spatialization challenges are arise and different spatial audio models are described.

2.1 Fundamentals of Sound

“It is impossible to investigate and evaluate surround systems objectively without first knowing how our brain processes sound, as it is this perceptual system that we are aiming to fool” [13].

Sound is created when it exists a wave motion in air or other elastic media (stimulus) and it is perceived when it exists an excitation of the hearing mechanism (sensation). We refer to a stimulus when we are talking about the physical characteristics of sound and to a sensation when referring to the psycho-physical ones.

For example Frequency is a characteristic of periodic waves measured in hertz (cycles per second), readily observable in a cathode-ray or countable by a frequency counter, while pitch may be quantified as a frequency, but is not a purely objective physical property. Pitch is a subjective psycho-physical attribute of sound. The same situation exists between intensity and loudness or between waveform (or spectrum) and perceived quality (or timbre).

2.1.1 Propagation of Sound

When an air particle is displaced from its original position, elastic forces of the air tend to restore it to its original position. As the displaced particle has inertia, it disturbs the resting position, bringing into play elastic forces in the opposite direction, and so on [14].

The intensity of sound decreases as the distance to the source is increased. In free space, sound from a point source is propagated uniformly in all directions. The same sound power flows out through the space, but the areas increase as the square of the radius, r. This means that the sound power per unit area (intensity) decreases as the square of the radius. Doubling the distance reduces the intensity to one-fourth the initial value, tripling the distance yield one-ninth and so one.

2.1.2 Complex Waves

Joseph Fourier (1768-1830) said that, theoretically any complex periodic wave can be synthesized from sine waves of different frequencies, different amplitudes and different relationships.

Harmonics: When three different frequencies are added, \( f_1 = 10 \text{ Hz}, \ f_2 = 20 \text{ Hz} \) and \( f_3 = 30 \text{ Hz} \), it is created a complex wave formed by a fundamental \( (f1) \), which is the lowest frequency, a second
harmonic (f2), which is the one with twice the frequency of the fundamental, and a third harmonic (f3), three times the fundamental.

**Phase:** When three different frequencies are added at the exact same time, it is created an in-phase complex wave (Fig. 2.1). In some cases, the time relationship between harmonics or between harmonics and the fundamental is not in phase (Fig. 2.2), producing a completely different wave shape. Curiously, even though the shape of the wave is dramatically changed by shifting the time relationships of the components, the ear is relatively insensitive to such changes. In other words, waves E of Fig. 2.1 and Fig. 2.2 would sound very much alike to us.

![Figure 2.1: In-phase complex wave](Taken from [14])

![Figure 2.2: Not in-phase complex wave](Taken from [14])

### 2.1.3 Sound Levels and the Decibel

The decibel is the fundamental unity used in sound. Levels in decibels make it easy to handle the extremely wide range of sensitivity in human hearing. A level in decibels is ten times the logarithm to the base 10 of the ratio of two power like quantities.

A **power level** of an acoustic power, electric power or any other kind of power $W_1$ can be expressed in terms of a reference power $W_2$ as follows:

$$ L_1 = 10 \log_{10} \left( \frac{W_1}{W_2} \right) \text{ decibels} \quad (2.1) $$

When levels other than power need to be expressed in decibels the follow equation is used:

$$ L_p = 10 \log_{10} \left( \frac{P_1}{P_2} \right)^2 = 20 \log_{10} \left( \frac{P_1}{P_2} \right) \text{ decibels} \quad (2.2) $$

**Sound pressure** is usually the most accessible parameter to measure in acoustics, even as voltage is for electronic circuits. For this reason, the Equation 2 form is more often encountered in day-to-day technical work [14].

Some standard reference sound pressure for $P_2$ is needed, and for sound in air it is $20 \, \mu\text{Pa}$ (micropascal). This is a very minute sound pressure and corresponds closely to the threshold of human hearing.
2.2 Psychoacoustics

“Psychoacoustics is an inclusive term embracing the physical structure of the ear, the sound pathways, the perception of sound, and their interrelationships” [14].

2.2.1 Ear Anatomy

The three principal parts of the human auditory system are the outer ear, the middle ear and the inner ear. Together they allow humans to hear sounds in a range of 20 to 20000 Hz [14].

The outer ear is composed of the pinna and the auditory canal (or auditory meatus). The auditory canal is terminated by the tympani membrane (or the eardrum). The middle ear is an air-filled cavity spanned by three tiny bones (or ossicles). The ossicles form a mechanical, lever-action connection between the air-actuated eardrum and the fluid-filled cochlea of the inner ear. The inner ear is terminated in the auditory nerve, which sends impulses to the brain [14] (Fig. 2.3).

The outer ear increases sound pressure at the eardrum through the pinna, for the important speech frequencies (2000 to 3000 Hz), in about 5 dB. Pinna also offers a certain differentiation of sounds from the front as compared to sounds from the rear and performs a very crucial function in imprinting directional information on all sounds picked up by the ear (see “Cone of Confusion” in [9]). The auditory canal produces a resonance effect that increases more or less the loudness depending on the size of the frequency.

The middle ear solves the problem of matching the sound in air to sound in fluid of the inner ear.

The inner ear deals with the standing waves being set up on the basilar membrane. Low frequency sounds result in maximum amplitude near the distant end of the basilar membrane and high-frequency sounds produces peaks near the oval window [14].
2.2.2 Loudness Vs Frequency

Loudness is a psycho-physical term. Therefore, in order to quantify loudness some work was done by Robinson and Dadson [15], and the family of equal-loudness contours has been adopted as an international standard (I.S.O 226). Each equal-loudness contour is identified by its value at 1 kHz, and the term loudness level in phons is thus defined. Observing Fig. 2.4 we can say that the ear is less sensitive to bass notes than mid-band notes at low levels.

2.2.3 Area of Audibility

The auditory area of the human ear is bounded by two threshold curves (see Fig. 2.5), (A) the threshold of hearing delineating the lowest level sounds the ear can detect, and (B) the threshold of feeling at the upper extreme. All of our auditory experiences occur within this area [14].

2.2.4 Loudness Vs Sound-Pressure Level

In order to have a better perception of human reaction to loudness of sound we need some sort of subjective unit of loudness. Many experiments conducted with hundreds of subjects and many types of sounds have yielded a consensus that for a 10-dB increase in sound-pressure level, the average person reports that loudness is doubled. Some studies report that this should be 6 dB, others 10 dB, so work on the problem continues [14].

2.2.5 Audibility of Loudness Changes

Detecting differences in intensity varies somewhat with frequency and also with sound-pressure level. At 1 kHz, for very low levels, a 3 dB change is the least detectable by the ear, but at high levels the ear can detect a 0.25 dB change. A frequency near 35 Hz at a very low level needs a 9 dB change to be detectable [14].

The duration of a sound burst influences the perceived loudness. Short bursts (less than 100 msec) need an higher sound-pressure level to sound as loud as long bursts or steady tones.
2.3 Spatial Audio Perception

“From an evolutionary standpoint, the horizontal placement of the ears maximizes differences for sound events occurring around the listener, rather than from below or above, enabling hearing of terrain-based sources outside the visual field of view” [9].

Although our monaural hearing capability is not to be underestimated, many research experiments like the one from Rayleigh [16], concluded that the position of the two ear mechanisms in the head is relevant in the localization of the sound sources.

2.3.1 Interaural Time Difference and Interaural Level Difference

In order to describe the frequency-dependent cues of *interaural time differences* (ITD) and *interaural level differences* (ILD), experiments are constrained according to a lateralization paradigm [9]. The word “laterized” indicate a special case of localization, where:

- The spatial percept is heard inside the head, mostly along the interaural axis between the ears;
- The means of producing the percept involves manipulation of ITD or ILD over headphones.

Lateralization mimics the interaural differences of a natural spatial hearing [9]. It is represented by a listener with a perfectly round head and no outer ears, placed at a certain distance in an anechoic chamber from a sound source at eye level (see Fig. 2.6).

**Interaural Time Difference** is the sound path length difference between the two ears. With the source at position A (at 0 degrees azimuth), the sound path lengths are equal and no interaural time difference is detected. At position B (60 degrees azimuth to the right of the listener) the paths are not equal, therefore the sound source wavefront will arrive first to the right ear and an ITD is detected.

If we take the speed of sound as 342 msec$^{-1}$ and consider an average head diameter of 16 cm we have a maximum time difference between the sound ears of 0.74 msec.
Maximum path difference between the left and right ear:

\[ \pi r = \pi \times 0.08 = 0.2513 \text{ m} \]  
(2.3)

Maximum time difference between the sound ears:

\[ \frac{\pi r}{\text{Speed of Sound}} = \frac{0.2513}{342} = 0.74 \text{ msec} \]  
(2.4)

**Interaural Level Difference** is the sound intensity difference between the two ears. With the source at position A no interaural level difference is detected. At position B the sound source will yield a significant ILD cue, but only for those waveform components that are smaller than the diameter of the head, i.e., for frequencies greater than about 1.5 kHz. Frequencies above this threshold will be attenuated at the left ear because the head acts as an obstacle, creating a “shadow” effect at the opposite side. The higher the frequency the higher is the attenuation.

If we take the speed of sound as 342 msec\(^{-1}\) and consider an average head diameter of 16 cm we have a minimum frequency that can be attenuated by the head of 2.1 kHz.

**Minimum frequency** that can be attenuated by the head:

\[ f = \left( \frac{1}{0.16} \right) \times 342 = 2.1 \text{ kHz} \]  
(2.5)

Until 1960's, researches of spatial hearing found results that suggested that the ITD and ILD cues operated over exclusive frequency domains (e.g. [17]). The theory was that one neurological component for spatial hearing responded to ITD cues only for low frequencies (below 1,5 kHz), and the other component evaluated ILD cues for the high frequencies (above 1,5 kHz). Later research revealed that timing information can be used for higher frequencies as well. This is because frequency phase differences are detected by the ear [18–20].

According to the maximum displacement to one side of the head on the interaural axis, and depending on the type of stimuli (short bursts, long bursts or steady tones), the effective range of interaural timing differences is approximately between 0.005 to 1.5 msec, and the effective range of interaural intensity differences is between approximately 1 do 10 dB [9].

Even though that lateralized sound sources are supposed to be heard only along the interaural axis, there is also a vertical component experienced by some listeners. As interaural differences are minimized, the source can travel upward [21]. In addition, the same lateralized sound source is heard sometimes behind, and sometimes in front.

### 2.3.2 The Precedence Effect

The ear and brain, have the ability to gather all reflections arriving within about 50 msec after the direct sound and combine them to give the impression that all sound is from the direction of the original sound. The sound energy integrated over this period also gives an impression of added loudness.
Haas [22], found that in the 5 to 35 msec delay range the sound from the delayed loudspeaker has to be increased more than 10 dB over the direct one before it sounds like an echo. This is the precedence effect and it explains an important inhibitory mechanism of the auditory system that allows one to localize sounds in the presence of reverberation.

2.3.3 Head Movement and Source Movement Cues

When we hear a sound we wish to localize, we move our head in order to minimize the interaural differences, using our head as a sort of a “pointer” [9]. Therefore, an important source of information for localization of sound sources are the head motion cues. Several studies concluded that allowing a listener to move his head can improve localization ability [23, 24].

With stereo loudspeakers and without head-tracking, a distortion of spatial imagery will occur when the head is turned to face a virtual sound source, since the processing to create the illusion depends on a known orientation of the listener [9].

A moving source causes dynamic changes for a fixed head [9]. Under controlled experimental conditions, the minimum audible movement angle (MAMA) can range up to 3 degrees while the minimum audible angle (MAA) for fixed sources is around 1 degree [25]. The MAMAs can increase as a function of movement velocity, location of movement and sound source type.

2.3.4 Head-Related Transfer Function

Our head and other features of the body like the outer ear, shoulder and torso contribute to a spectral filtering before the sound reaches the ear drum. This frequency filtering is termed as the head-related transfer function (HRTF).

The use of HRTF spectral shaping is typically featured as the key component of a 3-D sound system, from either direct measurement or modelling [9]. In theory, the most accurate means to produce a spatial sound cue is to transform the spectrum of a sound source at the ear drum as closely as possible to the way it would be transformed under normal spatial hearing conditions.

The complex shaping of the pinnae turns it as the most important HRTF cue in detecting the position of a sound source. Its asymmetry causes the spectral modification to be affected differently as a function of a sound source position. The asymmetrical and complex construction of the outer ears causes a unique set of microtime delays, resonances and diffractions that collectively translate into a unique HRTF for each sound source position.

**The main role of HRTF cues** from a psychoacoustic standpoint is thought to be the disambiguation of front from back, for sound sources on the cone of confusion and as an elevation cue for disambiguating up from down [9].

**Non individualized (generalized) HRTFs** can be derived from individualized HRTFs averages. This causes a significant decrease in azimuth localization accuracy [25].
Table 2.1: A broadband noise that is filtered to contain primarily the frequencies shown in the second and third columns, will be heard from the direction indicated in the first column.

<table>
<thead>
<tr>
<th>Perceived location</th>
<th>Center Frequency (kHz)</th>
<th>Bandwidth (kHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>overhead</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>forward</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>forward</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>rear</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rear</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

Spectral cues can influence the perception of direction independent of the location of a sound source [9]. Each sound source contains broadband spectral energy in order to be affected by the particular HRTF cue (see Table 2.1).

Elevation perception is considered to be very dependent of the pinnae’s directional effects. Several studies have shown that without the pinna, vertical localization is diminished [27, 28] and others demonstrated that to perceive an overhead sound it is required significant energy above 7 kHz [29]. Nevertheless, Noble [30] proved that around 25 percent of the population exhibits little or no sensitivity to elevation judgements based strictly on spectral cues, i.e., without cognitive cues such as familiarity.

2.3.5 Distance Perception

When studying auditory distance perception we need to make a separation between absolute and relative distance. Absolute distance perception refers to a listener’s ability to estimate the distance of a sound source upon initial exposure, without benefit of the cognitive familiarity. Relative distance perception includes the benefits gained from listening to the source at different distances over time, perhaps within an environmental context [9].

The loudness of a sound source is the primary distance cue used by a listener [9] and it plays a more important role with unfamiliar than with familiar sounds. Exposures to a particular sound source at different distances allows an integration of multiple cues over time for distance perception, but without this exposure, loudness is the only cue available to identify sound sources distances.

Loudness increments can only operate effectively as distance cues under conditions where other determining cues, like reverberation, are not present. In a reverberant context, the change in the proportion of reflected to direct energy functions is a stronger cue for distance than loudness scaling [9].

Expectation and familiarity cues are used when a listener tries to estimate the apparent distance of the virtual sound image according to a correlated visual object. This is called the ventriloquism effect and because of it, distances cues can be modified as a function of expectation or familiarity with the sound source. Any reasonable implementation of distance cues into a 3-D sound system will probably require some cognitive associations for a given sound source. For example, it is easier...
to simulate a voice whispering near the listener head than from ten meters away (normally people whispers right near the hear of the listener).

The spectral cue also takes an important role to simulate distance. A sound source’s wavefront is usually planar by the time it reaches the listener’s ears, but a closely proximate sound source will have a curved wavefront. This results in an added emphasis to lower versus higher frequencies of the sound source, also known as tone “darkening”. According to Békésy [21], for a sound source placed less than 1 meter away, an emphasis of low frequency sound energy, relative to high frequency energy, should indicate an approach of the sound.

In everyday listening we are more used to perceive an high-frequency damping as the distance increases, due to air humidity and temperature. So, and according to Butler [31], whom found that a low-pass stimuli in the headphones was consistently perceived to be farther away than a high-pass stimuli, in the absence of other cues, a low-frequency emphasis applied to a stimulus placed more than 1 meter away, would be interpreted as “more distant” compared to an untreated stimulus.

2.3.6 Reverberation

The wavefront that reaches the ears first by a linear path, without having bounced off a surrounding surface is called the direct sound. The reflected sound (or reverberation) is the energy of a sound source that reaches a listener indirectly, by reflecting from surface within the surrounding space occupied by the source and the listener. A particular reflection within a reverberant field is usually categorized as an early reflection or as late reflection, depending on time of arrival to the listener. Early reflections are those with a significant amplitude above the noise floor and reach the receiver within a period around 1 to 80 msec. The late reflections result from many subsequent reflections from surface to surface of the environmental context and contains less energy overall.

The perception of the environmental context is very dependent of the physical parameters of reverberation: reverberation time and level; absorptivity of the reflective surfaces; and the complexity of the shape of the enclosure. For example, a large reverberation time is usually associated with a large enclosure and a sound with no reverberation is associated with an anechoic chamber. The reverberation gives clues for categorizing sound sources location according to previous knowledge of the environment.

Reverberation improves distance perception but have negative effects on azimuth and elevation estimates [9], as can be seen in Fig. 2.7. Early reflections coming from the side degrade localization accuracy [32], this is because the precedence effect only partially suppresses the effects of reflected energy. The source width expansion, due to reverberation, is also a problem since there is a larger area on which to consistently place a center of gravity for azimuth and elevation judgements.
2.3.7 The Ear as an Analyser

Our ear can separate components arriving from many different wave systems but the eye cannot. When listening to a symphony orchestra we have the ability to focus our attention in the violins or in the percussion area even tough they are playing at the same time. This ability is a very powerful ear/brain combination of the human.

Humans can detect objects in the environment by sensing their echoes. Techniques such as tapping a cane, lightly stomp the foot or making clicking noises with the mouth, helps trained people to accurately identify the location and size of a nearby object. People with visual impairment use this technique to navigate.

2.3.8 Other Auditory illusions

There are several auditory effects that can effectively “fool” our brain. Sometimes these illusions aren’t premeditated and they can ruin an auditory experience, but most of the times they are consciously used because they can reduce the audio processing requirements and provide a better auditory experience.

The McGurk effect is a ventriloquism effect, this is, it happens when both vision and hearing are used together. Specifically the McGurk effect happens when vision and hearing are used together to percept speech [33]. A syllable can be completely misunderstood when the audio and video are mismatched suggesting that speech perception is multimodal.

The Octave illusion first discovered by Deutsch in 1973, is an effect that happens when two sequences of two notes, high to low and low to high, spaced an octave apart are played together in

---

Figure 2.7: Difference in localization accuracy for HRTF-processed speech between reverberant and anechoic stimuli. (Taken from [9])

---

separate stereo channels over headphones. These tones are two different sine waves of constant amplitude but they are perceived as a continuous two-tone chord, with the ear of input of each component switching repeatedly.

The Doppler Effect \[\text{3}\] is the change in frequency of a wave when a moving sound source passes by the observer. Therefore to simulate a moving sound source, it is necessary to apply an high-pass filter when it is approaching and a low-pass filter when moving away.

2.4 3-D Audio Spatialization Challenges

The perception of spatialized sound sources have some challenges. Humans have troubles in localizing sounds in height and in their back. In addition there are some auditory illusions that can influence the perception too. The existent sound systems aren’t perfect and that is too a challenge.

In order to simulate good 3-D sounds it is necessary to have a deep knowledge on psychoacoustics and in techniques to “fool” the auditory perception. Table 2.2 identifies some 3-D audio spatialization challenges and techniques to solve them.

Table 2.2: 3-D Audio Spatialization Challenges and Solutions

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disambiguation of front from back and up from down sounds</td>
<td>Allow the user to move is head; Use HRTFs; Spectral filtering according to psychoacoustics (Sounds are perceived at the top with higher frequencies, below with lower frequencies)</td>
</tr>
<tr>
<td>Distance Perception</td>
<td>Use reverberation; Make use of the Ventriloquism Effect; Use familiar sounds (when someone shouts, normally he/she is very far away); Spectral filtering according to psychoacoustics (sounds are perceived more distant if they have lower-frequencies, due to high-frequency damping)</td>
</tr>
<tr>
<td>Simulate Rooms</td>
<td>Use Reverberation in accordance with the visual environment</td>
</tr>
<tr>
<td>Reverberation improves distance perception but have negative effects on azimuth and elevation estimates</td>
<td>Make use of the Ventriloquism Effect</td>
</tr>
</tbody>
</table>

2.5 Spatial Audio Models

Spatial Audio Models are those that can “place” sound sources in any location around the listener. It is important not to confuse surround with 3-Dimensional sound systems. Surround sound systems (5.1, 7.1) generate sound around the audience, but only in 2-D. Sound is just differentiated between left and right, and between front and rear. 3-Dimensional sound systems include the height layer and therefore they are the only that can “place” a sound source anywhere around the listener.

Sound sources can be real or virtual. A real sound source is physical, like a loudspeaker. A virtual sound source represents an auditory object that is perceived in a location that does not correspond to any physical sound source.

3-Dimensional sounds can be reproduced through headphones or loudspeakers, and there exist four different spatial audio models to do it (see Table 2.3).

2.5.1 Binaural

Is an headphone spatial audio model. The sound signal is therefore reproduced directly to the ears. It is possible to manipulate the sound signal with ILDs, ITDs and HRTFs.

In VEs, the user typically likes to move around, so a real-time head-tracking and real-time processing of HRTFs are needed for an accurate reproduction.

2.5.2 Ambisonics

Is a recording and reproduction technique, invented by Gerzon with the help of his professor Fellgett [6, 7]. A sound signal is applied to all loudspeakers with different gains, in such a way as to fool the ears of listeners into thinking that they are hearing the original sounds correctly located. This can take place over a 360 degree horizontal only soundstage (pantophonic systems) or over the full sphere (periphonic systems/3-D sound systems).

Reproduction requires four or more loudspeakers depending on whether it is pantophonic or periphonic, size of area etc.. Practical minimums are four for horizontal only, eight if height is required as well.

Ambisonic decoders work best with a regular and symmetrical loudspeaker configuration. A first order Ambisonics system with 8 loudspeakers positioned in a perfect cube format, is capable of reproducing pretty good 3-D sound sources.

2.5.3 Vector Base Amplitude Panning (VBAP)

Is independent from the loudspeakers setup. It is a method used to calculate gain factors for pair-wise or triplet-wise amplitude panning [8]. VBAP enables arbitrary positioning of loudspeakers and the virtual source is reproduced using the three closest loudspeakers at a time. According to Gröhn [34], if the listener moves away from the supposed listening position, then the direction accuracy does not decrease with VBAP as much as it does with systems that apply a signal to all loudspeakers. VBAP can’t handle distance cues [35].
2.5.4 Wave Field Synthesis (WFS)

Is theoretically the perfect solution to reconstruct a whole sound field, but it is impractical in most situations. Sounds can only be played accurately if the loudspeakers are at a maximum distance of a half wavelength from each other [5]. Higher the frequency and higher is the number of loudspeakers needed to reproduce the sound sources. This means that to produce an accurate three dimensional wave field, hundreds of loudspeakers would be needed, which is not suitable for 3-D sound reproduction.

<table>
<thead>
<tr>
<th></th>
<th>Binaural</th>
<th>Ambisonics</th>
<th>VBAP</th>
<th>WFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener sweet spot[^1]</td>
<td>n/a</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Dependent from loudspeakers disposition</td>
<td>n/a</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hands-Free</td>
<td>n/a</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8 Loudspeakers or Less</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Requires Head-Tracking</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorporates Distance Cues</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[^1]Sweet spot is the area where the user should be, to perceive the sound sources exact locations.
Related Work

Contents

3.1 Virtual Environments .................................................. 21
3.2 Sonification Techniques ................................................ 30
3.3 The state of the art ....................................................... 37
In this section it is made a comparison between Visual-Only, Auditory-Only, Audio-Visual and other Virtual Environments that simulate more than the two main physical modalities (smell, touch or taste). The focus will be on the advantages of each one to enhance presence and awareness and their performance on collaborative, assembly, navigation and localization tasks. After that, some works that used sonification techniques to convey information or perceptualize data are presented.

3.1 Virtual Environments

3.1.1 Presence and Awareness in Virtual Environments

Vision, hearing, touch, smell and taste are what Immersive Virtual Environments (IVEs) try to computer-simulate so that the users can report the sense of having been somewhere other than where they really were at the time. Evaluating if the user felt a physical presence in the simulated environment and if he kept a continuous awareness of others presence, are two important factors to characterize a VE [36].

In a Virtual Environment like CEDAR, the user needs to feel that he/she is physically present and needs to keep a continuous awareness of others presence. If project CEDAR has a large-scale display, how can the user maintain awareness of people and objects when they fall outside his/hers field of view? To this case and others, 3-D sound spatialization takes a very important role in enhancing presence and awareness.

If developing Outdoor Virtual Environments, apart from a great visualization display and 3-D sounds, for sure that the other physical modalities are essential to enhance presence. Even for some VEs that only try to simulate an office, a conjunction of the other physical modalities can be helpful.

3.1.2 How to measure Presence in Virtual Environments

The feeling of presence in VEs has never benefited from a combined and generally established measurement approach [38]. Therefore, experiments are typically performed combining different measuring methods such as questionnaires, behavioural and physiological responses or comparison, and others. All of these approaches have their own set of problems, but when used together, they lead to clearer results. Depending on the type of VE, a correct mix of different methods seems to be the way to give us a more reliable definition of Presence.

Measuring Presence through Questionnaires

Presence is a subjective concept and since the objective is to understand if the person had the feeling of being in a place different from the one that he/she really is, questionnaires seem to be an obvious approach for measuring presence. Experiments evaluated by questionnaires are always

---

1Sonification: a form of auditory display, is the use of non-speech audio to convey information or perceptualize data.
based on the same paradigm: manipulate some particular aspects of the VE and elaborate questions related to that changes. The responses are typically based on Likert scales regarding how much the participants felt themselves to “be in another place”, e.g., scoring 1 for “not at all” and 7 for “very much” [12].

There are several examples of questionnaires [37–40], but they all seem to be insufficient when used alone. Amongst many problems, there are the methodological problems of analysing subjective data transformed in interval or ratio data [41].

**Measuring Presence through Behavioural and Physiological responses**

In a good virtual simulation, a person should manifest behavioural and physiological responses similar to real-world situations. To explore these characteristics, the experiments are normally designed to activate clearly measurable physiological or behavioural responses and then compare them under different experimental conditions.

The metric most widely used is based on the physiological characteristics of stress, mostly because it is easy to identify (through skin conductance and heart-rate variability). A good example is the experiment of a stress response to a visual cliff performed by Meehan et al. [42]. Nevertheless, not every VEs can have a situation to cause stress or other clearly measurable physiological responses [12].

Another possibility to measure presence is through the analysis of electroencephalography (EEG) data. Inexpensive systems like the *Emotiv EPOC EEG* headset, can detect different emotions off-the-box (engagement/boredom, frustration, meditation and instantaneous and long term excitement) and several facial expressions (blink, smile, laugh, smirk, etc.). Most of the experiments that use EEG are related to brain-computer interfaces (BCI) [43], but there are some studies that make use of these systems to measure users’ level of presence [44], or adapt components of the VE in accordance to emotions [45].

**Measuring Presence through Comparison**

If we define presence as the extent to which participants respond realistically to virtual events and situations, rather than their sense of “being in another place”, other methods can be used to measure it. For example if presence is higher when people behave as they would in the real world, then the extent to which this occurs is measurable in principle.

Slater and colleagues [12] define the notion of immersion in VEs: One system (A) is said to be more Immersive than another (B) if A can be used to simulate an application as if it were running on B. They demonstrated how this concept could be used as the basis for a psychophysics of presence in VEs, the sensation of being in the place depicted by the virtual environment (Place Illusion, PI), and the illusion that events occurring in the virtual environment are real (Plausibility Illusion, Psi).
Using the analogy of colour matching experiments to reach international standards on the representation of colour, Slater et al. proposed a method to measure presence by comparing the users’ Place Illusion (PI) and Plausibility (Psi) in a VE with variations on its components [12, 46]. For example, if the user feels the same PI in a VE-A with, and in a VE-B without a full body-tracked avatar, then there is equivalence between PI experienced in VE-A and VE-B.

### 3.1.3 Visual-Only Virtual Environments

The vision is known as the most important modality to navigate and localize objects in a VE, because it provides very detailed information of what we look at.

Head-Mounted Displays (HMDs) [? ] and CAVE-like systems [1 2] are very concerned in the visualization of 3-D models and in the navigation on virtual spaces.

Augmented Reality [3] is most concerned at bringing 3-D virtual images into the real world, or at substituting real images by virtual ones.

Allosphere [4] is the most powerful high-resolution visual Immersive environment to date (Fig. 3.1) and was considered essential to the next step in identifying quantum computers patterns.²

![Figure 3.1: Allosphere, the most powerful high-resolution visual Immersive environment to date. (Taken from [11])](image)

All these systems are very good in enhancing presence and awareness solely with photo-realistic images and visual cues, but having a 360 degree image reproduction is simply not enough.

The Allobrain [11] project is an interactive, Immersive, multimodal art installation that uses functional magnetic resonance imaging (fMRI) brain data to construct a virtual world. It was installed in the Allosphere, one of the largest VEs in existence. As this very huge display produces a visual overload [47], Allobrain used artificial semi-autonomous agents to include normally non-noticed stimuli (Fig. 3.2). These agents could enhance user’s presence and awareness but only if they were

sonified by the third-order 3-D Ambisonics system. All the experience depended on sound spatialization because of information being missed due to eyes focusing elsewhere or the agent being positioned behind the user's head.

In a comparison of visual-only and auditory-only information to navigate in a CAVE-like virtual world done by Gröhn et al. [48], it was also found that the user needed auditory cues, mostly outside its field of view, to navigate towards the objective faster (Fig. 3.3).

3.1.4 Auditory-Only Virtual Environments

Even tough that vision can provide very detailed information, it is fully dependent of the user focus area. Hearing is an always-open channel and it can capture information independently from the area that the eyes are looking at. A 3-D sound spatialization is a way to “place” the sound sources anywhere in the virtual world and to enable a 3-Dimensional information retrieval.

Nevertheless, even with real sound sources, the user isn’t as accurate as in a visual-only VE, in identifying the exact position from where the information is coming from [49]. Therefore, the sense of immersion in auditory-only VEs is normally lower, because users can’t feel so physically present when they are disoriented.

Most of the research in auditory-only VEs is focused on people who cannot see or on dark spaces, where vision is reduced almost to zero. That “blindness” condition plus a 3-D spatialized sound reproduction, enhances in a better way, presence and awareness.

In a VE with a haptics and visual interface where sighted and people who cannot see, needed to collaborate in order to build an object (Fig. 3.4), Moll et al. [50] found that assembly tasks are improved with the auditory feedback of the 3-D models. For example, reproducing a “joint sound” when the object fit in the construction or a “falling sound” when the object is no longer being grabbed, can provide some independence to the movements of the visually impaired user and keep the sighted user always aware, even when concentrated on other object. This fact is reported by Absar et al. [51]
as well, who also states that the performance of a sighted user in assembly tasks can be improved with sonic cues.

Also in Moll’s et al. [50] experiment (a collaborative virtual workspace used by blind and sighted people), it was found that sound can alleviate the problems that visually impaired users experience in maintaining awareness of other’s actions, increase the sense of social presence and make the grounding process easier.

3-D sound spatialization was found to be a very efficient method to allow visually impaired individuals to travel through familiar and unfamiliar environments [52]. Lokki et al. [53] also proved that sighted people can navigate in a VE with 3-D auditory-only cues.

In an effort to simulate a soccer game for blinds, Stockman et. al [54] tried to answer some questions like, how to audio simulate the sides of the pitch, positions of the goals and players who are not currently moving. For that he had the help of the British blind soccer squad (Fig. 3.5), and found that sounds of the environment such as calls and footsteps from other players, sounds from the ball, for example when it hits the sides or ends of the pitch, instructions from team coaches and the use of echolocation are enough to orient a blind football player. The auditory-only virtual game was considered particularly useful to convey ideas about specific aspects of soccer, such as where players should be and should react in set plays such as free kicks, throw in and corner kick situations.

In what concerns to localization of sound sources, Marentakis et al. [55] did an experiment to compare timbre, loudness and orientation updated cues with the objective of enhancing pointing efficiency in deictic spatial audio displays. He found that these feedback cues can reduce effective target widths, but they also reduce the speed of interaction. So he states that, with an appropriate design, this means, depending on if we want a fast interactive VE or an optimal awareness of the sound source’s position, it is possible to enhance pointing efficiency in a spatial audio display and overcome interaction uncertainty.

---

3Grounding process refers to how the collaborating partners were shown to share the same understanding of the task, the layout of the workspace and objects in it, in other words the common ground [50].
Even tough that nowadays spatial audio models are quite realistic, the information provided by sounds-only are normally not consensual to every users. People who can see are used to trust a lot in their vision to capture surrounding information, therefore when they can only count with the auditory system, presence and awareness might not be enough to provide a good engagement in a VE like CEDAR.

### 3.1.5 Audio-Visual Virtual Environments

Presence and awareness are fully correlated with the amount of physical modalities used in a Virtual Simulation [36]. The correct combination of audio and graphics in VEs will enhance presence and awareness and therefore improve collaborative or navigation tasks. If the user is capable of using both hearing and vision, a VE must use 3-Dimensional graphics and sounds.

In Boneel's et al. [56] study it was found that a better-quality sound improves the perceived similarity of a lower-quality visual material (Fig. 3.6). As in VEs the cost of improving graphics level of detail is very expensive, a trade-off can be made between audio and graphics for the same, or even better perceived material quality.

![Figure 3.6: The object B has a better quality than object A. Sound can improve the perceived quality of object A. (Taken from [56])](image)

Gröhn et al. [48] concluded that the combination of auditory and visual cues are very dependent of each other in a navigation task. Users use auditory cues to define the approximate location of an objective and visual cues for the final approach.

In Brown's et al. [47] study, where a user needed to map the natural sounds or the visual cues to a specific location on a screen with a 34cm diagonal measure, it was proved that it is possible to reduce the visual workload by presenting part of the information to the auditory modality. This is because humans can rapidly extract more than one piece of information from a sound, and then act on the information given.

The first project of Allosphere was to develop an Immersive and interactive software simulation of nano-scaled devices and structures, with atom-level visualization implemented on the projection dome [35]. They soon realized that sound would play an important role in such simulations and
visualizations. As nanoscale simulations provide so much information, **3-D sound can bring an important temporal phenomena to user's attention:** for example, to alleviate the difficulty of finding specific molecules in the vast visual space of the Allosphere, subtle auditory clues can alert the user to the emergence or presence of a specific molecular event in a particular direction.

Gröhn [57] found that **fast-changing or transient data** that might be blurred or completely missed by visual displays, may be easily detectable in even a primitive, but well-designed auditory display.

Nordahl [58] made an experiment to study if augmenting the environment with interactive sounds could improve **motion of the subjects.** This study was performed in an HMD photo-realistic VE with an 8 speakers sound spatialization, by the means of the VBAP algorithm (Fig. 3.7). He found that adding to the graphic visualization, a combination of 3-Dimensional ambient sounds and auditory rendering of user's own motion, more precisely the footsteps, significantly enhances the users' interest in moving around the Virtual Environment.

Rath [59] identified that a continuous sonic feedback of a virtual ball that could be balanced and rolled on a tiltable track, so that it could reach a target area, unconsciously **optimize the user's control movements.**

In a VE for **collaborative and creative music-making,** Shober [60] found that users didn’t felt uncomfortable with a remote situation, as long as auditory and visual cues were present. That is because an audio-visual situation can instantly feel remarkably like a fully copresent situation, this is, the user could felt many of the same sets of a copresent situation: breath, body leans, eyebrow raises, etc..

Kaltenbrunner et al. developed Marvin [61], an awareness support agent that provides audio cues, text-to-voice and voice recognition to operate in an Internet-based, shared 3-D VE. It acts as a bot to facilitate chance encounters between members of distributed work groups. Kaltenbrunner states that if each user has many avatar proxies, they can have a foot in many camps and therefore improve collaborative tasks. In order to enable that feature they found necessary, not only to have a visual
representation of the avatar, but also an audio one, so that users could be aware of others even when working with another application on-screen.

<table>
<thead>
<tr>
<th><strong>Table 3.1:</strong> Comparison between Auditory-Only, Visual-Only and Audio-Visual VEs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Visual-Only</strong></td>
</tr>
<tr>
<td>Users can navigate in the VE</td>
</tr>
<tr>
<td>Users can perform assembly tasks</td>
</tr>
<tr>
<td>Users can collaborate</td>
</tr>
<tr>
<td>Users can identify the exact position from where the information is coming from</td>
</tr>
<tr>
<td>Detailed information of objects</td>
</tr>
<tr>
<td>Independent from the Field of View</td>
</tr>
<tr>
<td>Can be used by visually impaired</td>
</tr>
</tbody>
</table>

3.1.6 Multi-Sensory Virtual Environments

A broader range of sensory cues promotes a greater sense of presence and therefore a better experience in a Virtual Environment. A VE which uses the best graphics and sound-systems available, usually accuses latency issues, specially in interactive activities. The combination of all the physical modalities means that a large number of cross-modal effects can be exploited to improve the performance. In some cases it is possible to reduce graphics or sound-systems quality without perceptiveness, just by adding a new modality to the VE [56].

Chalmers et al. [62], performed an experiment to identify how the different modalities (noise, temperature, smell and combinations of these) affect the visual component. The users were asked to chose between two animations: one with High-Quality and the other with Low-Quality or Selective Quality, which one had the best graphics quality. In each comparison some multi-modal distractors were used: noise (pink noise) only; high temperature only; strong perfume, noise and high temperature; strong perfume and high temperature; strong perfume and noise; mild perfume and noise; mild perfume, noise and high temperature; mild perfume and high temperature.

They found that the influence of the other modalities had an impact, specially when using a combination of perfume and high temperature.

According to Dinh et al., additional sensory input can increase both the sense of presence in a virtual environment and memory for the objects in the environment [63]. In their experiment, they discovered that overall the sense of presence increased with the addition of tactile and auditory
cues and that the olfactory cues were also a trend. They also state that surprisingly, adding auditory, tactile and olfactory cues increased the user's sense of presence, increasing the level of visual fidelity did not (Fig. 3.8).

Figure 3.8: A user with a HMD, headphones and some smell and tactile mechanisms. (Taken from [63])

Hollander et al. performed a very simple experiment [64] to test whether physically biting a virtual chocolate bar in an Immersive VE, was more realistic and more fun than imagining biting a virtual chocolate bar, and whether it would lead to a higher sense of presence.

They found that users had more fun, found the event more realistic, and showed a slight tendency to feel more present in the VE when they smelled and bit a virtual chocolate bar, than when they just smelled and imagined biting it.

Nordahl in [65] states that haptic feedback provided at feet level is appreciated by the participants and enhances perceived realism. This conclusion was proved with two experiments: the first showed that a user can better walk straight without falling from a virtual plank, when haptic stimuli are provided; the second proved that when looking down a virtual canyon, audio-haptic feedback enhances the sense of realism and in some cases intensifies the experience of vertigo (Fig. 3.9).

Figure 3.9: In the left: a user performing the experiment of walking in a virtual plank. In the right: a user performing the experiment of overlooking the virtual canyon. (Taken from [65])
3.2 Sonification Techniques

“Nowadays, human-object and human-human interactions are often facilitated by computational means. Interaction designers aim at designing such interactions to make them more effective and pleasant. Sonic Interaction Design emphasizes the role of sound as a mediator of meaningful interactions” [66].

3.2.1 Auditory Icons

Firstly introduced by Gaver [67], auditory icons are everyday sounds meant to convey information about computers events by analogy with everyday events. These sounds are intended to provide information about an event or object in the interface by representing the desired data using properties of the sound’s source, rather than properties of the sound itself. Auditory icons are easy to learn and remember as they are natural and related to our everyday lives.

SonicFinder [68] is the first application to use auditory icons to convey information in a computer. It was created to extend the visual desktop metaphor into the auditory dimension, adding sounds to essential events and objects.

The selection of a file was represented by a hitting sound and depending on the type of the file (file, application, folder, disk, trash) a corresponding sound was played (wood, metal, etc.). Opening a file plays a whooshing sound, dragging a scapping sound, and so on.

Gaver felt the need to use auditory icons, because he states that providing auditory information that is consistent with visual feedback, is one way of making the model world more vivid.

A deep evaluation of existent interfaces that made use of auditory icons, performed by Absar et al. [51], found that providing auditory information consistent with visual feedback enhances presence and awareness. According to his evaluations, auditory icons can facilitate tasks in:

- **Desktop interfaces**: navigation; picture categorization; hypermedia and web interfaces.
- **Complex systems**: monitoring tasks; collaborative tasks; peripheral awareness cues, ambient sound; navigation tasks.
- **Immersive VEs**: localization and navigation tasks; assembly tasks.

Nevertheless auditory icons aren’t always the right solution to sonify actions or events. This is because not every computer interface functions and objects have real world equivalents and it may be difficult to find a correct metaphor to represent them. A bad metaphor usually leads to issues of ambiguity, loss of context and annoyance factors in users.
3.2.2 Earcons

Earcons are abstract audio messages used in the user-computer interface to provide information and feedback to the user about computer entities [70]. In contrary to auditory icons, earcons are harder to remember and learn because they have no natural link or mapping to the objects or events they represent. On the other hand they are highly structured and can easily create families and hierarchies of objects and actions, with very simple audio messages.

To design earcons it is important to have in mind that humans prefer complex sounds rather than simple tones. Brewster et al. [71] found that users have better results in identifying families of icons and menus, when presented with earcons designed with more complex musical timbres.

Some tests to investigate the possibility of using structured earcons to provide navigational cues in a menu hierarchy (Fig. 3.10, also performed by Brewster [69], proved that not only earcons were a powerful method of communicating hierarchy information, but also that after a week users could recall earcons to identify menus location with the same accuracy. Two important factors also reported were that the recall of earcons depends a lot of their sound quality and of the training method (weather the training is purely textual or by experimenting).

After analysing some experiments that used earcons to provide information, Absar et al. [51] found that this type of sonification have better results in:

- **Desktop interfaces**: widgets; menu hierarchies; business and workplace applications; graphics and tables; hypermedia and web interfaces.

- **Alarms and warning systems**: vehicle-collision detection.

- **Immersive VEs**: assembly tasks.

- **Mobile systems**: mobile phone menus.
3.2.3 Data Sonification

In order to correctly sonify an information display, it is important to take in account the nature of the data to be presented and the task of the user. Also, it is necessary to understand that some dimensions of sound are perceived as categorical (e.g., timbre) and other attributes of sound can only be perceived along a perceptual continuum (e.g., frequency, intensity) [72]. Not less important to have in mind, is that the interpretation and scale of a data-to-display relationship, varies with the conceptual dimension used (size, temperature, price, etc.) [73].

Data can be discrete (e.g., events or samples) or can be represented by a continuous flow of information. A correct sound mapping, depends on the data type: for example, categorical data types should be represented by categorically changing acoustic variables (e.g., timbre) while interval data is better represented by continuous acoustic variables (e.g., pitch or loudness) [72]. Data sonification also depends on the user's functions and goals within a system:

- **Monitoring tasks** requires the user to do an action upon receipt of the sonified message (like answering a ringing phone).

- **Awareness-related tasks** need sound mainly to embellish the occurrence of a process, this is, sound is less action-oriented and more a tool to provide ongoing feedback regarding task-related processes.

- **Data exploration tasks** are more related to the quantification, comparison and visualization of data structures, being sound a tool to improve data interpretation (e.g., auditory graphs [4]).

![Figure 3.11: Map of a design process of Parameter Mapping Sonification (PMSon). (Taken from [72])](image)

**Parameter Mapping Sonification (PMSon)** [72] is the translating of data features into sound synthesis parameters, for the purpose of data display (Fig. 3.11). This is a process fully dependent from

---

4 Auditory graphs are the auditory equivalent of mapping data to visual plots, graphs and charts.
the human perception (green areas). In fact, this subjective component is what turns the design process not so straightforward as the grey and blue areas of (Fig. 3.11) may suggest.

PMSon has the potential to represent both physical and data spaces in a variety of ways that allow for exploration (exploratory PMSon) and observation monitoring (observational PMSon). Therefore PMSon is potentially useful in a wide range of applications from data mining to assistive technology for the visually impaired. There are several tools like the Sonification SandBox [74] or SonART [75] that turns PMSon quite feasible for a host of applications.

A good example of the use of exploratory PMSon to display complex data, is the Cassidy’s et al. [76]. An image of a colon cell is represented both visually and aurally (Fig. 3.12). When the user presses the mouse button over the image, there exists an auditory representation of the pixel. This enables the user to easily identify which are the malignant or benign tissues.

The Allobrain’s [11] sonified semi-autonomous agents (Fig. 3.2) are a good example of an observational PMSon.

**Figure 3.12:** An Hyperspectral image of the a colon cell, that can be visually and aurally explored. (Taken from [76])

Model-based sonification (MBS) provides a means of systematically sonifying data in the absence of an inherent time domain. It is a technique that looks at how acoustic responses are generated in response to the user’s actions. Here the data neither determines the sound signal nor features of the sound, instead it determines the architecture of a dynamic model which in turns generates sound (Fig. 3.13).

5Audio examples S15.8, S15.9 and S15.10 in [http://sonification.de/handbook/index.php/chapters/chapter15](http://sonification.de/handbook/index.php/chapters/chapter15), accessed on 28/12/2011
Model-based sonification (MBS) is a technique that looks at how acoustic responses are generated in response to the user’s actions. (Taken from [72])

Model-Based Sonification addresses our everyday and analytical listening skills, of two types of sounds:

- **Passive sounds**: sounds that are not directly caused by the user’s own activity. These sounds give information about the environment and may direct user’s attention as well as alert them.

- **Active sounds**: sounds that are caused by the user’s physical activity. These can be contact sounds caused by an indirect or direct manipulation of physical objects, the rustle of clothes while moving, the clip-clop of footsteps, the hiss of breathing, etc..

A useful sonification model to reduce the complexity for a given high-dimensional data set is the Growing Neural Gas (GNG) [77]. GNG is an undirected graph of vertices called neurons (or nodes) and connecting edges. This graph is an audio-visual representation of an adaptation process to the topological structure of the data distribution (Fig. 3.14).

Sonifying neurons, according to their energy level (see [77]), gives a sound signature for each network (in the case of Fig. 3.14 there are three networks and therefore three sound signatures). This is very useful to do rapid A/B comparisons between data clusters, particularly when data has four or more dimensions since it is impossible to have visually representations of dimensions higher than three.

**Figure 3.13**: Model-based sonification (MBS) is a technique that looks at how acoustic responses are generated in response to the user’s actions. (Taken from [72])

3.2.4 Motion Synthesis

Sometimes, when at a meeting, you don’t have to look at a person to understand if he/she is interested in what you are saying. Perhaps, you have already noticed that you can “perceive someone’s interest” just by hearing their movements (a deep breath for example, is a clearly demonstration of boredom).

**Body Motion** Synthesizing humans body motion is important to enhance presence and awareness [58, 78–81]. Effenberg [79] considers that the sonification of human movement is a new approach to support motor perception, motor control and learning in sports and rehabilitation. Analysing fMRI data, he proved that perceptual and motor control mechanisms both benefit of additional acoustic information. Valbom [82] with his project WAVE showed the benefits of using a movement-tracking system and a graphical/sound representation while playing a virtual musical instrument. He found that movement, sound and graphic feedback enriches musical decisions.

**Footsteps** sonification is a vast area of study [83]. Several experiments have proved the importance of footsteps sonification to enhance the sense of immersion in VEs. It is almost mandatory that every VEs, need this sonification to offer a good experience to the user.

As title of example, Giordano et al. [84] tested whether untrained listeners were able to estimate several properties of a walking sound source: the gender and emotion of a walker (anger, happiness, sadness and fear), and the size and hardness of the shoe soles. The majority of participants recognized each of these attributes at a higher-than-chance levels. The recognition of gender and sole hardness and size, parameters strongly correlated with each other (female walkers wear smaller shoes with harder soles), were more accurate than the recognition of emotions.

**Impact Sounds** The study of the perception of isolated impacts is the most developed area within the field of sound source perception [85].

- Perception of the material of struck objects, is linked with energy decay related properties of the sound signals [84].
- Perception of geometrical properties of struck objects, is linked with the frequency of the spectral components [86].
- Perception of the materials of impacting objects, is linked with the spectral properties of the early portions of the sounds [85].

Assembly tasks can be substantially improved with the sonification of Impact sounds. That was proved with the Moll’s et al. [50] research, where she found that reproducing a “joint sound” when the
3-D object fit in the construction or a “falling sound” when the object is no longer being grabbed, can provide some independence to the movements of a visually impaired user.

Integrating “real” impact sounds, can also benefit a Virtual Environment. Lopes [87] and others found that interacting with a surface, using sound gestures, can add more expression to user defined gestures.

**Ancillary Gestures** have been found particularly important to augment musical instruments [88–90]. According to Lähdeoja's [88, 89] research, the gestures that do not directly inject energy into the sound production process, are useful to augment part of the instrument while not requiring full conscious control. His system tries to enrich electric guitar effects with natural gestures, rather than with an independent interface like a foot-switch.

### 3.2.5 Soundscapes

Everything that is heard in a VE, is part of the soundscape. Nevertheless, if what we see doesn’t “fit” with what we hear, the sense of immersion is drastically affected. When we see images from a beach, we expect a soundscape with the seagulls screech and the waves hitting the sand.

When designing a VE, sounds must be carefully chosen. If users generate their own objects and sounds are produced in their interaction with the VE and other users, it is necessary to automatically characterize a given soundscape and search for sounds that best fit that characterization [91].

A work done by Finney at al. [92], showed that the Immersive properties of generated soundscapes from online databases, are preferred to the soundscapes done with real recordings and edition. This is very important, because every VEs can now focus their attention in characterizing soundscapes rather than in recording or synthesizing sounds.

### 3.2.6 Emotions

“The emotional dimension of sonic interaction is expressed through subtle variations within a single sound object and across articulations of sound events, with possible interaction with the other senses.” [66]

The experiment made by Vitale [93], proved that everyday actions such as knocking on door can communicate emotions. This fact may lead to effective technological augmentation of everyday objects.

Polotti [94] performed some studies that tested rhetoric and non-rhetoric earcons for actions such as copy, paste, undo, and so forth.. They found that using a rhetoric scheme led to a number of correct associations (earcon with action) much larger than what could be achieved with non-rhetoric earcons.
3.3 The state of the art

According to the related work, we can assume that sound can enhance presence and awareness in VEs. We can also state that hearing and vision complement each other and in order to simulate realistic environments, 3-D sounds are important.

Increasing the modalities of sensory input in a VE can also increase the sense of presence. This is not a surprise, since vision, hearing, touch, smell and taste are what we experience in the real life. Previous works advise us that we must look to every physical modalities when building a VE, rather than trying to exhaustively improve the graphics or the sound systems with a trade off with latency problems.

Sonification techniques must be wisely chosen. Many experiments already stated that aspects like the learning curve and the relation with day life activities, must be considered before choosing earcons or auditory icons. More, the kind of data to be sonified (if it is a size, a temperature, a price, etc.) or the perception that we want to transmit (geometrical proprieties, material types, emotions, body motions, etc.) have their related timbre, pitch or loudness that when incorrectly chosen may adversely affect the VE. Nevertheless sonification techniques are a great tool to enhance presence in a VE.

3.3.1 What is being forgotten

After our analysis of the related work we found out that the following topics are being left behind by the community:

1. Even though that there exist many experiments that clearly prove the importance of sounds to enhance presence and awareness, they lack on the analysis of the differences between using Stereo and 3-D sounds.

2. Another reality is that there aren’t experiments which weight the importance of each physical modality to enhance presence in Virtual Environments. These kind of analysis are important for example, to decide what to do after implementing 3-D graphics and sounds: introduce smell or haptics stimuli?

3. Least but not less important, there aren’t any standards to mount a 3-D sound system in a VE. Most of the times the sound system needs to adapt to the room: which can drastically influence sound localization (stated by many experiments). Also, this adaptation can contribute to a visual pollution (lots of cables and loudspeakers surrounding users). Never any experiment had evaluated the real impact of this pollution on sound localization or even in the sense of presence.
Assembling the Virtual Environment

Contents

4.1 Spatial Audio Model Used .............................................. 39
4.2 Simulation Room ......................................................... 40
In this section we will describe the spatial audio model used and the room where we performed all the experiments.

4.1 Spatial Audio Model Used

CEDAR is an hands-free Avatar-based Teleconference Environment, therefore binaural spatial audio model wouldn’t suite (it makes use of headphones). Binaural technique wasn’t considered, also because it needs head-tracking and calculations of HRTFs, for every users present in the VE, a process that uses too much resources. The drawbacks of VBAP are that it does not directly answer the question of how to handle distance cues. Using Wave Field Synthesis to produce an accurate three dimensional wave field, would need hundreds of loudspeakers, therefore this spatial audio model is not practical for 3-D sound.

We chose a first order Ambisonics system with 8 loudspeakers because it enables us to place a sound source anywhere around the users, allows an hands-free interaction, doesn’t need users head-tracking and the loudspeakers disposition “fits” in the CEDAR VE and also in our simulation room. This layout also allows us to reproduce sound in stereo (we call it 2-D sound), if only the front loudspeakers are enabled.

Figure 4.1: The room where the Virtual Environment was simulated. Pictures 1 and 2 show the front and back side of the simulation area without curtains. Picture 3: front-side with curtains and other components needed for the first experiment. Picture 4: back-side with the video projector. Picture 5: the control room separated by a curtain.
4.2 Simulation Room

Our simulation room had 3,10 x 5,70 x 3,20 m³. We placed our 8 loudspeakers distanced by 3,10m width, 3,80m length and 2,40m height. This disposition didn’t create a perfect cube, needed for a perfect Ambisonics sound reproduction. Nevertheless, and after some localization tests we found that a normal user could disambiguate sounds coming from the left, right, front, back, up or down positions.

All around the area covered by the sound system (the simulation area), there were black curtains that could hide the loudspeakers and create a separation from the technical area where were a secretary, the sound card and a desktop computer. This curtains could be moved so that the loudspeakers were unhidden (necessary for the third experiment). In the room there was also a projector that could produce a 2,50m x 2m image to the front-side (see Fig. 4.1).
5

Experiments

Contents

5.1 Experiment 1 - Multi-Sensorial Stimulation and Presence in Outdoor VEs . . . 42
5.2 Experiment 2 - Navigation Task . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 53
5.3 Experiment 3 - Machine-Less Ventriloquism . . . . . . . . . . . . . . . . . . . . . . 65
5.4 Summary of Experiments 1, 2 and 3 . . . . . . . . . . . . . . . . . . . . . . . . . . 73
In this section we will look at each experiment alone: motivation, participants, VE specification, procedures, experimental design, results and discussion.

The experiments presented encourage the use of a 3-D sound system when developing an IVE. The first elucidate us about the importance of each physical modality (smell, haptics, hearing and vision) to enhance presence on an Outdoor Virtual Environment (OVE). The second shows the importance of 3-D sounds for navigation tasks. The third experiment proves that it is worth hiding the loudspeakers from the users’ field of view, to improve the subjective perception of sound.

In the end of each experiment we make a summary of the most important points and in the end of this section we summarize every experiments.

5.1 Experiment 1 - Multi-Sensorial Stimulation and Presence in Outdoor VEs

A previous work performed by Slater [12], introduced a new method to measure presence in Immersive Virtual Environments. In this experiment we followed Slater’s methodology to measure presence in an Outdoor Virtual Environment where four physical modalities were activated (vision, hearing, haptics and smell). We also used the Emotiv EPOC EEG headset in order to remove some of the ambiguity in data.

Ten participants alternately categorized the relative contribution of each modality to enhance presence: ones were instructed to pay attention to the illusion of being present in the place depicted by the VE (Place Illusion, PI), others to the illusion that events occurring in the VE were real (Plausibility Illusion, Psi). First they experienced the full environment (where all the physical modalities were activated), and then after deactivating some stimuli, they had to increment the VE with the modalities they thought more important to achieve the same illusion as the full environment.

The results showed that participants who were trying to achieve PI, gave a lot of importance to 3-D sounds and wind. The others who tried to achieve Psi gave more importance to wind and images than 3-D sounds. Nevertheless EEG data showed that this preference on images didn’t meant that they were more engaged in the VE when video-projection was enabled. Smell was considered the less important physical modality, but it was usually activated before feeling PI or Psi.

We believe that these results clearly showed the importance of the other physical modalities rather than vision, to enhance presence in Outdoor Virtual Environments. But more important, we found that in our research Slater’s method is not entirely related to the participant’s subconscious feelings, as measured by the EEG.

5.1.1 Motivation

Virtual Environments are becoming increasingly popular and one niche of application that is attracting attention is related to Outdoor Virtual Environments (OVES), since they allow a museum visitor or tourist to feel immersed in sites that are otherwise unreachable or difficult to visit, such as natural parks protected by law, or touristic places that can become destroyed by human presence.
In designing these environments, vision, hearing, touch, smell and taste are what VEs try to computer-simulate so that users can report the sense of having been present somewhere else other than the place they really were. In this context, evaluating if the user felt a physical presence in a VE is an important factor to characterize it [36]. However, subjective as it is, the feeling of presence in VEs has never benefited from a combined and generally established measurement approach [36]. Therefore, experiments are typically performed combining different measuring methods such as questionnaires, behavioural and physiological responses or comparison, and others. All of these approaches have their own set of problems, but when used together, they lead to clearer results. Depending on the type of VE, a correct mix of different methods seems to be the way to give us a more reliable definition of Presence.

If one is developing a VE for people that can see, then obviously vision is the most important modality to navigate and localize objects, because they rely more on what is seen than on what is heard, touched, smelled or tasted. So perhaps, the sense of presence is greater when visual stimulus is available. But what if the visual stimulus isn’t considered in VEs? Can the feeling of presence become equal or even greater? VEs that try to simulate Outdoor Real Environments (OREs) normally need to give much more importance to other physical modalities. In a journey through the sea or forest, what one expects to feel is wind, heat, surrounding sounds or the whiff of eucalyptus and sea smell. Therefore, we set out our research question as the following: In VEs highly dependent of other physical modalities rather than vision, can a conjunction of sounds, haptics and smells, enhance the sense of presence in the same way (or even more) as a conjunction of sound and images (a typical VE)? Perhaps more importantly, we were interested in designing an evaluation method that could prove useful and more reliable when measuring the sense of presence.

Slater and others [12] define the notion of immersion in VEs: One system (A) is said to be more Immersive than another (B) if A can be used to simulate an application as if it were running on B. They demonstrated how this concept could be used as the basis for a psychophysics of presence in VEs, the sensation of being in the place depicted by the virtual environment displays (Place Illusion, PI), and also the illusion that events occurring in the virtual environment are real (Plausibility Illusion, Psi). This configures an ideal framework for setting our experimental approach. We performed an analysis of the influence of four different physical modalities (vision, hearing, haptics and smell) in enhancing the sense of presence on a virtual journey through the sea and the Laurissilva forest of Funchal, Portugal. Most of the forest is found in the northern part of Madeira Island due to the lower temperatures and higher humidity. This forest is classified by the UNESCO as World Heritage Site, and is a very protected environment, not easily reachable by humans in some areas. These characteristics turn it into an ideal scenario content for designing an OVE and evaluating it.

After installing the OVE, we coupled Slater’s method [12] with an approach based on EEG data gathering and analysis using an Emotiv EPOC EEG headset in order to remove some of the ambiguity in data.
5.1.2 Participants

Ten users (seven of them males) participated in the experiment. Their average age was 26 ± 10 (S.D.) years. Only one had running nose in the day of the experiment. Four users had some prior knowledge of the simulation room, but all of them did the experiment for the first time.

5.1.3 Virtual Environment Specification

Apart from the characteristics previously stated, the simulation room was composed by: a chair in the middle of the simulation area with two lights (blue and green) bellow it. At the bottom right and left side of the user there were four more lights. At the front-left lower side there was one fan and one heater behind it. At the front-right lower side there was another but bigger fan. Other two smaller fans were positioned in the front-top side of the user. These fans had car air fresheners attached to them, and they were turned on and off to simulate the smell of forest and sea (see Fig. 5.1).

All the electrical components were controlled through the X10 protocol. The lights could be dimmed, but the fans and the heater could only be turned on and off.

The soundscape was previously composed for a first order Ambisonics sound system through mono sounds. It was deployed in the 6DSpaces Science Center in Funchal, Madeira. The synchronization of sound, images and on/off or dimming of electrical components was achieved through MIDI messages sent by Ableton Live.

Before starting each session participants were asked to put the Emotiv EPOC EEG headset on their head. About 15 minutes were spent to setup it and to explain the user the apparatus and its purpose. After, they were asked to sit on the chair and told that they could move freely, with the condition of being sited all the time.

The Virtual simulation consisted of an approach through the sea to the island of Madeira. When arriving at the coast the user was presented with several local fish birds flying around him. Then the journey continued to the Laurissilva Forest, where there were birds, waterways, footsteps, wind, etc. Finally he/she entered inside a cave and finds a water cascade. This journey had a duration of about four minutes.

5.1.4 Properties

The property vector is \( P = [V, H, Ha, S] \), where V refers to vision (no visual stimulus, lamps, lamps and video projection), H to hearing (2-Dimensional Audition or 3-Dimensional Audition), Ha to haptics (no haptics stimulus, wind, wind and heat) and S to smell (no smell, forest and sea smell).

It is important to understand that the smell of car fresheners in the room was constant. Nevertheless, when the smell fans are turned on, the sea or forest smell becomes more intense.

\(^1\)http://www.x10.com/homepage.htm, accessed on 13/08/2012
\(^2\)http://www.ableton.com, accessed on 13/08/2012
Figure 5.1: The room where the Virtual Environment was simulated. In the top most we can see the user sited on the chair with some green and blue lights beneath and with the Emotive in his head. Picture 3: low wind fan with the heater behind it. Picture 4: two fans with car air fresheners attached. Picture 5: green and blue lights dimmed. Picture 6: strong wind fan.

(V) Vision

• (V=0) No visual stimulus.

• (V=1) Lamps. In this mode it were used 3 blue lights and 3 green lights, in the left and right lower side of the user and below the chair. Their intensities and colour could be dimmed in accordance with the place being simulated.

• (V=2) Lamps and video projection. This mode had, apart from the lights, a projection of the sea and the forest in front of the user. It is important to note that the projection was composed by several static images.

(H) Hearing

• (H=0) 2-Dimensional Audio. Four loudspeakers positioned in the front of the user reproduced the sound.

• (H=1) 3-Dimensional Audio. The sound could be heard in azimuth and altitude. For this, a first order Ambisonics system with 8 loudspeakers was used.

(Ha) Haptics

• (Ha=0) No haptics stimulus. In this mode, both wind and heat were dismissed.

• (Ha=1) Wind. In this mode, two fans produced wind in accordance with the place being simulated (strong wind or breeze).

• (Ha=2) Wind and Heat. In this mode, apart from the wind, heat could be produced by the means of a heater in accordance with the place being simulated (hot or not hot).
(S) Smell

- (S=0) No smell. In this mode, both sea and forest smell were dismissed.
- (S=1) Smell. In this mode, two small fans produced wind in the direction of the users’ nose. Those fans had car fresheners attached, one with sea smell and the other with forest smell.

5.1.5 Procedures

When the participants arrived at the laboratory they were given verbal information about the experiment and some instructions to read. Then they were assisted to don the *Emotiv EPOC EEG* headset and given some information about the apparatus and its purpose:

- Since this is an unknown “gadget” for the most participants, they spent some time testing it with the *Emotiv EPOC* interface. They were able to watch the correspondent movements of the eyebrows, mouth or eye blink in a virtual head. They could also observe a real-time graphic of their engagement/boredom and instantaneous excitement. We asked each participant if they thought that the graphic values corresponded with their engagement or excitement. Most of them weren’t sure, but some identified a correlation.

Once the user was seated, the back curtain was closed and the simulation area was partially in darkness.

Before starting the experiment they were able to listen short periods of sounds in 2-D and in 3-D, feel the two wind differences, the heat, watch some projected images and the different lights available as well as to experience the two different smells (sea and forest). After that, the user learned the configurations that he could ask for.

At the end of the experiment the user had to respond to a questionnaire.

5.1.6 Stimuli Configurations

After getting familiarized with the VE components and capabilities, as described above, the users experienced the simulation with all stimuli in their maximum configuration (full environment) and the EEG data was recorded. It is important to note that before experiencing the full environment the users were given one of the following two instructions:

- (PI) “What are the most important stimuli, to make me feel that I am no longer in this room, but in somewhere else?” Later (when some physical stimuli are disabled) we will ask you to get to that feeling, by enabling the stimulus or stimuli you think are essential for achieving it.

- (Psi) “What are the most important stimuli to make me feel that, what is happening is very close to reality?” Later (when some physical stimuli are disabled) we will ask you to get to that feeling, by enabling the stimulus or stimuli you think essential for achieving it.

After the full environment simulation, participants had one trial that, depending of the order of arrival, started from a different initial configuration. For example if the participant was the third to
Table 5.1: Initial Configurations

<table>
<thead>
<tr>
<th>Initial Configuration</th>
<th>Vision(V)</th>
<th>Hearing(H)</th>
<th>Haptics(Ha)</th>
<th>Smell(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

arrive, first he/she would experiment the full environment simulation with the instruction PI and then he/she would be given the second initial configuration. Else if the participant was the sixth to arrive, first he/she would experiment the full environment simulation with the instruction Psi and then he/she would be given the third initial configuration. The initial configurations were as shown in Table 5.1.

They were encouraged to make transitions and stop whenever they reached the same level of PI or Psi felt in the full environment. In each configuration the EEG was recorded, with markers like (strong wind on, smell off, image changed, user felt Psi, etc...) for an easier interpretation later on.

In order to encourage participants to think carefully about their transitions some rules were imposed:

- Transitions could only be in one direction, this means that when a higher order of a property was chosen, it was impossible to undo it or go backward.
- Only one-step transitions could be made. It was impossible to jump from ‘no visual stimulus’ to ‘lamps and video projection’.
- After the participants had chosen the configuration at which they felt the same PI or Psi, they were asked to continue until they had completed five transitions.

5.1.7 Experimental Design

Depending of the order of arrival to the laboratory, participants were given the PI or Psi instructions (see above the instructions).

Our experiment had one factor with two levels, PI or Psi. For those, three variables were considered:

1. The 4-tuple [V,H,Ha,S] at which a participant declared a match (with PI or Psi).
2. The transitions, i.e. the set of all transitions from the initial configuration to the fifth and last configuration.
3. The engagement/boredom and the instantaneous excitement in each configuration.

Half of the users were given the instruction PI and the other half Psi. Overall, users spent 10 minutes to don the Emotiv Epoc EEG headset and understand its apparatus; 5 minutes learning which were the several stimuli available and about 40 minutes in the experiment (from full environment until configuration 5).
5.1.8 Results

Method of Analysis  For each user, different initial configurations and instructions (PI or Psi) were given. In that way we guarantee that every stimuli are activated in an initial configuration for each instruction. This is important because we force the user to pay attention to every stimuli and it prevents the monotony of the results: hearing and vision as the first choices (a normal VE and what people are used to).

We followed three methods of analysis. First, we considered the configuration at which users declared a match. We used the Bayes’ Theorem to estimate the probability of a match being declared when the user is experiencing the configuration \([V,H,Ha,S]\). Our population had 5 users for each instruction so: \(P(PI)\) and \(P(Psi) = 0.5\). The probabilities for having a match on PI or Psi instruction are estimated with the following equations:

**PI instruction:**

\[
P(v, h, ha, s|PI) = \frac{P(PI|v, h, ha, s)P(PI)}{P(PI|v, h, ha, s)P(PI) + P(Psi|v, h, ha, s)P(Psi)} \tag{5.1}
\]

**Psi instruction:**

\[
P(v, h, ha, s|Psi) = \frac{P(Psi|v, h, ha, s)P(Psi)}{P(Psi|v, h, ha, s)P(Psi) + P(PI|v, h, ha, s)P(PI)} \tag{5.2}
\]

Second, we individually looked at every stimulus and estimated the probability of one being selected by the user on each transition and the probability the stimuli being active when the user reported feeling PI or Psi. Third, we made an analysis of the EEG data and used the Pearson correlation to find a relationship between different environments. We consider a different environment when a VE-A has less or more activated stimuli than a VE-B. The users experienced every stimuli in their maximum configuration (full environment) in the beginning and in the end of the experiment. We made the assumption that the EEG levels of engagement/boredom and the instantaneous excitement are equal at both the beginning and end of the experiment. Therefore we estimated the correlation \(r\) between each different configuration (initial, 1, 2, 3, 4, as per table 1) and the average between the first and last full environments.

Probability of a match configuration. The probability distributions for having a match on the instructions PI and Psi, are shown in Fig. 5.2a (only configurations with at least a probability > 0 are shown). The PI group activated 3-D Audio and Smell together most of the times while Psi never (60% compared to 0%). The Psi group activated 2-D Audio and lights+video projection (V=2) together most of the times while PI never (60% compared to 0%). Overall there is a 50% probability that Wind and Smell are activated together, to feel PI/Psi.

Probability of activating a stimulus on each configuration and when feeling PI/Psi. The probability distributions of having each stimulus activated when felt PI or Psi are shown in Fig. 5.2b. When
analysing this graphic it is important to note that, haptics(2) is a sum of wind and heat and vision (2) a sum of lights and video projection. For the PI group, 3-D Audio and Smell exhibit the highest probabilities of being activated. Also for the PI group, haptics(1) (only wind) has a higher probability than haptics(2) (wind and heat). Considering the Psi group, we can see that to feel Psi, a conjunction of lights and video projection is more important than lights only. The probability is equal for both vision(2), haptics(1) and smell(1).

The probability distributions for choosing a stimulus on each configuration are shown in Fig. 5.3. For the PI group there were two major choices on the first transition: 3-D Audio or wind (hearing(1) and haptics(1) respectively). The Psi group selected only the haptics modality, wind or heat (heat only for the ones who started with an initial condition with Ha=1). For the PI group, 40% of the times vision(2) (video projection) has only been chosen in the last configuration. For the Psi group the last choices were usually the Smell and the Heat (40% and 50% of the times respectively).
Correlation between full environment and the other configurations. We considered that the user has the best experience possible when every stimulus is in its maximum configuration. Therefore the values of engagement/boredom and instantaneous excitement must have an higher correlation to the first and last full environments, in order to say that the configuration [V,H,Ha,S] causes the same effect as the full environments. The correlations $r$ between every configurations (initial, 1, 2, 3, 4) and the average of the first and last full environment are shown in Fig. 5.4.

Both for engagement/boredom and instantaneous excitement in the two groups, the values differ a lot from user to user. This is because different users have chosen different transitions and have started from different initial configurations and instructions. Also it is clear that the highest correlation isn’t always a synonymous of feeling PI or Psi. In these values we can also see that the correlation doesn’t have a progressively increase from transition to transition.

- **Important Note**: In the graphic “Engagement PI” the User2 exhibited very high values for each transition. This was because “Emotiv EPOC EEG” wasn’t properly mounted on the user’s hea. Therefore we discarded the values of User2.

Since there isn’t much valuable information on an analysis of the correlation through each configuration, we decided to look at the most significant increases from a configuration A to a configuration B. We cross-checked the correlation value with the stimulus that was activated in that configuration: the most significant increase due to the addition of a given stimulus is shown in Fig. 5.5. For PI and Psi, smell was never related to a significant increase in the two emotions. Vision(2) (video projection) was neither responsible in the PI group. Wind (haptics 1) was the most important stimulus to significantly
increase engagement on both groups. Regarding instantaneous excitement, haptics(2) was the most relevant for PI group and vision(1) for PSI. This data suggests that vision isn’t as relevant for the PI group as it is for the Psi group.

5.1.9 Discussion

3-D audio is important to feel PI and less important, although not significantly, to feel Psi. We consider that a configuration with vision at its maximum and hearing at its minimum, means that 3-D audio is not so important. Only some of the Psi group had chosen a configuration like [2,0,*,*] on a match (see Fig. 5.2a). From those users, only one had just one more stimulus activated (Haptics). The rest had at least two more stimuli activated (Smell and Haptics).

It is also possible to see the importance of 3-D Audio in Fig. 5.2b, where hearing(1) has a 20% probability to be active in PI group and 5% in Psi. Nevertheless, for Psi group vision(2) wasn’t significant, because it had the same probabilities as haptics(1) and smell(1). This means that, even though that for the Psi group, 3-D audio seems to be not so important, the other two physical senses are essential.

Curiously enough, excitement never exhibited a significant increase due to the activation of the video-projection (Fig. 5.5). For the PI group, engagement didn’t had it either, while in Psi group video-projection was responsible for a significant increase. Another interesting observation is that engagement in the Psi group had a larger correlation value than in PI group, due to the activation of 3-D audio. This might mean that, **Psi users weren’t well aware of the differences conveyed by 3-D sounds and just realized it after saying that they had felt Psi**. This was an observation that great part of the users made, when doing the experiment and after activating 3-D audio: “Now I realized the differences on a 3-D sound system”.

First choices are Haptics and Hearing. If we look to Fig. 5.4 we can see that in the first configuration, PI users activated Hearing(1) or Haptics(1), while the Psi users, Haptics(1 or 2). Haptics(2) only for those who started from an initial configuration like [0,0,1,0]. Since the first choice is considered to be the most relevant when trying to feel PI/PSi, and considering that in the questionnaires, only 20% of the users had answered that if they repeat the experiment, they would change the order of choices,
we conclude that wind (Haptics(1)), was the most important stimulus to enhance presence in this Outdoor Virtual Environment. The second one was the 3-D Audio (Hearing(1)).

Wind increases the Engagement as measured by the EEG. It was responsible for the highest values both for Psi and PI users (Fig. 5.5). If we had to choose the stimulus that brings the greatest benefit for an OVE, we would say wind.

5.1.10 Summary

We described an experiment that combined Slater’s method for measuring presence with data gathered from an EEG headset, which proved useful for removing data ambiguity, and therefore reducing some of the subjectivity inherent to Presence as an attribute. One of the most important findings in our research is the verification that Slater’s method is not entirely related to the participant’s subconscious feelings, as measured by the EEG. In particular, we noted that the most significant increases in the EEG values of excitement and interest (in correlation with the senses’ configurations) was not performed at the same time as the user claimed a PI or Psi response.

This experiment demonstrates the importance of including the EEG dimension as a way to improve the accuracy of using PI and Psi responses to measure presence in a VE. It is important to note that we compared the relative weight of the different senses involved in a virtual experience. Therefore, this observation only applies to these conditions.

Our assumption was that the other physical modalities rather than hearing and vision (a typical VE), are so or even more important to enhance presence in OVEs. While not necessarily an absolute truth (because presence is subjective), the different inputs (EEG data and the different stimuli configurations), proved that our assumption was correct.

Another interesting and unexpected conclusion was that the EEG excitement indicator never exhibited a significant increase due to the activation of the video-projection. Also we noted that engagement in the PI group had a larger correlation value than in the Psi group, due to the activation of 3-D audio. This probably meant that Psi users weren’t well aware of the differences conveyed by 3-D sounds and just realized it after saying that they had felt Psi.

The combination of both methods suggests that tactile perception of wind is the most important stimulus in OVEs. The experiment also elucidates us about the importance of each physical modality (vision, hearing, haptics and smell) in an OVE setting. These conclusions are particularly important for designers interested in crafting similar Virtual Environments.
5.2 Experiment 2 - Navigation Task

In this experiment we carried out a navigation test in CEDAR. The users needed to identify and mark colliding pipes in three different sets: video projection only, video projection and 2-D sound, video projection and 3-D sound. Depending on the set, the user had some visual or audio-visual cues of the collision nearest to him or her. If the set was the one with video projection only, audio cues were disabled, else audio-visual cues were enabled.

We focused our attention on four factors: number of found collisions; search time between collisions; normalized path length (travelled path length/distance between collisions); and rotation degrees between collisions. The results showed that audio cues make navigation effortless and improve sense of orientation. We found that with 3-D sound, users chose the shortest paths; without audio cues, users lost orientation more often and with 3-D sound users rotate more than with 2-D. We also found that 3-D sound approaches real life behaviours to the Virtual Environment, therefore enhancing presence.

![Figure 5.6: The Virtual Environment Scenario. Pictures 1, 2 and 3 shows the oil tanker and platform. Underwater we can see the coral regions with some oil pipes (4), the Christmas trees (5 and 6) and the pipes connecting the tanker with the platform (7).]

5.2.1 Motivation

Navigation in Virtual Environments needs to be effortless: the user must always be aware of where he/she is (orientation), how to get to a desired place (way-finding) and be able to move through the space (travel). As the major part of VEs are related to the visualization of 3-D models (like CEDAR), navigation is normally based on visual information. This information can be a radar, a compass, an highlight of the important objects, amongst others. Less often, a VE may also use auditory cues to help navigation [11, 48].

In this experiment we made an evaluation of the ease of navigation in the CEDAR VE. We used three different sets: video projection only, video projection and 2-D sound, video projection and 3-D sound. The objective was to identify the places where oil pipes were colliding with other differently
coloured oil pipes or coral regions. The user had 3 minutes to find the most number of collisions. From the measured data we analysed four factors: the **number of found collisions**, **search time between collisions**, the **normalized path length** (travelled path length/distance between collisions) and the **rotation degrees between collisions**.

Our assumption was that navigating in a VE with audio-visual cues is easier than with just visual cues. We also wanted to understand if 3-D sounds contributes to an effortless navigation.

### 5.2.2 Participants

Six users (five of them males) participated in the experiment. Their average age was 26 ± 2,2 (S.D.) years. Only one used glasses and none of them suffered from hearing problems. Two users had some prior knowledge of the simulation room, but all of them did the experiment for the first time.

### 5.2.3 Virtual Environment Specification

The VE was a simulation of an oil extraction location in the middle of the ocean. Above the water were an oil tanker and an oil platform. Both the the platform and the tanker had some oil pipes connecting each other, and others connected to the oil wells by the means of Christmas trees[^3]. In the seabed were also some coral regions (see Fig. 5.6).

![Navigation control method in the VE. The user has the Wii remote control in his hand to navigate and kinect is tracking his right-hand position.](http://www.xbox.com/en-gb/Kinect/)  

handled the pointing position of the user’s hand (see Fig. 5.7). During the experiment, the user had some visual or audio-visual cues of the collision nearest to her. If the set was the one with video projection only, audio cues were disabled, else audio-visual cues were enabled. It is important to note that, in the setup with 2-D sound and video projection, the user could only listen to the sounds coming from her front.

- **Visual Cues**: 1) The nearest collision blinked each 2 seconds with a green colour; 2) There was a radar in the bottom-right corner, that showed the position of the nearest collisions (see Fig. 5.8).

- **Auditory Cues**: 1) Each object in the environment had its correspondent sound: for example the oil tanker was composed by the noise of the engine and the chirping of seagulls, the sea was composed by the sound of waves. Depending from the zone, the reverberation had different values and some filters were used: underwater, the sounds of the environment objects were low-pass filtered and different reverberation values were used; 2) The sonar metaphor was used to sonify the nearest collision: apart from the correct positioning of the collision sound, some filtering was applied in accordance to Table 2.1. Also, a collision sound further away had an emphasis on its low-frequencies [31].

In (audio2) you can listen the environment sounds and the audio cues. In the first 30 seconds you are above water, and you can listen the sounds of the waves, the oil platform and the boat. When you immerse in the ocean you can notice that a low-pass filter is being used. The sounds of the collisions nearest to you are always present “plom...plom” or “plim...plim” depending if it is a collision with a coral or another pipe respectively.

**Figure 5.8**: Visual Cues. The radar is on the bottom-right showing the collisions near to the user. The green aura represents the nearest collision.
5.2.4 Procedures

When the participants arrived at the laboratory they were given verbal information about the experiment and some instructions to read B.1.

Before starting the experiment on each different set (video projection only, video projection and 2-D sound, video projection and 3-D sound), the user was encouraged to test and learn the navigation mechanism and to mark some collisions in a test environment using the different visual or audio-visual cues (see Fig. 5.9). This adaptation had no time-limit and it could be repeated until the user felt comfortable with the navigation system and the cues.

The user had 3 minutes to mark as much collisions as possible in each different set. The time left was always visible in the top right corner of the screen. At the end of the experiment the user had to respond to a questionnaire B.2.

5.2.5 Task

The task was to find and mark as many collisions as possible in 3 minutes time. A collision is when pipes from different colours are touching each other, or when a pipe is on top of a coral region. If pipes from the same colour are touching each other, the collision isn’t considered. The same happens for pipes colliding with other objects rather than different coloured pipes or coral regions (Fig. 5.10).

The users could “fly” around the environment by the means of a Wii Remote Control (see Fig. 5.11 for a description of each button). The collision was considered found when marked: to mark a collision the user needed to be relatively close to it, because he/she needed to trigger a button just visible when near (see Fig. 5.10 picture 2).
Figure 5.10: Valid and Invalid Collisions. The first picture shows two oil pipes with the same color colliding with each other (this is an invalid collision). Picture 2 shows a collision between two pipes as well as the trigger button. Picture 3 shows valid collisions between pipes and coral regions.

There were 22 collisions possible to find: 5 of different colored pipes and 17 of pipes with coral regions. The minimum and maximum values for the \([x, y, z]\) position of the collisions were \([x=-65,1068]\), \(y=47,453\) and \(z=-684,920\). The distances between near collisions ranged between 32 and 560 with an average of 215.

5.2.6 Experimental Design

The first set of the VE was dependent of the order of arrival of each Participant. The first participant made the experiment in the following order: video projection only, video projection and 2-D sound, video projection and 3-D sound. The fifth participant: video projection and 2-D sound, video projection and 3-D sound, video projection only. There were also three starting positions that differed from set to set, but the number of collisions and the disposition of the pipes and other objects were equal in each set.

Figure 5.11: Wii Remote Control. Method to navigate in the Virtual Environment.

In each set, the user had 3 minutes to find the most number of collisions as possible. From the
measured data we focused our attention on four factors:

1. Number of found collisions, the user needed to be relatively close to the collision so that it could be marked and considered found;

2. Search time between collisions, the user wasn’t forced to mark the collision that the visual or audio-visual cues indicated;

3. Normalized path length (travelled path length/distance between collisions);

4. Rotation degrees between collisions, the distance between the user and the display screen wasn’t considered for this measure.

Overall, users spent 2 to 5 minutes on the test environment before starting the experiment in the three different sets. So all the experiment had a duration of about 20 minutes.

5.2.7 Results

**Number of found collisions.** Each user was able to find and mark more collisions when audio cues were enabled. According to Table 5.2, on average users found ±3 more collisions on the setups with 2-D and 3-D sound. Also the maximum number of collisions to mark was almost achieved when 3-D sound was enabled (21 out of 22). Looking at the total number of found collisions for all users, we see that audio cues increased the number of found collisions in nearly 20%. When observing the standard deviation values on Fig. 5.12 we can say that for the 3-D setup the minimum values were higher ($V = 11.7$ (STD 3 collisions), $2D = 14.5$ (STD 6 collisions), $3D = 14.2$ (STD 5 collisions)).

<table>
<thead>
<tr>
<th>Setup</th>
<th>Average</th>
<th>Min</th>
<th>Max</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video projection only</td>
<td>11.7</td>
<td>7</td>
<td>15</td>
<td>70</td>
</tr>
<tr>
<td>Video and 2-D</td>
<td>14.5</td>
<td>5</td>
<td>20</td>
<td>87</td>
</tr>
<tr>
<td>Video and 3-D</td>
<td>14.2</td>
<td>9</td>
<td>21</td>
<td>85</td>
</tr>
</tbody>
</table>

**Table 5.2:** Number of found collisions in three minutes for each set

![Figure 5.12: Average of found collisions in three minutes for each set. V for Video projection only, 2D for Video and 2-D Sound, 3D for Video and 3-D Sound.](image)
Search time between collisions. The time spent between marking one collision and another was significantly lower when the audio cues were enabled. The setups with 2-D and 3-D sound improved the searching time between collisions in ± 6 seconds. Video and 3-D setup established a maximum of 20 seconds between one collision and another (see Table 5.3).

Looking at Fig. 5.13 we can see the averages for each set and the correspondent Confidence Interval (CI): \( V = 14,8 \) (CI 3,3 seconds), 7,9 (CI 1 second), 8,7 (CI 1 second).

The data samples did not present a normal distribution, therefore the Friedman non-parametric test showed that there were significant differences in the time spent between collisions depending on which setup used, \( X^2(2) = 109,1; p < 0,001 \). The Wilcoxon post-hoc test also presented significant differences between V-2D \( (Z = -7,271; p < 0,001) \), V-3D \( (Z = -7,271; p < 0,001) \) and 3D-2D \( (Z = -6,014; p < 0,001) \).

Table 5.3: Search time between collisions in three minutes for each set

<table>
<thead>
<tr>
<th>Setup</th>
<th>Average</th>
<th>Min</th>
<th>Max (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video projection only</td>
<td>14,8</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>Video and 2-D</td>
<td>7,9</td>
<td>8</td>
<td>31</td>
</tr>
<tr>
<td>Video and 3-D</td>
<td>8,7</td>
<td>8,5</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 5.13: Average of search times between collisions in three minutes for each set. V for Video projection only, 2D for Video and 2-D Sound, 3D for Video and 3-D Sound.

Normalized path length between collisions. The normalized path length on Video and 3-D sound setup is lower. If we look at Table 5.4 and Fig. 5.14 we see that the average is equal for both 2-D and 3-D setups, but the Confidence Interval (CI) is much lower for 3-D. This lets us say that the average of normalized path length for 3-D setup varies less and therefore the value is more reliable.

The data samples did not present a normal distribution, therefore the Friedman non-parametric test showed that there were significant differences in the normalized path length depending on which setup used, \( X^2(2) = 64,794; p < 0,001 \). The Wilcoxon post-hoc test presented significant differences between V-2D \( (Z = -5,949; p < 0,001) \) and V-3D \( (Z = -6,901; p < 0,001) \), between 3D-2D there weren’t significant differences \( (Z = -1,343; p = 0,179) \).
Table 5.4: Normalized path length in three minutes for each set.

<table>
<thead>
<tr>
<th>Setup</th>
<th>Average CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video projection only</td>
<td>1.99 0.32</td>
</tr>
<tr>
<td>Video and 2-D</td>
<td>1.35 0.34</td>
</tr>
<tr>
<td>Video and 3-D</td>
<td>1.35 0.11</td>
</tr>
</tbody>
</table>

Figure 5.14: Average of normalized path length in three minutes for each set. V for Video projection only, 2D for Video and 2-D Sound, 3D for Video and 3-D Sound.

Rotation degrees between collisions. The users rotated significantly more between collisions when audio cues were disabled. Observing Table 5.5 and Fig. 5.15 we see that in the 2-D set the users rotated less and with 3-D sound they rotated more 20 degrees on average.

The data samples did not present a normal distribution, therefore the Friedman non-parametric test showed that there were significant differences in the rotation degrees between collisions depending on which setup used, $X^2(2) = 121,463; p < 0.001$. The Wilcoxon post-hoc test presented significant differences between V-2D ($Z = -7.115; p < 0.001$), V-3D ($Z = -7.722; p < 0.001$) and 3D-2D ($Z = -6.409; < 0.001$).

Rotation and Distance through Time. To understand the users’ behaviour, we analysed the three longest paths between two marked collisions in the three minutes time. If we pick the total distance or the total rotation and analyse it through the total time spent between collisions, we can have an overview of a typical behaviour on each setup.

As seen before, the setup with Video projection only presented the worst results for number of found collisions, search time, normalized path length and rotation degrees between collisions. Data in 2-D and 3-D was very similar in the previous analysis, so on the following analysis we focused our attention on those setups.

In a first look to Fig. 5.16 we see that the 3D line is always beneath 2D until $t = 0.75$. Also if we look at theirs Confidence Interval (CI) positive values, we can say the same.
Table 5.5: Rotation degrees between collisions in three minutes for each set.

<table>
<thead>
<tr>
<th>Setup</th>
<th>Average</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video projection only</td>
<td>143,7</td>
<td>44,9</td>
</tr>
<tr>
<td>Video and 2-D</td>
<td>67,2</td>
<td>12,9</td>
</tr>
<tr>
<td>Video and 3-D</td>
<td>87,2</td>
<td>17,7</td>
</tr>
</tbody>
</table>

Figure 5.15: Average of rotation degrees between collisions in three minutes for each set. V for Video projection only, 2D for Video and 2-D Sound, 3D for Video and 3-D Sound.

The three samples that we chose to perform statistics analysis presented a normal distribution, therefore we used a one-way repeated-measures ANOVA to assess differences between the normalized path length in the three setups (Table 5.7). There were significant differences only for t = 0.5 and t = 0.75. A Bonferroni post-hoc test revealed significant differences between V-2D and V-3D on both samples.

In a first look to Fig. 5.17 we see that the 3D line is always above 2D until t = 0.75. Also if we look at theirs Confidence Interval (CI), we can say the same.

The three samples that we chose to perform statistics analysis presented a normal distribution, therefore we used a one-way repeated-measures ANOVA to assess differences between the normalized path length in the three setups (Table 5.7). There weren’t significant differences.

Table 5.6: One-way repeated-measures ANOVA for three samples of the average of the normalized path length through time (see Fig. 5.16).

<table>
<thead>
<tr>
<th>Sample</th>
<th>ANOVA</th>
<th>V-2D</th>
<th>V-3D</th>
<th>2D-3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = 0.25</td>
<td>F_{1,771} = 0.885; p = 0.412</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>t = 0.5</td>
<td>F_{1,599} = 5.832; p = 0.012</td>
<td>p = 0.010</td>
<td>p = 0.014</td>
<td>p = 1</td>
</tr>
<tr>
<td>t = 0.75</td>
<td>F_{1,574} = 21.355; p &lt; 0.001</td>
<td>p &lt; 0.001</td>
<td>p = 0.001</td>
<td>p = 1</td>
</tr>
</tbody>
</table>

*With a Greenhouse-Geisser correction.
*Post hoc tests using the Bonferroni correction.
Figure 5.16: Average of the normalized path length through time. For each user and each setup, it was chosen the three longest baths between collisions marked. V for Video projection only, 2D for Video and 2-D Sound, 3D for Video and 3-D Sound.

Table 5.7: One-way repeated-measures ANOVA for three samples of the of the normalized rotations degrees through time. (see Fig. 5.17).

<table>
<thead>
<tr>
<th>Sample</th>
<th>ANOVA [a]</th>
<th>V-2D</th>
<th>V-3D</th>
<th>2D-3D [b]</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = 0.25</td>
<td>F[1,637] = 1,532; p = 0,234</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>t = 0.5</td>
<td>F[1,895] = 0,657; p = 0,517</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>t = 0.75</td>
<td>F[1,543] = 0,246; p = 0,726</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

\[a\] With a Greenhouse-Geisser correction.
\[b\] Post hoc tests using the Bonferroni correction.

Figure 5.17: Average of the normalized rotations degrees through time. For each user and each setup, it was chosen the three longest baths between collisions marked. V for Video projection only, 2D for Video and 2-D Sound, 3D for Video and 3-D Sound.
Questionnaires. The radar was considered the most important visual cue. Overall the users gave more importance to sounds than the radar for navigation. They strongly agreed that collision sounds are important for orientation. Also they agreed that 3-D sounds were more important than 2-D sounds for navigation. Most of them stated that have felt differences between 2-D and 3-D sounds.

5.2.8 Discussion

Although having three different starting positions that differed only on the rotation of the camera, and the collisions being positioned in the same place, the users didn’t recognize it. If did, we would had a progressive increase of found collisions from set to set.

Navigation controls were not much appreciated by overall users (in the questionnaires, the average was 2.7 from a Likert Scale which varied from 0 to 5). This might have influenced the performance, but since the controls were equal for both 3 setups, we considered that it didn’t influence our results.

Audio cues are essential to improve users’ orientation, way-finding and ease of travel in a Virtual Environment. The results are clear, audio cues can make navigation in Virtual Environments, an effortless task. The objective in this experiment was to find and mark as much collisions as possible in three minutes time. Therefore the objective was best achieved, for the video and 2-D setup which had an higher average and total number of found collisions (see Table 5.2 and Fig. 5.12).

In what concerns to the performance on achieving the objective, we can also state that the two setups with audio cues, presented better results. Less time spent between collisions means that the user got lost less often. Lower normalized path lengths means that the user chose more often the shortest paths. More rotation means more disorientation or can also mean more awareness on what’s surrounding. Having this we can say that:

- With audio cues, users spent less time between collisions so they got lost less often.
- With 3-D sound, users chose shortest paths.
- Without audio cues, users lost orientation more often.
- With 3-D sound users rotated more than with 2-D. We believe that in this case, more rotation means more awareness on what’s surrounding.

3-D sound helps finding the shortest paths more rapidly. Since the values of search time and normalized path length between collisions presented very similar results, it is worth looking at the normalized path length through time (Fig. 5.16). Being closer to the next collision in a shortest period of time, means that the user decided where to go more rapidly. Also the first periods of time are when some subconscious decisions are made, if they aren’t correct, most likely a user changes his mind before being half way from the next collision. When comparing two very similar results, it matters if in the initial moments, the user was or was not closer to objective.
The results showed that the 3D line and CI positive values, were always beneath 2D in the initial moments. This lets us say that not only 3-D sound helps finding shortest paths more rapidly (remember that on Fig. 5.14 we saw that 3-D had the lowest average), but also approaches real life behaviours to the Virtual Environment. Our subconscious decisions are related to our day life experiences, they are effective on a Virtual Environment if the user doesn’t need to change his mind after taking them.

**3-D sound approaches real life behaviours to the Virtual Environment.** When people that can see are lost and need to decide where to go, the first thing that they normally do is look around and see what's surrounding. To make the best decision possible, one most first “take a clear picture” of the world at 360 degrees.

The results on Fig. 5.17 showed that the 3D line and CI, were always above 2D in the initial moments. This lets us do the following observation: most likely the users with 2-D sound, as soon as they had a visual match of a collision, they moved forward to it. It didn’t matter if it was the nearest collision or not, they avoided looking around because they weren’t sure if it worth the effort. On the other hand the users with 3-D sound felt more comfortable to look around and follow what they were hearing, because before rotating they already knew that a collision was there.

Like said before, with 3-D sound subconscious decisions are more effective. Also people seem to be more comfortable to look around and pay attention to what's surrounding, because before rotating they already have an idea of what's there.

### 5.2.9 Summary

In this experiment we evaluated the influence of audio-visual cues on making navigation in a Virtual Environment effortless. We also evaluated the differences between three setups: video only projection, video and 2-D sound, video and 3-D sound. Our assumption was that audio cues could help navigation, especially in the setup with 3-D sound. This was proved to be true.

Video only projection is insufficient for tasks that need a rapid judgement and awareness of what's surrounding. We saw that with audio cues, users spent less time between collisions so they got lost less often; with 3-D sound, users chose shortest paths; without audio cues, users lost orientation more often and with 3-D sound users rotated more than with 2-D.

Analysing the three longest paths between two marked collisions and comparing the total distance or rotation through time, gave us a better understand of users’ behaviours. With 3-D sound subconscious decisions seemed to be more effective and people seemed to be more comfortable to look around and pay attention to what was surrounding, because before rotating they already had an idea of what was there.

In conclusion audio cues make navigation effortless and improve sense of orientation. 3-D sound approaches real life behaviours to the Virtual Environment, therefore enhancing presence.
5.3 Experiment 3 - Machine-Less Ventriloquism

The ventriloquism effect is the maximum unperceived distance between audio and visual stimuli. It is also the association of something physical to a sound source. This means that, even tough that a virtual sound source can be perceived as coming from elsewhere rather than the loudspeakers, the human will always associate the sound to two sources: the image and the loudspeakers.

In this experiment we evaluated the ventriloquism effect with the loudspeakers hidden and visible. The visual source was a talking-head placed in the middle of the screen, the audio-source was increasingly shifted away from the image [listen audio1].

The results proved that hiding the loudspeakers is an easy and effective way to improve the subjective perception of sound in a VE. This is because the maximum unperceived distance between audio and visual stimuli becomes greater, and therefore users tolerate greatest positioning errors (intrinsic to every sound systems). Results also proved that users’ confidence and correctness are greater when judging where the sound is coming from.

![Figure 5.18: Front-face of the Ambisonics Cube with the loudspeakers visible.](image)

5.3.1 Motivation

In the real world, when we listen to someone speaking, we intrinsically associate the sound Source to the movements of the mouth. Our brain is trained to “connect” what we see to what we hear.

A VE can have a 3-Dimensional sound system that is able to position the sound sources relatively close to the visual sources, so that humans perceive sound as coming from the exact same place as the images [10]. Normally this types of sound systems are composed by several loudspeakers or by headphones.

When using headphones, the listener assumes that he/she is “entering” in a virtual space, where the sound is coming from the physical sound sources attached to the ears. This is also true in a VE that uses loudspeakers. Even tough that a virtual sound source can be perceived as coming from elsewhere rather than the loudspeakers, the human will always associate the sound to two sources: the image and the loudspeakers. We, as intelligent beings, connect what we hear to the loudspeakers.
surrounding us. We unconsciously, don’t believe that the sound source is really coming from the image facing us, because what is virtual is not real.

With this experiment we wanted to understand if hiding the loudspeakers and components worth the effort.

### 5.3.2 Participants

Seven users (six of them males) participated in the experiment. Their average age was $27 \pm 2.6$ (S.D.) years. None of them suffered from hearing problems. Five users had some prior knowledge of the simulation room, but all of them did the experiment for the first time.

### 5.3.3 Virtual Environment Specification

The Virtual Environment had two different sets: with the 8 loudspeakers hidden and with the 8 loudspeakers visible.

Since each user has its own particular perception of sounds position, in the beginning of the experiment a personal calibration with pink noise[^6] was made. It is important to note that this calibration was performed only to identify the user perception of the front-left, front-right, front-top and front-bottom position, this is the front-face of the Ambisonics Cube (Fig. 5.18).

Pink noise was used because according to [34, 48, 53], a normal user can use interaural and level differences due to pink noise's high energy at low (below 1.5kHz) and high frequencies (above 1.5kHz). After the calibration, it was possible to precisely shift the sound source 5 centimetres at each time until it reaches the Ambisonics Cube's front-face limits.

There was a video-projection of a talking-head at the center of the screen. We chose solely a talking-head so that there weren’t any more distractors which could influence the perceptiveness of audio and visual stimuli and because, faces are what people are most audible and visually familiarized. The image was always in the same position during all the experiment, the correspondent sound was progressively shifted away from the image by the following manner:

Variables: \( \text{Shift\_Value} = 0 \text{cm} ; \text{Shift\_Positions\_Vector} = \{ \text{Left}, \text{Right}, \text{Top}, \text{Bottom} \} \)

1. Place the sound source in the same position as the talking-head's mouth;
2. Increase 5 centimetres to \( \text{Shift\_Value} \).
3. Shift the sound to one of the positions in the \( \text{Shift\_Positions\_Vector} \);
4. Repeat the shift until all positions are chosen;
5. If the user perceives a shift, the position in the \( \text{Shift\_Positions\_Vector} \) is removed;
6. Repeat 1 to 5 until \( \text{Shift\_Positions\_Vector} \) is not empty.

The user was seated in the middle of the simulation area: his head position was approximately at
the same distance from each loudspeaker (the perfect sweet spot for the Ambisonics Sound System).
He also had a remote control, to point where did he perceived the sound as coming from.

In each shift, the user needed to say if he perceived the sound as coming from the talking-head’s
mouth or from elsewhere. If the user perceived it as coming from elsewhere, he/she was asked to
point out where did the sound was coming from. The size of the pointing aura could be bigger or
smaller, reflecting the amount of confidence that the user had when pointing that the sound was
coming from a certain position (bigger = less confidence) Fig. 5.19.

5.3.4 Procedures

When the participants arrived at the laboratory they were given verbal information about the ex-
periment and some instructions to read C.1 Then they were told to seat in the chair and that they
could move their heads’ freely. They were also told that when answering if sound was coming
from the talking-head’s mouth, the eyes needed to be open. We imposed this condition because
we were analysing the ventriloquism effect, the unperceived spatial difference between the visual
and audio stimuli.

Before starting the experiment, the user needed to calibrate the sound system. By the means
the remote control, they needed to place the calibration sound-source (pink noise) in 4 different posi-
tions: the front-left-most, front-right-most, front-top-most and front-bottom-most. When beginning to
calibrate each position, the pink noise started in a middle position and only one position sound could
be listen at a time. The users could re-listen the position sounds and compare with each others, until
they were sure that the sounds were placed in the correspondent positions.

After the calibration, it was possible to precisely shift the sound source 5 centimetres at each time
until it reaches the front-right,left,top,bottom-most positions.

At the end of the experiment the user had to respond to a questionnaire C.2.
Figure 5.20: Values of Predicted Sound Pressure Level of the room with and without curtains. The values in the legend 64.4dB and 63.5dB are the SPL for the frequency 2.5kHz (which belongs to the most important speech frequencies).

5.3.5 Experimental Design

Half of the participants started the experiment with the 8 loudspeakers hidden and the other half with the 8 loudspeakers visible. Before starting the experiment they needed to calibrate the system (explained above).

Since there exists an absorption of sound by the curtains, we made some tests to understand if this could drastically change the room response in order to influence the experiment. We placed an hyper-cardioid microphone (AKG-C568B) in the middle of the simulation area and measured the room response with Room Eq Wizard. We concluded that there was no need to use equalisation, since there were no big differences with and without curtains (see Fig. 5.20).

- **Important Note**: The measurement of the room response is usually done with an omni directional microphone. Since the purpose was only understand if room’s response differed a lot with and without curtains (rather than perfectly equalise it), we believe that this measurement is valid.

In each set the user needed to answer if he perceived the sound as coming from the talking-head’s mouth, until he identified shifts in every positions (Left, Right, Top, Bottom). He or she were told that, in each 3 shifts they needed to listen and see all the loop at least one time (one, two, three, four, five, six, seven, eight, nine, ten, zero). This was necessary because, different numbers (different sound frequencies) might be perceived as coming from different positions (spectral cues).

The user was told that he could rest 10 minutes between the experiments in each set. We believe that resting is essential, so that fatigue wouldn’t distort the results. This time was also used to hide/unhide the loudspeakers.

From the measured data we focused our attention on the following:

---

1. The maximum unperceived distance between the audio and visual stimuli (ventriloquism effect);

2. The correctness on identifying a shift, this is if the user identified a shift but he pointed it as coming from another position;

3. The number of times that a shift was perceived when the sound was coming from the same place as the talking-head’s mouth, useful to identify users’ fatigue;

4. The values of calibration, so that they could be compared with other users’ calibration.

5.3.6 Results

**Maximum unperceived distance.** When the loudspeakers were hidden, users unnoticed greatest distances between the audio and visual stimuli. The average when the loudspeakers were visible was 22 (STD 9.8 cm), when hidden 37 (STD 30 cm) (see Fig. 5.21). Looking at single user values on both setups: 4 tolerated greatest distances when the loudspeakers were hidden, 1 tolerated equal distances and just 2 tolerated greatest distances when the loudspeakers were visible.

![Figure 5.21: Average and standard deviation of maximum unperceived distance when the loudspeakers were hidden and visible.](image)

**Confidence and distance error.** When the loudspeakers were hidden, users pointed where the sound was coming from with more confidence and less error. The average when the loudspeakers were hidden was 61 (CI 12 cm) for distance error and 111 (CI 19 cm) for circle diameter. When the loudspeakers were visible 73 (CI 8 cm) and 119 (CI 15 cm), see Fig. 5.23.
A paired T-Test showed that there were significant differences between both setups regarding distance error and circle diameter ($t_{35} = -4.978; p < 0.001$) and ($t_{32} = -3.051; p = 0.005$) respectively.

Looking at Fig. 5.23 we see that for a single user, both the distance error and circle diameter had lower values in the setup with the loudspeakers hidden.

**Fatigue.** According to Fig. 5.24 only one user exceeded the number of wrong shifts perceptions in the hidden loudspeakers setup. Nevertheless the sum of all users wrong shifts was nearly the same (8 for hidden and 9 for visible). In a Likert Scale with values from 0 to 5, the average was 2 for the question if fatigue was felt during the experiment.

**Calibration values.** According to the average of each user’s calibration, our 3-D sound system needed to shift its center 27cm to the left and 21cm down (see Fig. 5.25). This could have been due to numerous reasons related to the room acoustics and individual’s sound perception.
5.3.7 Discussion

Hiding the loudspeakers can improve the subjective perception of sound in a VE. We believe that this simple technique can really improve the way people perceive sounds in a Virtual Environment with a first order Ambisonics system. Not only helps reducing the imperfections intrinsic to every sound systems, but also helps creating a better illusion, i.e. a better ventriloquism effect.

After looking at the results we can say that, when the loudspeakers were hidden users unnoticed greatest shifts between the image and the sound. The average was 15cm greater (see Fig. 5.21), this meaning that the user tolerated greatest distances between sound and visual stimuli. Also the confidence and correctness were improved (see Fig. 5.22 and 5.23), which lets us say that users are influenced by the loudspeakers when judging where the virtual sound source is coming from.
Users' calibration average is useful to identify room acoustics problems. The values from each user' calibration are very useful so that we can calculate an average and therefore identify some problems that the room might have. There exist several tools and techniques to do this kind of acoustics corrections, but this method is cheap and effortless.

In our case, the average of users' calibration told us that our room was probably influenced by some sound reflections (due to its short dimensions), resulting in a shift of 27cm right and 21cm up.

5.3.8 Summary

In this experiment we evaluated the influence of hiding the loudspeakers to improve the maximum unperceived distance between audio and visual stimuli (ventriloquism effect). Our assumption was that the physical sound sources influence the belief that a virtual sound source is coming from the projected image. That was proved to be true.

In a VE our unconscious associates what we hear to the virtual image and also to the physical sound sources. Forbidding the users from knowing where those physical sound sources are, makes them tolerate greatest positioning errors between the image and the sound, this is an improvement on the subjective perception of sound. Also this simple, cheap and effortless technique improves users’ confidence and correctness, when judging where the sound is coming from.

We consider that every VEs, that use a sound system, should make an effort to hide the loudspeakers from the users’ field of view, so that they can have a better sound perception.
5.4 Summary of Experiments 1, 2 and 3

In Experiment 1 we proved that other physical modalities rather than hearing and vision (a typical VE), are so or even more important to enhance presence in Outdoor Virtual Environments. We combined two measuring methods (Slater’s \[12\] and EEG) and discovered that they complement each other. The gathered data suggested that tactile perception of wind is the most important stimulus in OVEs and that to feel Place Illusion users needed 3-D sounds. But most important we found out that our conclusions are particularly useful for designers interested in crafting similar Virtual Environments.

In Experiment 2 we proved that audio cues, especially if they are 3-D sounds, make navigation effortless and improve sense of orientation. From the three setups evaluated: video only projection, video and 2-D sound, video and 3-D sound, the first was the one with the worst results. We found that with 3-D sound, users chose the shortest paths; without audio cues, users lost orientation more often and with 3-D sound users rotate more than with 2-D. We also found that 3-D sound approaches real life behaviours to the Virtual Environment, therefore enhancing presence.

In Experiment 3 we proved that it is worth hiding the loudspeakers from the users’ field of view, to improve the subjective perception of sound in a VE. In the setup where the loudspeakers were hidden, users tolerated greatest distances between audio and visual stimuli. This simple, cheap and effortless technique improves users’ confidence and correctness, when judging where the sound is coming from. It also enables greatest positioning errors between the image and the sound (very useful because it doesn’t exist a perfect sound system).
In the beginning of this work, a deep analysis on the fundamentals of sound, psychoacoustics, spatial audio perception, 3-D audio spatialization challenges and spatial audio models was performed. Understanding the characteristics of the sound waves as well as the human who listens and the loudspeakers that reproduce them, was essential to evaluate the related work and discover what was missing so that this thesis could be a contribute for the community.

We divided the related work in two great areas: Virtual Environments and Sonification Techniques. First we were interested in looking at works that tried to evaluate presence. Secondly to others that evaluated the importance of visual and/or audio cues in VEs. After we paid special attention to experiments that used several physical modalities on their environments. Finally we looked deeply at almost every sonification techniques and their applications on VEs.

We found out that: the existent contributions lack on the analysis of the differences between using Stereo or 3-D sounds; there aren’t experiments with a focus on benchmarking the value of each physical modality to enhance presence in Virtual Environments; there aren’t any standards for mounting a 3-D sound system on an IVE and never any experiment had evaluated the impact of the visible distractors (like the loudspeakers) on sound localization or in the sense of presence.

In an effort to contribute to the community we carried out three different experiments. They were independent from each others but all together tried to answer to our initial question: “Is it worth having 3-D sound systems to enhance presence in VEs?”. After analysing others’ experiments we decided to use a first order Ambisonics system with 8 loudspeakers. For all the three experiments we used the same room. During this time we also mounted a replica of our simulation room in the island of Madeira, which currently serves as a ludic place for tourists and others experiment a travel through the Laurissilva forest of Funchal, Portugal.
In the first two experiments, we answered our initial question: 3-D sound proved to be essential not only to enhance presence (experiment 1) but also to improve navigation, orientation and way-finding (experiment 2). In the third experiment we proved that a simple, cheap and effortless technique can improve the subjective perception of sound. Amongst others, these were the most important conclusions from the three experiments:

Wind and 3-D sound are important to feel Place Illusion (experiment 1); Wind increases the Engagement in Outdoor Virtual Environments (experiment 1); Users’ Excitement doesn’t increase due to the activation of visual stimuli in OVEs (experiment 1); With audio cues, people get lost less often (experiment 2); With 3-D sound people chose the shortest paths more often (experiment 2); With 3-D sound people rotate more than with 2-D (experiment 2); With 3-D sound subconscious decisions are more effective (experiment 2); With 3-D sound people seem to be more comfortable to look around and pay attention to what’s surrounding (experiment 2); When the loudspeakers are hidden, people tolerate greatest distances between audio and visual stimuli (experiment 3); Hiding the loudspeakers, improves users’ confidence and correctness, when judging where the sound is coming from (experiment 3).

Proving that it is worth the effort to mount a 3-D sound system in a Virtual Environment, is rewarding. The results were enthusiastic and will hopefully be a reference to researchers on these subjects. In our point of view, these three experiments are very important to reinforce the fact that sound systems (2-D and especially 3-D) must not be a last minute add-in, but yes an integral part of the architecture of an Immersive Virtual Environment.
Our work studies 3-D sound and its importance on Virtual Environments to enhance presence. The three experiments that we did were essentially to encourage the use of a 3-D sound system, and for that we focused our attention on comparing environments with and without 3-D sounds. These kind of comparisons are essential to reinforce that not only all IVEs should use 3-D sounds, but that their design should depend on the sound-system and sonifications used.

We believe that there are numerous experiments that can be reused and truly contribute to the 3-D sound ascension, just by comparing their previous results when using a 3-D or a 2-D sound system or none. Localization, navigation and orientation are three important variables to analyse, but there are several others: collaborative tasks, sense of presence, emotions, task correctness, interactive gestures, multimodal interfaces, and so on.

It is easy to introduce 3-D sounds in any Virtual Environment and it is important to think on the sound system in the firsts architectures. But for the purpose of research, it doesn’t matter if we mounted a 3-D sound system in the beginning or in the end, there are always possibilities to evaluate the benefit that 3-D sounds brings to a VE.

As told before, there aren’t many experiments that report the good and bad between using different audio/visual configurations to perform a specific task. Experiments that wish to contribute in this area can start by comparing 2-D and 3-D sounds, but can also go a little further and compare other physical modalities like smell, taste or touch. These kind of evaluations are a need, so that some standards to mount VEs can be made.

Finally and very important, we look forward to see others doing the same machine-less ventriloquism experiment we did. We are convinced that our results are true, independently of the sound system or the simulation room used, but this premise can only get strong with some more identical results.
Bibliography


Procedures of Experiment 1

A.1 Instructions

You will experience a virtual journey through the sea and the Laurissilva Forest of Funchal, Madeira. First you will approach through the sea to the island and then you will travel right up to the forest. Each trial has a duration of 4.30 minutes. This room can simulate 4 different physical stimuli (hearing, vision, haptics and smell), that are intended to enhance your immersion in the Virtual Environment. When experiencing the Virtual Environment with all the 4 physical stimuli enabled, pay attention to the following (depending from the user number we showed the PI or Psi instruction):

- (PI) “What are the most important stimuli, to make me feel that I am no longer in this room, but in somewhere else?” Later (when some physical stimuli are disabled) we will ask you to get to that feeling, by enabling the stimuli you think essential for it. Please have in mind that we don’t expect you to completely forget that you are in the simulation room. The important thing is to identify the stimuli that have a higher influence, even though that you are in a simulation room.

- (Psi) “What are the most important stimuli to make me feel that, what is happening is very close to reality?” Later (when some physical stimuli are disabled) we will ask you to get to that feeling, by enabling the stimuli you think essential for it. Please have in mind that we don’t expect you to forget that everything that is happening is a simulation. The important thing is to identify the stimuli that have a higher influence, even though that everything that is happening isn’t real.

In this experiment you must remain sited in the chair during all the time and you can’t take off the Emotiv EPOC EEG headset. You are free to move your head through whatever position you wish as long as you remain sited. Also, feel free to ask for a pause whenever you feel tired.
A.2 Questionnaires

**General Questions** - Likert Scale (1 2 3 4 5)
- I was impressed with this Virtual Environment (4,2).
- I felt bored during the experiment (2,2).
- If I weren’t using the Emotiv EPOC headset my stimuli enumeration would be different (1,7).
- I felt sickness or fatigue (1,7).

**Virtual Environment Design** - Likert Scale (1 2 3 4 5)
- SoundScape was well designed (4,7).
- Images corresponded to the soundscape (4,1).
- Lamps corresponded to the soundscape (3,7).
- Smell corresponded to the soundscape (3,7).
- Wind corresponded to the soundscape (4,7).
- Heat corresponded to the soundscape (3,1).

**Stimuli** - Likert Scale (1 2 3 4 5)
- 2-D sounds are important (4,2).
- 3-D sounds are important (4,3).
- Images are important (4).
- Coloured Lamps are important (3,1).
- Different smells are important (4,2).
- Wind is important (4,7).
- Heat is important (3,3).

**Expectation** - Likert Scale (1 2 3 4 5)
- I expected to feel other wind intensities (2,1).
- I expected to feel other smells (2,8).
- I expected to hear other sounds (1,6).
- I expected to see other images (2,8).
- I expected to see other colours in the lamps (2,8).
Other Questions

• If you repeat this Experiment would you change the order of the stimuli?

• What did you dislike most on the Experiment?

• What would you like to change on the Virtual Environment?

• What would you like to add on the Virtual Environment that is not currently available?

• Would you like to leave any suggestion or comment regarding this experiment?
B.1 Instructions

You will be asked to navigate in a Virtual Environment, in order to find and mark collisions between pipes of different colours and between pipes and coral regions.

This VE represents an oil extraction location in the middle of the ocean, where some engineers are trying to figure out where to position the pipes. Since the simulation of maritime currents, influence the pipes positions over the time, it is likely that some pipes collide with each other or with coral regions in the first iterations.

Your job is to identify as much collisions as possible in 3 minutes time, so that the engineers can place the pipes in a position that will guarantee 0 collisions.

You will do this experiment in 3 different sets: video projection only, video projection and 2-D sound, video projection and 3-D sound. Before starting on each different setup, you will have the possibility to test and learn the different audio-visual cues available. You are encouraged to be always positioned in the middle of the room (inside the circle marked on the floor). Nevertheless you can move freely all around the simulation room. Please feel free to ask for a pause whenever you feel tired.
B.2 Questionnaires

General Questions - Likert Scale (1 2 3 4 5)
- I felt differences between 2-D and 3-D sound (4,2).
- I felt bored during the experiment (1,3).
- I felt sickness or fatigue (1,2).

Navigation Controls - Likert Scale (1 2 3 4 5)
- I liked the navigation control (2,7).
- Pointing mechanism was good (2,7).
- Navigation system forbade me to pay attention to the audio-visual stimuli (2).

Audio Cues - Likert Scale (1 2 3 4 5)
- The sounds sea, oil boat and other environmental sounds are important for orientation (3,7).
- The collision sounds are important for orientation (4,8).
- Sounds are important for navigation (4,7).
- 2-D sounds are important for navigation (3,7).
- 3-D sounds are important for navigation (4,8).

Visual Cues - Likert Scale (1 2 3 4 5)
- The graphics were well designed (4).
- I had difficulties to see where were the collisions because of the colors used (1,2).
- The blinking green aura of the closest collision is important for orientation (3,3).
- The radar is important for navigation (3,3).
- I would like to had more visual cues (3,2).

Other Questions
- What did you dislike most on the Experiment?
- What would you like to change on the Experiment?
- What would you like to add on the Experiment that is not currently available?
- Would you like to leave any suggestion or comment regarding this experiment?
Procedures of Experiment 3

C.1 Instructions

In this experiment you must remain sited in the chair during all the time. You are free to move your head through whatever position you wish as long as you remain sited.

The talking head will always be placed in the middle of the screen. In each new shift, the sound source can be shifted from the head's mouth or not.

Please have in mind that we are studying the Ventriloquism Effect (the maximum unperceived distance between the visual and audio stimuli), therefore you must have your eyes open when answering YES or NO. Also, it is important that you listen and see the entire loop (one, two, three, four, five, six, seven, eight, nine, ten, zero) in each 3 shifts, I will remind you.

There is no right or wrong answers, since sound perception differs a lot from human to human (especially in altitude and behind us). Feel free to ask for a pause whenever you feel tired.
C.2 Questionnaires

General Questions - Likert Scale (1 2 3 4 5)

- With the curtains the sound was different as without it (2,71).
- I felt bored during the experiment (2).
- I felt sickness or fatigue (2).
- Without curtains, I looked a lot to the loudspeakers in order to help my answer (3).

Behavioural Questions - Likert Scale (1 2 3 4 5)

- I felt confident when saying that the system was calibrated (4,1).
- I had troubles to maintain my eyes open when answering Yes or No (1).
- I felt confident when answering Yes or No (2,9).
- I felt that sound was never in the same position (2,9).

Other Questions

- What did you dislike most on the Experiment?
- What would you like to change on the Experiment?
- What would you like to add on the Experiment that is not currently available?