UNIVERSIDADE CATÓLICA PORTUGUESA

THE ROLE THAT SOUND SPATIALIZATION PLAYS IN IMPROVING PERFORMANCE IN AN INTERACTIVE INSTALLATION: STUDY OF THE CORRELATION BETWEEN GESTURE AND LOCALIZATION OF SOUND SOURCES IN SPACE

Thesis submitted to the Portuguese Catholic University for the Doctoral Degree in Science and Technologies of the Arts – Interactive Art

by

Diogo Leichsenring Franco

ESCOLA DAS ARTES

March 2018

UNIVERSIDADE CATÓLICA PORTUGUESA

THE ROLE THAT SOUND SPATIALIZATION PLAYS IN IMPROVING PERFORMANCE IN AN INTERACTIVE INSTALLATION: STUDY OF THE CORRELATION BETWEEN GESTURE AND LOCALIZATION OF SOUND SOURCES IN SPACE

Thesis submitted to the Portuguese Catholic University for the Doctoral Degree in Science and Technologies of the Arts – Interactive Art

By Diogo Leichsenring Franco

Supervised by Prof. Doutor Álvaro Manuel Mendes Barbosa

ESCOLA DAS ARTES
March 2018

Abstract

The main objective of this research work is to study the correlation between gesture and localization of sound sources in space within the framework of interactive installations, based on theories of hearing and gesture.

We have therefore chosen the experimental method by developing an interactive installation with which we carry out three different experiments, in which a subject's hand is tracked by a Microsoft Kinect depth camera (motion capture) and a deictic gesture is used to trigger recorded music sounds and identify their localization in the horizontal plane. Thus, we manipulate the direction of sound and we measure the percentage of correct perceptual sound source localizations resulting from the participant's responses in an Inquiry Mode Questionnaire in comparison with the actual directions of the gesture and perceptual sound sources provided by software.

Descriptive and inferential statistics is applied to the collected data. The main results show that it is easier to define the origin of sound and that auditory perception is more accurate when its incidence is frontal in the horizontal plane, just as sound source localization theory predicts. Whereas 86.1% of all volunteers consider that their gesture coincides with the origin of sound in experiment 1, in which the use of their gesture in a certain direction produces a sound from that direction, only 58.1% admit the same in experiment 3, in which the same gesture is used to identify the system-predetermined localization of a perceptual sound source in an angle of 260^{o} around a subject. At least 55.9% of all participants do not perceive that their gesture cannot coincide with the origin of sound in experiment 2, since sound is produced from the opposite surround direction, which seems to demonstrate that, when sounds are produced frontally or from the back and a person has the task of controlling their motion with a deictic gesture at the same time, his or her ability to identify the origin of sound generally diminishes, in addition to the already well-known reduced ability to identify it when it is in the median plane, if the head is not rotated.

We therefore conclude that there is a relatively high correlation between gesture and localization of sound sources in space, but this is not as perfect as it could be owing to the limitations of the human auditory system and to the natural dependence of head movement on gesture.

Keywords: Sound spatialization, sound source localization, sound, gesture, motion capture, interactive installation

Resumo

O objectivo principal deste trabalho de pesquisa é o de estudar a correlação entre gesto e localização de fontes sonoras no espaço, no âmbito das instalações interactivas, com base nas teorias da audição e do gesto. Na ocasião em que começámos a nossa investigação, verificámos que havia vários estudos que abordavam os assuntos "gesto" e "localização de fontes sonoras" de diversas maneiras: 1) de forma independente um do outro e/ou noutros contextos distintos dos das instalações interactivas, como por exemplo em Blauert (1997), Pulkki (1999) Pulkki & Karjalainen (2001), Pulkki (2001a), Bates et al. (2007), Hammershøi (2009), McNeill (1992), Coutaz & Crowley (1995), Choi (2000), Cadoz & Wanderley (2000), Nehaniv (2005), Campbell (2005), ou Godøy & Leman (2010); 2) de um ponto de vista mais técnico, como por exemplo em Harada et al. (1992), Jensenius et al. (2006), Marshall et al. (2006), Schacher (2007), Neukom & Schacher (2008), Zelli (2009), Marshall et al. (2009), Bhuiyan & Picking (2009), ou Schumacher & Bresson (2010); ou 3) de um ponto de vista mais artístico, como em Bencina et al. (2008) ou Grigoriou & Floros (2010). Havia, no entanto, muito poucos estudos a envolver ou a abordar ambos os assuntos e a analisar de maneira conjugada as suas relações de um ponto de vista mais perceptual, como por exemplo em Gröhn (2002), de Götzen (2004) ou Marentakis et al. (2008). Foi esta última perspectiva que decidimos seguir e que aqui exploramos.

Desta forma, optámos pelo método experimental, aplicando um desenho de medidas repetidas e desenvolvendo uma instalação interactiva com a qual realizamos três experiências diferentes, em que a mão de um sujeito é rastreada por uma câmara de profundidade Microsoft Kinect (captação de movimento) e um gesto díctico é usado para activar sons de música gravada e para identificar as suas localizações no plano de escuta horizontal. Assim, manipulamos a direcção do som e medimos a percentagem de localizações de fontes sonoras perceptuais correctas, resultante das respostas dos participantes num Inquérito Por Questionário em comparação com as direcções reais do gesto díctico e das fontes sonoras perceptuais fornecidas pelo software que utilizamos no nosso trabalho. Para população-alvo pensámos em pessoas com conhecimentos musicais e pessoas com poucos ou nenhuns conhecimentos musicais, o que nos levou a solicitar a um grande número de pessoas a sua participação voluntária, anónima e sem constrangimentos no nosso estudo. Isso foi levado a cabo sobretudo através do envio de correio electrónico para amigos, para estudantes de diferentes áreas a frequentar e para colegas a trabalhar na Escola de Artes da Universidade Católica Portuguesa (EA-UCP), na Escola Superior de Música e Artes do Espetáculo do Instituto Politécnico do Porto e na Academia de Música de Espinho. Para além disso, foi também crucial falar-se com amigos e familiares e informar tantas pessoas quanto possível sobre a nossa investigação, através da colocação de cartazes informativos nas paredes dos corredores da Universidade Católica, alguns dias antes de as experiências terem sido realizadas no Laboratório de Captação de Movimento da EA-UCP.

Por fim, é efectuada uma análise estatística descritiva e inferencial dos dados recolhidos. Os principais resultados apontam no sentido de ser mais fácil definir a origem do som quando a sua incidência é frontal no plano de escuta horizontal, para além de a percepção auditiva ser mais precisa nessa direcção, tal como a teoria da localização de fontes sonoras prevê. Enquanto 86.1% de todos os participantes consideram que o seu gesto díctico coincide com a origem do som na experiência 1, em que o uso desse gesto numa determinada direcção faz despoletar um som proveniente dessa direcção, apenas 58.1% admitem o mesmo

na experiência 3, em que o mesmo gesto é usado para identificar a localização de uma fonte sonora perceptual predeterminada pelo sistema num ângulo de 260º em torno de um sujeito. Esta última percentagem parece dever-se ao facto de a maior parte dos sons ser produzida a partir de direcções laterais na experiência 3, tendo a posição da cabeça voltada para a câmara como referência. Pelo menos 55.9% de todos os voluntários não percebem que o seu gesto não poderia ter coincidido com a origem do som na experiência 2, já que o som é produzido a partir da direcção envolvente oposta. Este facto parece demonstrar que, quando os sons são produzidos frontalmente ou de trás e uma pessoa tem a tarefa de controlar os seus movimentos com um gesto díctico ao mesmo tempo, a sua capacidade para identificar a origem do som é, em geral, ainda mais baixa, para além da já conhecida capacidade reduzida para identificá-la quando o som se encontra no plano mediano, se a cabeça não for rodada.

A maior parte dos participantes sente um controlo imediato sobre o som nas experiências 1 e 2, mas os tempos estimados pelos próprios são bastante superiores aos aproximadamente 650 milissegundos necessários para o ser humano ouvir e reagir a um som na nossa instalação interactiva.

Descobrimos também que o tempo médio necessário para localizar sons com o uso de um gesto díctico na nossa experiência 3 é de cerca de 10 segundos, o que corresponde a um tempo bastante mais longo do que os 3 segundos que supusemos. Para além disso, os voluntários fazem em média 2 tentativas para localizar sons com os seus gestos dícticos, tendo a necessidade de ouvir apenas uma vez em média cada som na íntegra para o localizar.

Os desvios à esquerda e à direita efectuados pela maior parte dos participantes relativamente às direcções verdadeiras do som, quando estes tentam identificar as localizações predeterminadas pelo sistema das fontes sonoras perceptuais com os seus gestos dícticos na zona periférica do corpo, são em média de 7.97º e -7.19º, respectivamente. Desta forma, o desvio médio absoluto é de 7.76º. Comparando esses desvios com aqueles levados a cabo pelos participantes usando a mão esquerda (desvios de 6.86º para a esquerda e -6.35º para a direita das direcções verdadeiras do som) e com aqueles usando a mão direita (desvios de 8.46º para a esquerda e -7.38º para a direita das direcções verdadeiras do som), concluímos que os resultados são bastante parecidos entre si.

Descobrimos que a maior parte dos voluntários estima um tempo muito mais longo do que os 2 segundos que supusemos experimentalmente para entender cada uma das três experiências. Para além disso, esse tempo estimado pelos participantes diminui da primeira para a última experiência, aparentemente devido à familiarização, conscientemente provocada por nós através da mesma sequência de realização das experiências imposta a cada participante, com o nosso sistema interactivo, embora considerem ter entendido cada uma das três experiências rapidamente.

Acresce que a maioria dos voluntários interage facilmente com a nossa instalação e concorda que o gesto sugerido por nós foi adequadamente seleccionado para qualquer uma das três experiências.

Também constatamos que os participantes consideram a resposta do sistema ao gesto como sendo imediata nas nossas três experiências, ou seja, estimam cerca de 1 segundo, o que é consistente com o resultado da medição da latência do sistema de cerca de 470 milissegundos.

Além disso, verificamos que a maioria dos voluntários se sente envolvida pelo som na nossa instalação interactiva usando Ambisonics Equivalent Panning.

Portanto, concluímos que, usando uma instalação interactiva como a nossa com um público-alvo semelhante àquele que tivemos, há uma correlação relativamente elevada entre o gesto e a localização de fontes sonoras no espaço, mas que esta não é tão perfeita como poderia ser devido às limitações do nosso sistema auditivo e aparentemente à dependência natural do movimento da cabeça do gesto. Assim, parece que a espacialização sonora pode melhorar o desempenho numa instalação interactiva, mas de forma moderada. Mesmo assim, defendemos que um sistema como o nosso pode vir a ser aplicado com vantagem em domínios diversos como os que apresentamos como exemplos.

Palavras-chave: Espacialização sonora, localização de fontes sonoras, som, gesto, captação de movimento, instalação interactiva

Acknowledgments

Firstly, I would like to thank the scientific coordinator of the PhD course in Science and Technology of the Arts, taught at the School of Arts of the Portuguese Catholic University in Porto (EA-UCP), Portugal, Paulo Ferreira-Lopes, for the opportunity I had to put into practice the knowledge that I have acquired not only during the course, but also throughout my student, professional, and personal life, for the possibility of being able to devote myself to the present investigation, and for the creation of favourable conditions for the realization of this type of study. The Motion Capture Laboratory that was put at my disposal and the loan of equipment (cf. chapter 3), such as eight loudspeakers, a TV set, cables, among others, were absolutely decisive. In this sense, I would also like to express my gratitude to Pedro Oliveira, who helped me with this process.

I would like to express my deep gratitude to my doctorate advisor Álvaro Manuel Mendes Barbosa, who kindly agreed to accept the supervision of this study, to indicate objectively the course to follow through his advice and suggestions, which greatly contributed to enhance this investigation, and who helped me in my worst personal moments.

For their knowledge and teaching of subjects with such a rich and important contents for me and for the imparting of knowledge without any kind of constraints, I am most thankful to all the teachers I had during the curricular year of the PhD course in the EA-UCP, hoping not to forget anyone: Aaron Williamon (Research Methods), Ajay Kapur (Human-Computer Interaction), Álvaro Mendes Barbosa (Human-Computer Interaction), Andre Bartetzki (Programming Languages), Andy Farnell (Human-Computer Interaction), Antonio Camurri (Interactive Art), António Ramires (Virtual and Augmented Reality), Arturo Castro (Programming Languages), Beatriz Sousa Santos (Perception and Cognition), Bill Verplank (History and Aesthetics), Carlos Sena Caires (Interactive Art; Virtual and Augmented Reality), Christa Sommerer (Interactive Art), Christoph Seibert (Perception and Cognition), Claudia Robles (Interactive Art), Cristina Sá (History and Aesthetics), Daniel Tércio (Interactive Art), Daniela Coimbra (Research Methods), David Rokeby (Interactive Art), Douglas Edric Stanley (Interactive Art), Guilhermina Castro (Research Methods), Ido Iurgel (Virtual and Augmented Reality), Isabel Galhano Rodrigues (Perception and Cognition), Ivan Franco (Virtual and Augmented Reality), João Paulo Costeira (Virtual and Augmented Reality), Jorge Cardoso (Programming Languages; Virtual and Augmented Reality), Kònic Thtr (Rosa Sanchez and Alain Baumann) (Interactive Art), Luís Gustavo Martins (Research Methods; Perception and Cognition; Programming Languages), Luís Filipe Teixeira (Virtual and Augmented Reality), Marcelo Wanderley (Human-Computer Interaction), Martin Kaltenbrunner (Interactive Art), Natalie Woolf (Interactive Art), Paulo Ferreira-Lopes (Perception and Cognition; Interactive Art), Peter Higgins (Interactive Art), Suguru Goto (Interactive Art), Thomas A. Troge (Research Methods), Victor Gama (Interactive Art), Xavier Serra (Perception and Cognition), and Yann Orlarey (Research Methods).

I would like to thank my classmates at EA-UCP for the shared experiences and ideas during the PhD course, too, hoping not to forget anyone as well: André Gonçalves, André Perrotta, Carla Almeida, Diana Cardoso, Fernando Martins, Fernando Monteiro, João Castro Pinto, João Alves de Sousa, João Tiago Silva, Jorge Gonçalves, Jorge Manuel Mota, José Vasco Carvalho, Luís Sarmento Ferreira, Luís da Costa Monteiro, and Samuel van Ransbeeck.

I owe a great debt of thanks to my colleagues at the School of Music and Performing Arts of the Polytechnic

Institute of Porto (ESMAE-IPP), Portugal: Daniela Coimbra for her assistance in the preparation of the Inquiry Mode Questionnaire (InQ) (see appendix A); Inês Vicente for her suggestions relatively to the InQ and practical experiments; and Rui Coelho for his advices on the InQ as well.

I am also most indebted to the volunteers who accepted to participate in my practical experiments and without whom it would not have been possible either to carry out this study: André Baltazar, André Gomes, André Perrotta, António Franco, Beatriz Garcia, Bruno Pereira, Celine Valente, Daniel Pinheiro, Eduardo Sousa, Florian Pertzborn, Francisco Pereira, Gustavo Almeida, Igor Fonseca, Inês Franco, Inês Moura, Inês Vicente, Joana Gomes, João Lóio, Jonas Pinho, Jorge Cardoso, José Vasco Carvalho, Laura Franco, Luís Sarmento Ferreira, Luís Freitas, Luís Gustavo Martins, Luís Silva, Manuel Cunha, Mariana Rêgo, Miguel Gonçalves, Miguel Vieira, Nuno Pinto, Regina Castro, Ricardo Leão, Rui Coelho, Samuel Guimarães, Samuel van Ransbeeck, Sofia Lourenço, Telmo Marques, Tiago Bôto, Tiago Sarmento, Tiago Vouga, Tulio Pezzoni, Victor Rodriguez, and Vitor Joaquim.

I am deeply grateful to the former PhD professor at the Faculty of Arts of the University of Porto (FLUP), Portugal, António Franco, for his reading and the very precise and insightful English correction and criticism of the text. The English teacher graduated from the University of Cambridge, Laura Franco, and the English teacher at the Music Academy of Espinho (AME), Portugal, Jonathan Fox, were also valuable readers and correctors of the text at the beginning of this journey, and so I thank them as well.

I want to express my thanks to the director of the Music Academy of Espinho (AME), Alexandre Santos, and the technical producer, Manuel Cunha, for the loan of a sound level meter (see chapter 3).

A special word of thanks also goes to Igor Fonseca, Manuel Cunha, and Ricardo Leão for motivation and professional support.

A big thankyou to Mestre Paulo Araújo who helped me to recover from a rather complicated health problem that arose at the time when I was finishing this document.

I would like to emphasize that it would not have been possible to carry out my study without the financial support, which was granted in the academic years 2012/2013 and 2013/2014 with the consent of the president of the Polytechnic Institute of Porto (IPP), Rosário Gambôa. In this sense, I am also extremely grateful to the former president of ESMAE-IPP, Francisco Beja, who sadly died recently, for the constant encouragement and support in the advanced training process and in the development of part of my professional skills.

Finally, I am deeply grateful to my wife, Laura, my children, Inês and Diogo, and my parents, António and Irmtraud Franco (the latter having unfortunately passed away almost at the beginning of this long task), for the stimulus, endless love and moral and material support, and their endless patience and understanding.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Context and Motivation

Research within the scope of subjects such as the human hearing system, sound, sound sources, sound source localization, sound spatialization, gesture, computer science, real-time controllers, mapping, motion capture, interactive applications, among many others, has been carried out in numerous centres around the world, such as the Center for Computer Research in Music and Acoustics (CCRMA) in Stanford (California, USA), the Center for New Music & Audio Technologies (CNMAT) in Berkeley (California, USA), the Center for Research in Electronic Art Technology (CREATE) in Santa Barbara (California, USA), the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) in Montreal (Quebec, Canada), the Centro de Investigação em Ciência e Tecnologia das Artes (CITAR) in Porto (Portugal), the Digital Media & Arts Research Centre (DMARC) in Limerick (Ireland), the Input Devices and Music Interaction Laboratory (IDMIL) in Montreal (Quebec, Canada), the Institut de Recherche et Coordination Acoustique/Musique (IRCAM) in Paris (France), the Institut für Elektronische Musik und Akustik (IEM) in Graz (Austria), the Sonic Arts Research Center (SARC) in Belfast (UK), and the Zentrum für Kunst und Medientechnologie (ZKM) in Karlsruhe (Germany), just to mention a few. A more complete list of research centres can be found, for example, at `http://smcnetwork.org/resources/centers` (visited on 12/12/2017).

With regard to gesture and sound, they are often influenced by each other (Odowichuk, 2012, pp. 1, 6), as, for instance, gesture can be used, on the one hand, to produce sound, and sound can be followed, on the other hand, by sound-accompanying gestures (Jensenius et al., 2010, p. 30). Although sound localization, or more specifically sound source localization, as it is called in Psychoacoustics, and gesture have already been thoroughly studied respectively by authors like Blauert (1997) and Moore (2013), and McNeill (1992) and Godøy & Leman (2010) –, whose respective published works we have read with great interest for personal and professional reasons –, among others, but somehow separately from each other and/or in different contexts other than relating both from a more perceptual point of view, we, therefore, felt the need to investigate the relationship between these two subjects from this latter perspective, because at the time when we began our investigation there were too few studies focusing on this aspect. Thus, the aim was to study the correlation between gesture and localization of sound sources in space, based on theories of hearing and gesture, known

until then, in order to perceive the role that sound spatialization plays in improving performance in an interactive installation.

1.2 Main Research Question and Hypothesis

Our main research question is: ”What sort of relationship is there between a deictic gesture and localization of perceptual sound sources in the horizontal plane in space?” Accordingly, the main hypothesis to be tested is: ”There is a significantly high relationship between a deictic gesture and localization of perceptual sound sources in the horizontal plane in space.”

Thus, our main contributions towards the improvement of scientific knowledge in the context of gesture and sound spatialization in interactive installations, in general, and in our own interactive installation, specifically developed for the purposes we propose, are as follows:

1. We investigate the correlation between a deictic gesture and localization of perceptual sound sources in the horizontal plane in space;
2. We investigate the impact caused on a human being’s sound source localization capacity by his or her deictic gesture, when he or she has control over the motion of perceptual sound sources in the horizontal plane in space with this gesture;
3. We investigate the impact caused on a human being’s sound source localization capacity by his or her deictic gesture, when he or she uses this gesture to control the motion of perceptual sound sources in the horizontal plane in space that are originated from the opposite surround direction;
4. We investigate to what extent a human being is capable of localizing static perceptual sound sources in predefined directions in the horizontal plane in space with the use of his or her deictic gesture.
5. We evaluate the impact of our interactive installation on human beings, that is, of the selected sound reproduction system and of the suggested gesture on human beings.

To achieve this aim, we cannot do without resorting to information from diverse sources, that will be dealt with in chapter 2, a chapter that in consequence has turned out to be rather long. We recognize this fact, but we are above all interested in extracting the most relevant elements from that information, in order to come to a consistent, well-formed and solid conclusion.

1.3 Outline of the Thesis

Our thesis follows an organization somehow similar to that proposed by Azevedo (2004, p. 15), and it consists, therefore, of five chapters, the first of which is the present one.

More than a mere review of the literature, chapter 2 is intended to be a systematic and pedagogical summary of, or review on, distinct points of view of different authors on topics directly related to the subject of the thesis. Thus, the terms *sound*, *physical sound*, and *perceptual sound* are defined first. In addition, the most

common listener independent and listener dependent sound attributes are addressed and dealt with. Then, auditory test signals, physical, perceptual, static, and moving sound sources, sound spatialization, physical space, perceptual space, and sound reproduction systems are thoroughly described. The human hearing system, the head-related transfer function, the cone of confusion, and the sound source localization process are also described and explained afterwards. Gesture and its relationship with sound and sound spatialization are then taken into account as well. Finally, gestural controllers, mapping, and interactive installations, where gesture and sound are important aspects involved, are addressed.

In chapter 3, we describe, explain, and justify the various options taken to develop and install our interactive installation, which was used for our research purposes: the selected room, equipment, sound reproduction system, test sounds, type of gesture, and software architecture. In addition, we introduce our experimental methods, as well as complementary research questions and hypotheses to the main hypothesis, already presented in section 1.2, taking some considerations on the reliability and validity of the used tools into account. The test sounds and the complete code can be found in a DVD-ROM, in appendix E.

Chapter 4 deals with data results and analysis. The method of analysis of data we follow is presented and justified first. Afterwards, we describe the target population, the accessible population, the sample definition process, and the sample of participants used, from which two different groups, one with and the other without any musical knowledge, emerge, in order to test hypothetical deviations in the results of the analysed data. In addition, we describe and discuss the obtained results by reference to all hypotheses we formulated, and information about the confirmation or non-confirmation of the initial previsions is given. Finally, the usefulness of an interactive installation like ours in practical life is also described. The data we collected and all resulting statistical information we worked on can be found in the DVD-ROM, in appendix E.

In chapter 5, we present the final conclusions. Limitations of our interactive installation and process, the main contributions, and the most significant results of our research are presented in this chapter, at the same time suggesting possible directions for future work and the use of our interactive system in many other possible contexts.

Chapter 2

Theoretical Foundations

Chapter 2 is actually a double chapter, with one part about sound and the other about gesture. We could have basically put together a third chapter with this second aspect, but we have chosen to include gesture in this same chapter, because we consider it, for our purposes, to be closely related to the matter of sound, not treating it in a different part.

2.1 Sound

For the human being, *sound* can currently be approached from different points of view. For example, from the acoustical or physical, perceptual, musical, immersive, or conceptual (design) viewpoint, among others. In the following paragraphs, some standpoints of different authors are by way of introduction presented briefly which help to describe these approaches. It should, however, be clear that we are not going to define or thoroughly explore every point of view, but we will rather fundamentally concentrate our attention on the first two aspects.

Therefore, in the first case, as a physical phenomenon, sound is understood as being related to the sound source (Henrique, 2007, p. 6) which causes a mechanical disturbance of the medium of propagation, which may be air, or a solid, liquid, or another gas (Howard & Angus, 2001, p. 1). Further details from this perspective are presented in section 2.1.1.

As a psychophysical phenomenon and from the perceptual angle, sound refers to the sensation that it causes to the human being, that is, sound is related to the way the human ear and brain absorb and interpret it (Sonnenschein, 2001, p. xxi). See also section 2.1.2.

From a musical point of view, sound is a means whereby ideas of musical structure and harmony can be expressed (Gibbs, 2007, p. 8).

Nowadays, technology and the increasing noticeable presence of noise in everyday life, from motorized vehicles, machines, phones, toys, among others, have come to compel increasingly the post-industrial revolution listener to a less attentive or to a less concentrated way of listening. Consequently, the listener is being urged to the participation or to the subjective experimentation of sound. This means that the listener is testing less and less his or her imaginative identification capability. From the immersive point of view, the listener is therefore immersed in a medium of sound propagation which is more and more contaminated by noise. Today,

music, which was once an object of a studiously quiet listening space (R. Brown, 2010, p. 2), is considered to be a part of the general noisy environment of a specific place as a result of its diffusion from sound sources, such as musical instruments or electroacoustic systems (headphones or loudspeakers). This music can reach the listener, for example, by escaping through doors and windows of buildings.

Conversely, the noisy environment may also be admitted as part of music. It is therefore no wonder that sound effects and recordings of the daily sounds are commonly used in multiple musical genres. The Italian painter, composer and inventor of the *Intonarumori* or Noise Intoner, Luigi Russolo (April 30, 1885 - February 6, 1947)[1], proposed early in 1913 in his manifesto *L'Arte dei Rumori* (*The Art of Noise*), that one should not make any distinction between sounds of instrumental or musical origin and sounds from industry or, generically, from the surrounding space (Gibbs, 2007, p. 23). In the 1970's, Brian Eno (born May 15, 1948)[2] proposed the idea of creating sound as an integral part of the sound environment of one location, instead of being heard in an isolated way. The concept of ambient music was born (Gibbs, 2007, p. 39).

From a conceptual or design point of view, sound is regarded as contents to be used in a theatrical performance, in a concert or in any other performance. It is determined by aesthetic decisions or artistic concepts. It involves the abstraction of sound itself from the immersive environment in order to develop a conceptual working model, although one returns to this environment afterwards (R. Brown, 2010, p. 6).

As already mentioned, we will focus on the points of view that, for example, Bregman (1990, p. 10), Everest (2001, p. 1) and Henrique (2007, p. 6) report about *sound*, that is, sound has two meanings depending on whether it is considered as a physical phenomenon or as a psychophysical phenomenon. In this sense, Toole (2008, p. 4) points out that the inclusion of sound as both a physical event and a perceptual event is remarkable, because it answers the riddle: "If a tree falls in a forest and nobody is there to hear it, does it make sound?" This can be interpreted as being basically a sort of alternative formulation of Protagoras' "Man is the measure of all things" (5th century B. C.). So, sound is created when the tree falls, but if the sound does not reach any ears, then there can be no perception of the physical event. In addition, Blauert (1997, p. 1) refers that there is no perception without both a subject and an object. For that reason, we differentiate between physical sound and perceptual sound, just like Klapuri & Davy (2006, pp. 302–303) did, a differentiation that will be explored in sections 2.1.1 and 2.1.2, respectively.

2.1.1 Physical Sound

Blauert (1997, p. 2) argues that human beings are primarily visually oriented and that the concepts and descriptions are based primarily on visual objects. As a consequence and from the physical point of view, he reports that the notion of *sound* was defined in 1959 by the German Standard DIN 1320 firstly from a visual perspective as mechanical vibrations and waves of an elastic medium. Therefore, *sound* describes the physical aspect of the phenomenon of hearing, and terms like *sound wave*, *sound signal* and *sound source* describe physical phenomena which are characteristic of sound events. Sound events consist of one or more sound signals, which may be the same or different, radiated into space by one or more sound sources (see section 2.2) at different positions in space (Blauert, 1997, p. 22). The acoustic contents of these sound events

[1] Retrieved 31/01/2016, from `https://en.wikipedia.org/wiki/Luigi_Russolo`
[2] Retrieved 31/01/2016, from `https://en.wikipedia.org/wiki/Brian_Eno`

informs us about physical happenings and relates to the physical cause (Bregman, 1990, p. 10).

Following the same line, Rossing (1990, p. v) defines *sound* as the disturbance in a medium, and Everest (2001, p. 1) notes that it can be defined as a wave motion in air or other elastic media (stimulus). The ANSI/ASA Acoustic Terminology S1.1.-1994 document (revised in 2004) defines *sound* as an oscillation in pressure, stress, particle displacement, particle velocity, etc., in a medium with internal forces (e.g., elastic and viscous), or the superposition of such propagated oscillations (Alvarsson, 2013, p. 2), emphasizing that there is always the need for a material medium for sound to propagate (Salema & Ferreira, 2009, p. 29).

In order to clarify that the context in which the word *sound* is used is the physical one, Klapuri & Davy (2006, pp. 302–303) introduce the term *physical sound* and define it as the vibration of the medium itself, that is, sound as a physical entity.

Therefore, the way this disturbance travels through a physical medium like air is described as expanding longitudinal pressure waves which are liable to a reduction in their level as they spread (Malham, 1998, p. 167). Furthermore, these waves are absorbed, reflected, refracted, diffracted, or scattered by objects, and the higher frequencies are increasingly lost with distance remarkably due to the absorption of water vapour in the air and, to a lesser extent, due to air density and temperature (Pedroso de Lima, 2012, pp. 170–171) (see section 2.1.3.1).

As Schacher (2007, p. 358) reports, sounds have even obtained the condition of object from the moment when swarming behaviours were introduced artificially by Bisig et al. (2007) to create emergent structures, which gives them different properties from the traditional acoustical ones.

On the other hand, sound waves may be compared with each other. This can be accomplished mathematically using two methods: by the correlation method or by the coherence method (Guevara & Corsi-Cabrera, 1996, pp. 146–148). The former is more sensitive to phase (see section 2.1.3.1) and polarity changes which occur between signals, that is, it is the recommended method in situations where it is desired to measure the waveform similarity and the time coupling between two signals, independently of amplitudes (see section 2.1.3.1). The correlation quickly degrades with an increase in noise, and its values vary between -1 and 1. The coherence method in turn gives information on the stability of the similarity, or on the maintenance of the relationships with respect to power asymmetry and phase, between two signals. The coherence values vary between 0 and 1, since signals are squared, which means that the polarity information is lost. If either power or phase changes in one of the signals, that is, if the relation between the two signals is not maintained, the coherence value is affected, so that a coherence of 0 means that the two signals are totally unrelated, and a coherence of 1 means that the two signals have the same relationship at a given frequency. Coherence is not affected with an increase in noise.

Furthermore, the normalized cross-correlation function can be used to mathematically determine the similarity between these waveforms (Blauert, 1997, pp. 201–202; Brutti et al., 2008, p. 69). As a result, two physical sounds are said to be coherent if they are identical or if they differ only in one or more of the following ways:

1. Level differences which are independent of frequency;

2. A pure delay, which is independent of frequency, of one sound relatively to the other;

3. A phase difference of 180º between both sounds.

Partially coherent or incoherent sounds are those which differ from the above. The degree of coherence of sounds at the physical sound sources (see section 2.2.1) is usually different from that at the ear inputs. Partially coherent sound signals are always perceived at the ear inputs regardless of whether coherent or incoherent sounds are produced at the sound sources, except if sound is presented over headphones or by two loudspeakers which are symmetrical relatively to the median plane (Blauert, 1997, p. 240) (see section 2.3.2).

2.1.2 Perceptual Sound

From the perceptual point of view, Rossing (1990, p. v) remarks that the word *sound* describes an auditory sensation in the ear, and Blauert (1997, p. 2) observes that the German Standard DIN 1320 only includes the frequency range of human hearing (16 Hz to 20 kHz) to describe what is perceived auditorily.

The ANSI/ASA Acoustic Terminology S1.1.-1994 document (revised in 2004) defines *sound* in perceptual terms as the auditory sensation evoked by the oscillation referred to in 2.1.1 (Alvarsson, 2013, p. 2).

Blauert (1997, p. 2) and Pulkki (2001b, p. 3) state that the adjective *auditory* relates to the perception of physical phenomena and that the terms *auditory object* and *auditory event* describe an internal percept which can be caused by a sound source that emits sound signals (or sound waves).

However, auditory events are not always caused or determined by sound events (see section 2.1.1). When the acoustic nerve is artificially stimulated or when a ringing in the ears occurs due to certain disease conditions, like tinnitus, then the auditory events are not caused by, or connected with, sound events (Blauert, 1997, p. 3). As a consequence, sound events and auditory events differ from each other in time, space and in other qualities. In addition, Solmer (1999, p. 224) remarks that in this sense sound is a single personal experience, which does not have any intermediary.

According to Blauert (1997, p. 242), a single auditory event of relatively small extent is perceived at the median plane (see section 2.3.2) when the ear input signals are totally coherent (see section 2.1.1). When the degree of coherence is diminished, the area over which components of the auditory event are found increases. The centre of the auditory event remains at first unchanged until two spatially separated auditory events are perceived, one at each ear. Thus, the human auditory system is able to identify components of partially coherent ear input signals. Furthermore, it is able to form a separate auditory event for each one of these components. The localization accuracy of auditory events decreases if incoherent components are present as well. If the ear input signals do not include coherent components at all, a separate auditory event is formed from each ear input signal. These observations are valid for sound presented over headphones or in a free sound field over loudspeakers, that is, in an environment in which there are no sound reflections.

In order to clarify that the context in which the word *sound* is used is the perceptual one, Klapuri & Davy (2006, pp. 302–303) introduce the term *perceptual sound* and define it as a psychological entity that corresponds to what human ears perceive as one sound. Thus, many perceptual sounds produced by many musical instruments may be heard through a monaural loudspeaker. In this case, the number of physical sound sources (see section 2.2.1) is just one, against the number of perceptual sound sources (see section 2.2.2), which is

greater than one. Furthermore, what is heard as one sound depends on time, place, occasion, and even attention (see section 2.5.6).

2.1.3 Relevant Sound Attributes

Sound attributes are distinguished as having objective and/or subjective qualities. To be before an objective quality means that the sound attribute under study can be analysed by means of adequate equipment or apparatus, which yields rigorous numerical measurements of its physical properties that are independent of the listener (Henrique, 2007, p. 169). On the other hand, although a subjective sound attribute might also be measured, which is the fundamental goal of psychoacoustics, it is not likely to be measured with mathematical rigour (Pedroso de Lima, 2012, p. 290) and its judgement varies, or may vary, from person to person, or even for the same individual according to the circumstances (Henrique, 2007, p. 169). Thus, subjective attributes are dependent on the listener as well as on one or more physical parameters (Rossing, 1990, p. 80).

2.1.3.1 Listener Independent Sound Attributes

a) As an objective attribute, sound duration or time t is a quantity which can be measured with atomic clocks, chronometers or any other time measuring instruments. As a base quantity, it is normally expressed in the base unit *second*, represented by the symbol s, in the International System of Units (SI). A very precise definition of second is that it is "the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom" at rest, at a temperature of 0 K, with corrections from ambient radiation (BIPM, 2006, p. 113). Other non-SI units of time used to indicate the sound duration are the minute min, the hour h and the day d.

b) *Period* refers to a particular length of time and when cycles are observed, that is, when something repeats itself at regular time intervals, being periodic or cyclic, then the quantity period T measures specifically the minimum time required to complete one whole cycle or vibration of it (Rossing, 1990, p. 31). Thus, period is a particular time quantity for one cycle, as can be observed in figure 2.1. The starting point of a cycle is described by its initial phase or phase angle ϕ in degrees or radians. A phase difference between two sounds means that one sound is temporally delayed relative to the other, that is, they are out-of-phase, but if they start at the same point they are said to be in phase (Roads et al., 1996, pp. 18–19).

c) On the other hand, if one desires to measure how often something happens during a certain time interval, then the quantity frequency f is used. It is defined generically as the number of cycles per time unit (Henrique, 2007, p. 47). As the time unit second is utilized in most oscillatory phenomena, fundamental frequency or lowest frequency of a sound (Henrique, 2007, p. 180) is then defined as a component with the lowest number of cycles per second in that sound, being expressed in the derived unit *hertz*, abbreviated Hz, which means reciprocal second or s^{-1} (BIPM, 2006, p. 118) and that is used in honour of the German physicist Heinrich Rudolf Hertz (February 22, 1857 - January 1, 1894)[3], who determined it. A single frequency is then

[3] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Heinrich_Hertz`

represented graphically by a sinusoid or sine wave (see figure 2.1), named after the mathematical function sine, which describes its smooth repetitive oscillation.

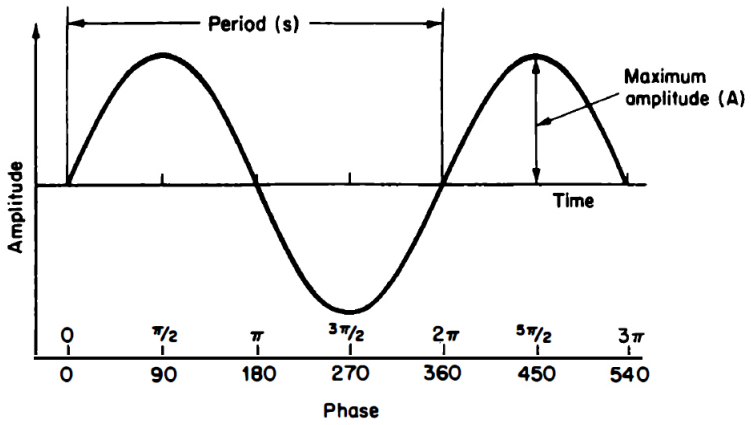


Figure 2.1: Graphical representation of a single vibration with period T: a sinusoid or sine wave (Moore, 2013, p. 3).

Thus, frequency is a rate quantity, which is the reciprocal of the period (Rossing, 1990, p. 20):

$$f = \frac{1}{T} \Leftrightarrow T = \frac{1}{f} \, .$$

A sound can be comprised of many frequencies in addition to the fundamental frequency, therefore called *complex sound* by Howard & Angus (2001, pp. 51, 89), because its waveform is more complex than a simple sinusoid. Each one of these individual frequency components is called a partial. If a partial has a frequency which is virtually an integer multiple of the fundamental frequency f, that is, a frequency equal to $n \times f$, where n is an integer, then it is referred to as the nth-harmonic (Roads et al., 1996, pp. 16–17). Other partials with frequencies that are not integer multiples of the fundamental frequency are designated as non-harmonics (Henrique, 2007, p. 178). The contents of a sound can be displayed by a frequency-domain or spectrum representation of it at a certain short instant in time (see figure 2.2), which permits the observation of the relations between the partials of a sound, or by a long-time average spectrum (LTAS), the latter allowing to detect frequency regions with higher energy or magnitude by the averaging process, as is the case in the lower frequency region in figure 2.3, whereas background noise averages towards zero (Henrique, 2007, p. 287).

In this regard, Pedroso de Lima (2012, p. 81) refers that a *musical sound* is generally speaking the result of the superposition of periodic or approximately periodic sounds, that is, harmonics. In addition, he mentions that *noise* corresponds to non-harmonic related sounds or to sounds that are very brief or which unpredictably change their characteristics, such as white noise and pink noise (see section 2.1.4).

The main difference between the spectrum of periodic and non-periodic sounds is that discrete partials are observed in periodic sounds, represented as vertical lines in the spectrum with a spacing which is inversely proportional to the period of the fundamental frequency (Howard & Angus, 2001, p. 55). In non-periodic sounds, such as noise (see section 2.1.4), the spectrum is continuous and does not have discrete components. Figure 2.2 shows that, in practice, both coexist more or less. The higher peaks in the spectrum of figure 2.2 indicate that there is a harmonic relationship between partials, starting at approximately 400 Hz. So, in this case, the first higher peak corresponds to the first harmonic of the sound with a frequency of about 400 Hz,

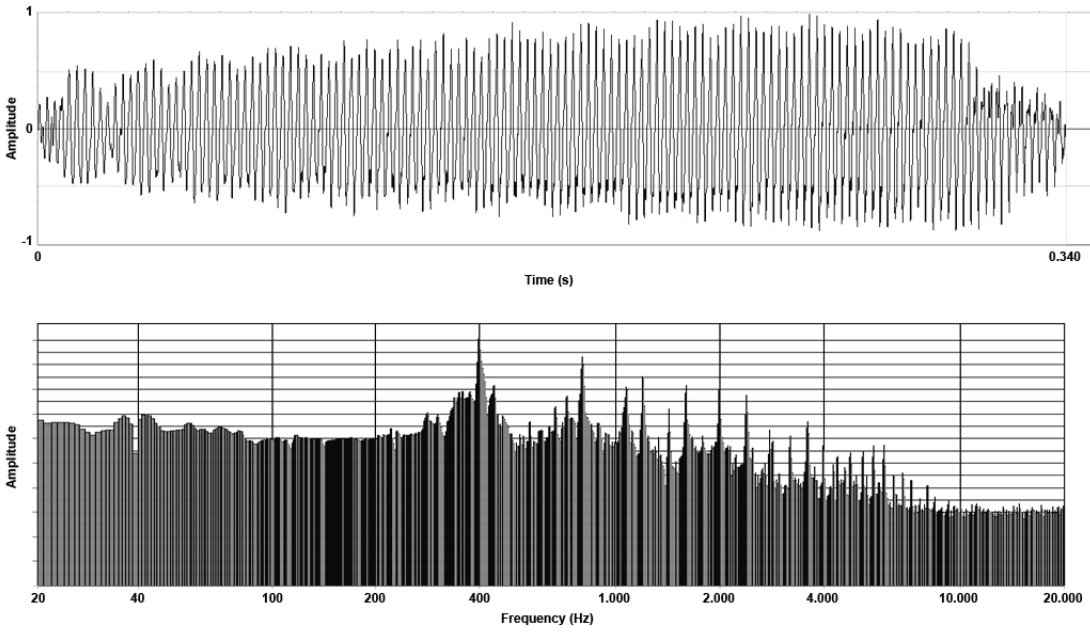


Figure 2.2: Time-domain representation of a sound with a short duration of 0.340 s (above), produced by a musical instrument, horn, and the corresponding spectrum representation (below).

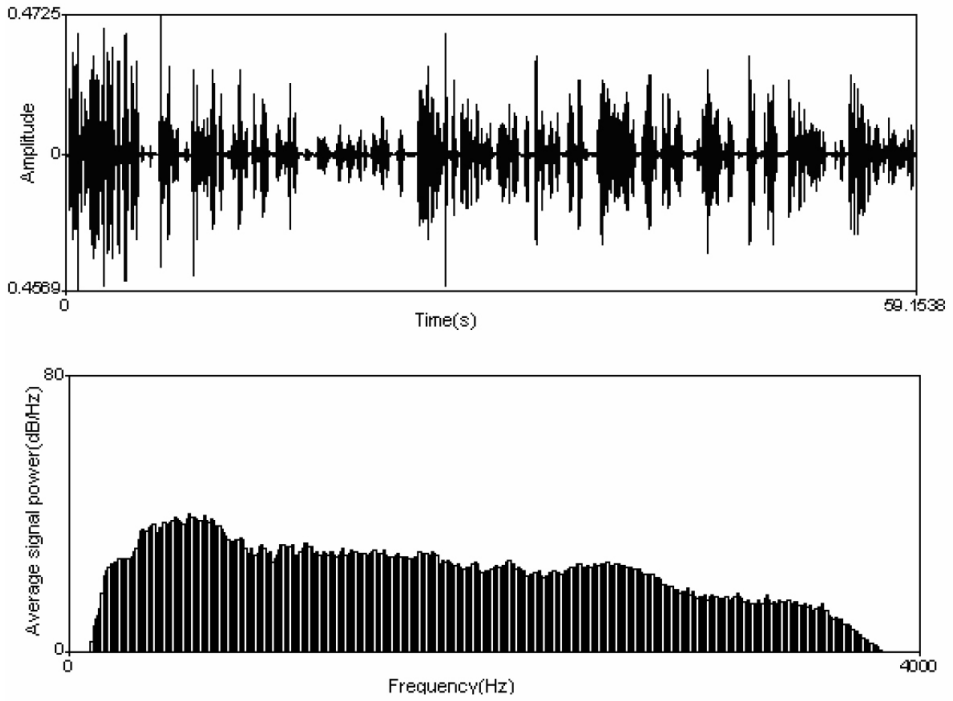


Figure 2.3: A speech signal with a duration of about 59 s (above) and the corresponding long-time average spectrum (below) (Pauk, 2006, p. 31).

the second higher peak to the second harmonic of about 800 Hz, the third to the third harmonic of about 1200 Hz, and so on. Nevertheless, there are also other frequencies that are non-harmonics and which present high peaks, too, such as 1070 Hz and 1425 Hz (these are approximate values). The remaining continuous spectrum corresponds to noise, such as ambient noise, noise produced by the instrument and instrumentalist,

and electrical noise captured in the recording.

The spectrum of frequencies in a sound varying over time can also be displayed by a spectrogram, where the horizontal axis describes time, the vertical axis represents frequency, and the colour indicates amplitude or level (explained later in this section) of a particular frequency at a certain time (Rossing, 1990, p. 331) (see figure 2.4).

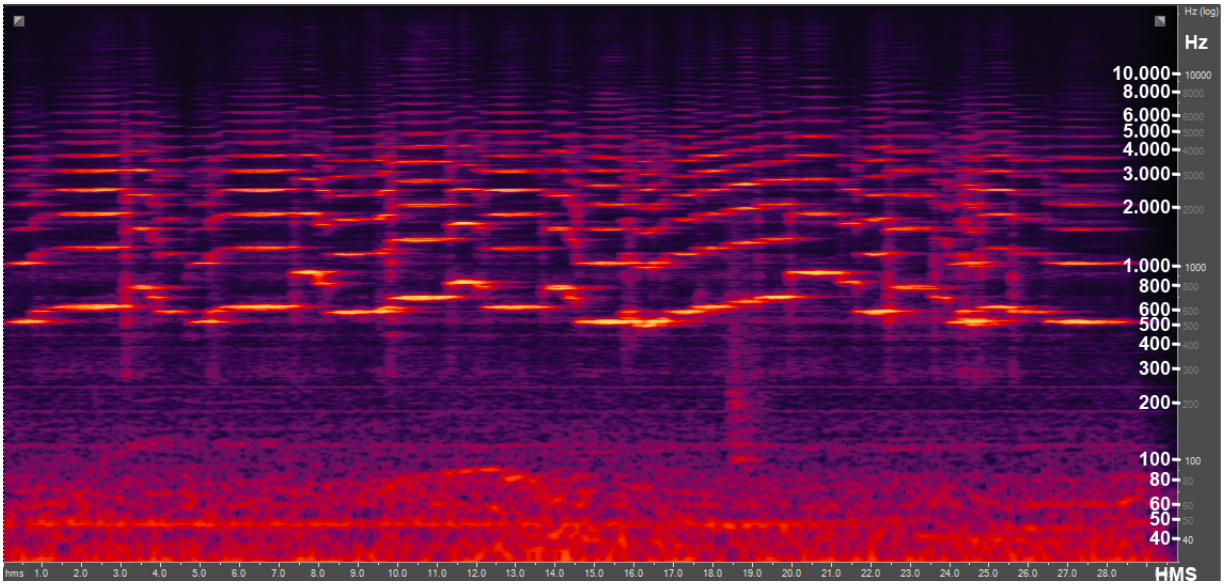


Figure 2.4: A spectrogram of sounds played on a violin. *HMS* in the horizontal axis stands for *hours*, *minutes*, and *seconds*.

The representations of a sound in the time domain and in the frequency domain, where time and frequency are the parameters on the horizontal axis of the corresponding graphs, both provide complementary information about it. Fourier analysis, named after the French mathematician and physicist Jean-Baptiste Joseph Fourier (March 21, 1768 - May 16, 1830)[4], is a family of mathematical techniques, all based on decomposing signals into sinusoids (Smith, 1997, p. 141). Thus, the transformation of any arbitrary periodic or non-periodic signal from the time domain to the frequency domain is achieved by the Fourier Transform, which is a mathematical operation that allows a function dependent on time to be transformed into a function dependent on frequency (Henrique, 2007, p. 260). The reverse is achieved by the Inverse Fourier Transform. In digital signal processing (DSP) systems, the best known Fourier Transforms are the Discrete Fourier Transform (DFT) and the Fast Fourier Transform, which is a class of fast algorithms for the computation of DFT (Pauk, 2006, p. 18). A more detailed explanation about this issue can be found in Smith (1997); this aspect goes beyond the scope of this thesis.

d) While a whole cycle of sound is completed in a time of one period, a distance equal to one complete wavelength λ is travelled by it. Wavelength is the minimum space required to complete a whole cycle or vibration of sound, and it is usually expressed in the base unit *metre*, represented by the symbol m, in the SI. The metre is defined as "the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second" (BIPM, 2006, p. 112).

[4] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Joseph_Fourier`

e) Therefore, as well as the derived scalar quantity *speed* (magnitude of the vector quantity *velocity* – see section 2.3.4.8 for the definition of *vector*) refers to the base quantity distance x travelled per time unit t, speed of sound in the air c_{air} is also equal to the wavelength per period or to the wavelength times frequency:

$$c_{\mathrm{air}} = \frac{x}{t} \Leftrightarrow c_{\mathrm{air}} = \frac{\lambda}{T} \Leftrightarrow c_{\mathrm{air}} = \lambda \times f \ .$$

In practical terms, due to its dependence on temperature and, to a lesser extent, on relative humidity, that is, water vapour content in percentage, in dry air (0% relative humidity) sound travels with a certain speed c_{air}, which is determined approximately by the following formula:

$$c_{\mathrm{air}} \approx 331.3 \times \sqrt{1 + \frac{^{\circ}\mathrm{C}}{273.15}} \Leftrightarrow c_{\mathrm{air}} \approx 20.05 \times \sqrt{273.15 + ^{\circ}\mathrm{C}} \Leftrightarrow c_{\mathrm{air}} \approx 331.3 + 0.606 \times ^{\circ}\mathrm{C} \ ,$$

where $^{\circ}\mathrm{C}$ is the temperature in degrees on the Celsius scale and c_{air} is a derived quantity expressed in the derived unit $\mathrm{m/s}$ or $\mathrm{m\ s^{-1}}$ (Wikipedia, 2015d). Therefore, speed of sound increases by about 0.6 $\mathrm{m\ s^{-1}}$ for each $^{\circ}\mathrm{C}$ rise in ambient temperature (Rossing, 1990, p. 41; Howard & Angus, 2001, p. 7).

f) The quantity pressure p exerted by air in the atmosphere of Earth is defined as a force F that air exerts per unit area A (Kane & Sternheim, 1988, p. 245):

$$p = \frac{F}{A} \ .$$

In this case, the quantity force is an external influence on air that causes it to change its velocity relative to an inertial reference frame (Tipler, 1999, p. 85). As force is expressed in the SI derived unit *newton*, symbol N, after the English physicist and mathematician Sir Isaac Newton (January 4, 1643 - March 31, 1727)[5], and an area is given in $\mathrm{m^2}$, pressure is expressed in the derived unit $\mathrm{N/m^2}$ or Pa (*pascal*), in honour of the French mathematician, physicist, inventor and theological writer Blaise Pascal (June 19, 1623 - August 19, 1662)[6]. The standard atmospheric pressure is equal to 101 325 Pa (BIPM, 2006, p. 127).

Similarly, sound pressure p is also defined as the force that sound exerts per unit area. A disturbance in the air such as an oscillation in atmospheric pressure, as already referred to in section 2.1.1, leads to alternating compression and rarefaction regions in the air. In the former case, the number of air molecules is greater than the normal air distribution, leading to a locally and temporarily slight increase of the atmospheric pressure (Henrique, 2007, pp. 202–203). In the latter case, the number of air molecules is smaller than the normal air distribution, resulting in a locally and temporarily slight decrease of the atmospheric pressure. In addition, the air particles that are liable to vibration only move by an infinitesimal amount, that is, a very short distance, on each side of its equilibrium or rest position (Arau, 1999, p. 9). Thus, sound causes small pressure fluctuations in the air, superimposed on the normal atmospheric pressure (BIPM, 2006, p. 107). The distance of the air particles from their rest position, the sound pressure deviation from the atmospheric pressure (Sundberg, 1991, p. 10), or amount of air pressure change (Roads et al., 1996, p. 15) over a certain interval of time, is called amplitude. A positive amplitude then corresponds to a compression, a negative amplitude matches a rarefaction and the atmospheric pressure is therefore considered as equal to zero amplitude. Figure 2.1 shows

[5] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Isaac_Newton`
[6] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Blaise_Pascal`

the maximum positive amplitude A of a sinusoid, which has an instantaneous amplitude given by $A \times \sin(2\pi ft)$ or $A \times \sin(360ft)$, where A is the maximum amplitude, f is the frequency, and t corresponds to time (Moore, 2013, p. 3). Furthermore, amplitude of a sound decreases with distance from the source mainly due to losses in heat dissipation (Pedroso de Lima, 2012, p. 80).

On the other hand, in a time-domain representation or graph of amplitude versus time, the way amplitude of sound changes over its duration is described by a curve that follows the amplitude, so to say delimiting it, which is called amplitude envelope (Roads et al., 1996, p. 95) or decay curve (Rossing, 1990, p. 22) (see figure 2.5).

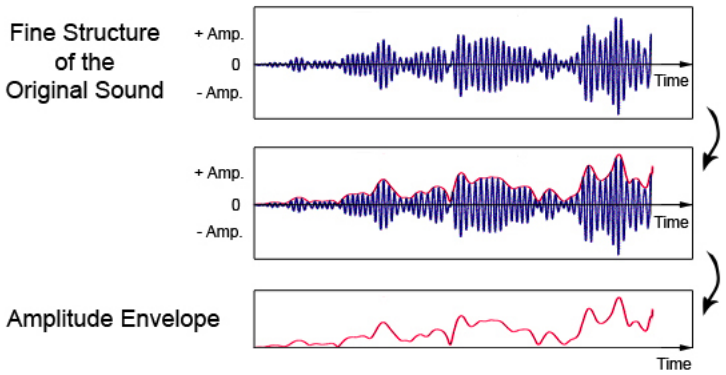


Figure 2.5: Amplitude envelope of sound as a result of positive amplitudes (+ Amp.).

The amplitude envelope of a single sound, consisting of one or more frequencies, usually includes the initial transient or attack portion at the beginning, the release or final decay at the end, and variations in amplitude in between (Rossing, 1990, p. 80) (see figure 2.6).

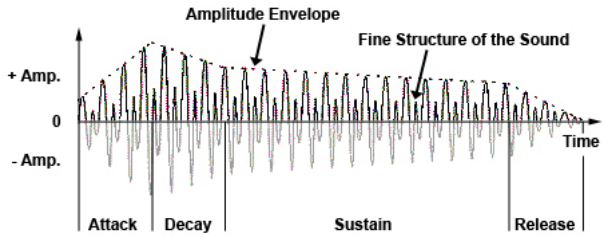


Figure 2.6: Amplitude envelope of a single complex sound (adapted from `http://www.zytrax.com/tech/audio/sound.html`).

The attack portion lasts for a certain time that is measured from the generation of the sound until it reaches the maximum amplitude, due to initial irregular vibrations (Henrique, 2007, p. 171). The variations in amplitude after the attack comprise the initial decay and the sustain portions, which can be identified depending on the overall sound duration. The initial decay portion comprehends the time it takes for the maximum amplitude of the attack to drop to the sustained amplitude, which is in turn the amplitude at which the sound more or less stabilizes or more regular vibrations can be observed. The release portion at the end also lasts for a certain time, but it is measured from the amplitude of the sustain portion until it reaches silence. The duration of each of these envelope phases depends on the sound source itself and on how the sound is originated. Furthermore,

a sound produced by a musical instrument has an attack time that ranges from about 5 ms for some percussive instruments to up to 200 ms for woodwind instruments (Lerch, 2012, p. 120), a sustain portion of about 100 ms to some seconds, and a release time that varies a lot from instrument to instrument and which also depends on the surrounding environment (Henrique, 2007, p. 173). Nevertheless, if the amplitude envelope of a single complex sound decays slowly, then its frequency spectrum occupies a narrow bandwidth, having less frequencies (Howard & Angus 2001, pp. 60–61; Moore 2013, pp. 8–9). On the other hand, if the amplitude envelope decays quickly, then the spectrum occupies a larger bandwidth, presenting more frequencies.

It is also possible to display sound in a three-dimensional graph, known as a waterfall plot, where each one of the three mutually perpendicular axes represent frequency, time, and amplitude or level, respectively (see figure 2.7).

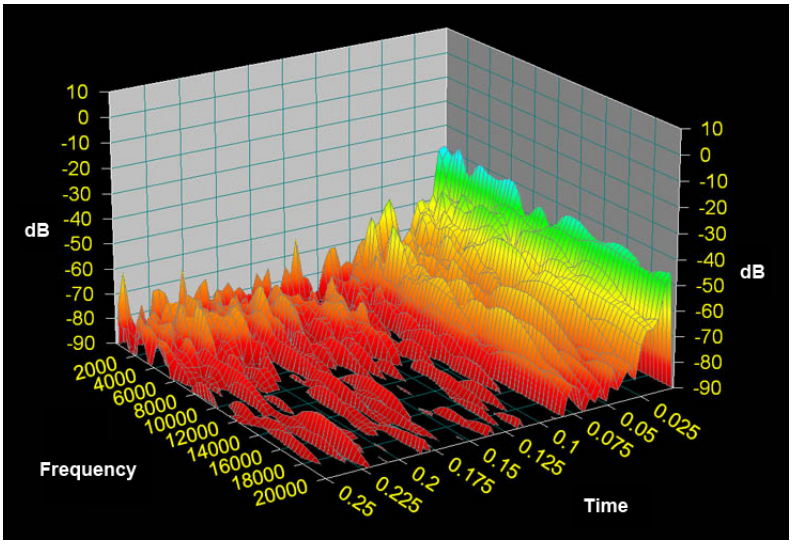


Figure 2.7: Waterfall plot (adapted from `http://www.silencertalk.com/tests/5-06-05-5.56mm/Spectrum/Wavelet/SRT-Hurricane-3D_1200.jpg`).

Due to the sensitivity of human ears to pressure and because it is easier to measure, pressure is used as a measure of the amplitude of sound (Howard & Angus, 2001, p. 17) and it is determined at a point in the air, at a certain distance from the sound source, and at a given instant. It should be noted here that the result of this measurement is the sum of all sound pressure waves at the measurement point. However, sound pressures that are able to stimulate the human auditory system vary over a wide range in the order of magnitude of more than a million, which is not very practical in everyday life. For a young and perfectly healthy human ear, sound pressure varies from the threshold of hearing, of about 20 micro-pascal, that is, 20 $\mu\mathrm{Pa}$ or 20 $\times$ 10^{-6} Pa, to the threshold of pain (see section 2.1.3.2), of approximately 20 Pa, at a frequency of 1 kHz (Pedroso de Lima, 2012, p. 164). Consequently, in order to express numbers with fewer digits, the quantity sound pressure level L_p is used instead (see figure 2.8).

Sound pressure level is expressed on a decadic logarithmic scale in decibel, symbol dB, based on the ratio of the actual sound pressure p_{actual} to a pressure reference p_{ref}, the latter being equal to the hypothetical threshold of hearing at 1 kHz of 20 $\times$ 10^{-6} Pa (Howard & Angus, 2001, p. 17):

$$L_p = 10\log_{10}\frac{p_{actual}^2}{p_{ref}^2} \Leftrightarrow L_p = 10\log_{10}\left(\frac{p_{actual}}{p_{ref}}\right)^2 \Leftrightarrow L_p = 20\log_{10}\frac{p_{actual}}{p_{ref}} \Leftrightarrow L_p = 20\log_{10}\frac{p_{actual}}{20\times10^{-6}\ \mathrm{Pa}}\ .$$

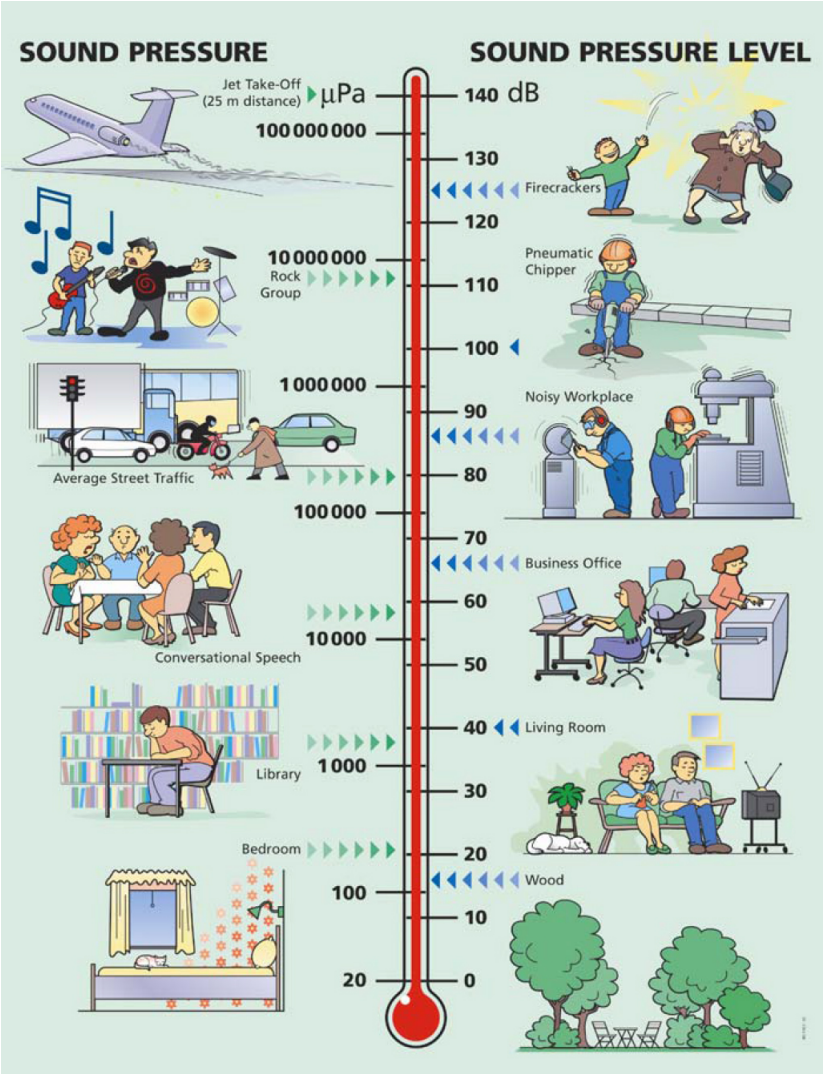


Figure 2.8: Comparison between sound pressure and sound pressure level (Brüel & Kjaer, 2015).

This is the reason why sound pressure levels are indicated as dB *re* 20×10^{-6} Pa, where *re* means reference. Thus, in this context, 0 dB corresponds approximately to the threshold of hearing and 120 dB equals roughly the threshold of pain. Furthermore, the value of the sound pressure level without the reference of the corresponding distance to the sound source is meaningless (Pedroso de Lima, 2012, p. 164). The use of logarithmic scales is related with the fact that the response of the human ear to acoustical disturbances is of the logarithmic type (Henrique, 2007, p. 243).

By definition, the decibel always describes a ratio of two quantities (G. Davis & Jones, 1989, p. 19) and it corresponds to one tenth of a bel, symbol B, a unit named in honour of the Scottish scientist and inventor Alexander Graham Bell (March 3, 1847 - August 2, 1922)[7], who patented the first functional telephone. It does not have an absolute scale, but rather a comparative scale, and is therefore said to be dimensionless (Pedroso de Lima, 2012, p. 161). Furthermore, it allows to manage quantities with very large ranges, and when the magnitude of a sound is specified in decibels, the use of the word *level* refers to its magnitude (Moore, 2013, p. 10). The units bel and decibel have been accepted by the Comité International des Poids

[7] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Alexander_Graham_Bell`

et Mesures (CIPM) for use with the International System, but are not considered as SI units (BIPM, 2006, p. 127).

g) Another scalar objective attribute is sound power W, which measures the total energy E radiated in all directions by a sound source per unit of time (Howard & Angus, 2001, p. 16):

$$W = \frac{E}{t} \ .$$

As energy is expressed in the SI derived unit *joule*, symbol J, sound power is represented by the SI derived unit $\mathrm{J/s}$ or W (*watt*), named after the Scottish engineer James Watt (January 19, 1736 - August 25, 1819)[8]. Sound power level L_W is used instead of sound power in order to express numbers with fewer digits on a logarithmic scale in decibel, being useful for comparing the total acoustic power radiated by sound sources. It is calculated from the ratio of the actual sound power W_{actual} to a power reference W_{ref} of 1 picowatt, that is, 1 pW or 10^{-12} W (Howard & Angus, 2001, p. 16):

$$L_W = 10 \log_{10} \frac{W_{actual}}{W_{ref}} \Leftrightarrow L_W = 10 \log_{10} \frac{W_{actual}}{10^{-12}\ \mathrm{W}} \ .$$

h) Sound is a phenomenon that occurs in three dimensions and it gets weaker as it spreads out in all directions from a sound source. Thus, it is useful to determine the energy transfer rate that occurs in a given area A in space, that is, sound power per unit area or power density of a sound propagating in a particular direction. This operation is called sound intensity I and it has a direction which is perpendicular to the area that the energy is flowing through (Howard & Angus, 2001, pp. 14–15):

$$I = \frac{\frac{\text{Quantity of energy transferred}}{\text{unit time}}}{\text{unit area}} \Leftrightarrow I = \frac{W}{A} \ .$$

In addition, sound intensity as a function of distance from a source, which emanates sound omni-directionally as increasing sound spheres, that is, a point source (see section 2.2.1), is given by:

$$I = \frac{W}{4\pi r^2} \ ,$$

where r is the distance from the source and $4\pi r^2$ is the growing spherical sound surface over which power is distributed. Thus, the intensity decreases by a factor of four as the distance r is doubled, but the total sound power remains the same (see figure 2.9). This inverse square relationship between intensity and distance from the source is called the *Inverse Square Law* and it is purely a function of geometry (Howard & Angus, 2001, pp. 28–29). Moreover, it is only valid in free field conditions, that is, when the sound source is far away from any surfaces so that there are no reflected sounds. In practice, sound intensity does not usually decrease so sharply due to multiple reflections on obstacles, on the floor and walls, but it decreases additionally due to water vapour and impurities in the air, being more pronounced at higher frequencies, as already mentioned in section 2.1.1.

Sound intensity level L_I is used instead of sound intensity in order to express numbers with fewer digits on a logarithmic scale in decibel. It is calculated from the ratio of the actual sound intensity I_{actual} to an intensity

[8] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/James_Watt`

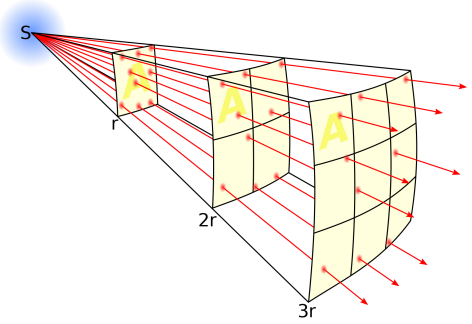


Figure 2.9: Inverse square relationship between intensity and distance from the source (Wikipedia, 2015b).

reference I_{ref} of 1 picowatt per square metre or 10^{-12} $\mathrm{W\ m^{-2}}$ (Howard & Angus, 2001, p. 15), which is equivalent to the notional threshold of hearing at 1 kHz of 20 $\times$ 10^{-6} Pa (Moore, 2013, p. 10):

$$L_I = 10\log_{10}\frac{I_{actual}}{I_{ref}} \Leftrightarrow L_I = 10\log_{10}\frac{I_{actual}}{10^{-12}\ \mathrm{Wm^{-2}}}\ .$$

Relatively to the above referred omni-directional sound source in a free field, the intensity decreases by about 6.02 dB as the distance r is doubled:

$$L_I = 10\log_{10}\left(\frac{\frac{W}{4\pi(2r)^2}}{\frac{W}{4\pi r^2}}\right) \Leftrightarrow L_I = 10\log_{10}\left(\frac{W}{4\pi(2r)^2}\times\frac{4\pi r^2}{W}\right) \Leftrightarrow L_I = 10\log_{10}\frac{r^2}{(2r)^2} \Leftrightarrow L_I = 10\log_{10}\frac{r^2}{4r^2}$$

$$\Leftrightarrow L_I = 10\log_{10}\frac{1}{4} \Leftrightarrow L_I = 10\log_{10}1 - 10\log_{10}4 \Leftrightarrow L_I = -10\log_{10}4 \Leftrightarrow L_I \approx -6.02\ \mathrm{dB}\ .$$

In a free field, if there is only a single pressure wave, that is, a direct sound, from the sound source at the measurement point, the sound intensity level and the sound pressure level are approximately equivalent (Howard & Angus, 2001, p. 18–19; Pedroso de Lima, 2012, p. 163):

$$L_I \approx L_p \Leftrightarrow 10\log_{10}\frac{I}{I_{ref}} \approx 10\log_{10}\left(\frac{p}{p_{ref}}\right)^2 \Leftrightarrow 10\log_{10}\frac{I}{I_{ref}} \approx 20\log_{10}\frac{p}{p_{ref}}\ .$$

In general, typical sound sources do not propagate sound omni-directionally but favour rather a certain direction. The directivity information of a sound source is so described by the unit-less directivity factor value Q at a point or by the directivity index D_I on a logarithmic scale, in decibel (D. Davis & Davis, 1997, pp. 106 and 108–109). They are both dependent on the radiation angle and frequency (D. Davis & Davis, 1997, p. 111), typically increasing with frequency and with smaller radiation angles. Howard & Angus (2001, p. 18) demonstrate that sound intensity is proportional to the square of pressure and thus the directivity factor can be written as (D. Davis & Davis, 1997, pp. 108–109; Arau, 1999, pp. 31–32; Howard & Angus, 2001, p. 31):

$$I_\theta = \frac{W}{4\pi r^2}\times Q \Leftrightarrow Q = \frac{I_\theta}{\frac{W}{4\pi r^2}} \Leftrightarrow Q = \frac{I_\theta}{\bar{I}} \Leftrightarrow Q = \left(\frac{p_\theta}{\bar{p}}\right)^2 \Leftrightarrow Q = 10^{\left(\frac{L_{p_i}-\bar{L_p}}{10}\right)} \Leftrightarrow Q = 10^{\left(\frac{D_I}{10}\right)}\ ,$$

where I_θ is the measured sound intensity, p_θ is the measured sound pressure, and L_{p_i} is the sound pressure level measured at point number *i* in space, all three at distance r and angle θ from the source, and $\bar{I}$, $\bar{p}$, and $\bar{L_p}$

are, respectively, the average sound intensity, the average sound pressure, and the average sound pressure level, all three over a spherical surface at the distance r, with

$$\bar{L_p} = 10 \log_{10} \frac{\sum_{i=1}^{n} 10^{\frac{L_{p_i}}{10}}}{n} ,$$

where n is the integer total number of measured points in space (cf. section 2.2.1).

The directivity index is then given by:

$$D_I = 10 \log_{10} Q \Leftrightarrow D_I = L_{p_i} - \bar{L_p} .$$

If the sound source propagates sound omni-directionally, then $Q = 1$ and $D_I = 0$ dB. For hemispherical radiation, $Q = 2$ and $D_I = 3$ dB. In the latter case, the sound pressure level measured on the hemisphere is 3 dB higher than for spherical radiation (Ostergaard, 2003, p. 25; Almeida, 2013, p. 51).

2.1.3.2 Listener Dependent Sound Attributes

The relationships between physical stimuli and the subjective sensations they produce on human beings are studied quantitatively by Psychophysics (Rossing 1990, p. 77; Wikipedia 2015c). *Psychophysics* is a term proposed by the German philosopher, physicist, and experimental psychologist Gustav Theodor Fechner (April 19, 1801 – November 18, 1887)[9] in his 1860's published book entitled *Elemente der Psychophysik* (*Elements of Psychophysics*), where he describes research intended to determine the quantitative relationships between stimuli and perceived sensations, based on earlier work by the German physician Ernst Heinrich Weber (June 24, 1795 – January 26, 1878)[10] (Rossing, 1990, p. 77).

On the other hand, *Psychoacoustics* is a branch of Psychophysics, which studies specifically how humans perceive sound (Howard & Angus, 2001, p. 65), that is, it studies the relationships between objective physical properties of acoustic sensory stimuli and subjective physiologically evoked responses of the human ear system (Pedroso de Lima, 2012, p. 286). One of the main areas of interest of Psychoacoustics is the identification and localization of sound sources, which will be discussed in section 2.5. The analysis and interpretation of sounds and the codification of frequencies and pressures are other issues dealt with by Psychoacoustics.

Furthermore, Pedersen & Zacharov (2008, p. 1217) refer that human beings have the need to describe sounds in a detailed manner by using a large number of words in different sound domains. Nevertheless, many common terms and concepts are encountered in many studies of the same domain, although there is frequently a different and more refined set of sensory descriptors needed to characterize the perceptually prominent characteristics of the stimuli under test (Pedersen & Zacharov, 2008, p. 1219). In any case, according to Rossing (1990, pp. 63 and 80), loudness, pitch, timbre, and duration are four subjective attributes often used to describe sound, especially musical sound, each one depending on one or more physical parameters, as already mentioned in section 2.1.3. In order to differentiate two sounds, Pedroso de Lima (2012, p. 159) states that the four fundamental subjective attributes are virtually the same Rossing proposes, except for the latter, which is the localization in space instead of duration. In other domains where sound quality testing is an important design concept and where the audible suitability of a product when compared with a user's

[9] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Gustav_Fechner`
[10] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Ernst_Heinrich_Weber`

expectation is searched for (University of Salford, 2015), which is not the case in this thesis, the most used attributes are the classical psychoacoustic attributes, such as loudness, sharpness, roughness, fluctuation strength, tone prominence, and impulse prominence, supported by instrumental procedures giving estimates of the attributes (Pedersen & Zacharov, 2008, p. 1218).

The relationships between stimuli and the evoked responses are described with relative precision by the Weber-Fechner's Law proposed by Fechner, which is an extension of Weber's Law. Whereas Weber's Law states that the Just Noticeable Difference (JND), Difference Limen (DL), or Differential Threshold (Moore, 2013, p. 425), that is, the smallest detectable change in a quantity, between two stimuli is proportional to the magnitude of the stimuli, denoting the ability of the human ear to perceive the relative stimulus, Weber-Fechner's Law declares that the subjective sensation is proportional to the logarithm of the stimulus' magnitude or that the sensation increases logarithmically, as the stimulus is increased linearly, relating the global sensation with the stimulus (Pedroso de Lima, 2012, p. 289–290).

The determination of the JND of a quantity is carried out based on a number of attempts, that is, it is determined statistically, usually as the stimulus difference that produces 75% correct responses, since the difference perceived by the human being varies subjectively from trial to trial (Moore, 2013, pp. 143–144). In that sense, Pedersen & Zacharov (2008, p. 1217) refer that if it is possible to hear a difference between two sounds, then the perceived magnitude of one or more fundamental attributes are different for the two sounds.

Thus, for wideband or for bandpass-filtered noise, the JND relatively to sound intensity, ΔI, is roughly proportional to the magnitude of the original sound intensity I:

$$\Delta I = C \times I \Leftrightarrow \frac{\Delta I}{I} = C \ ,$$

where the parameter *C* is an estimated constant, which is equivalent to the fixed proportion or percentage increase (if multiplied by 100) the original sound intensity is subject to by the JND someone is able to reliably detect (Henrique, 2007, p. 860). The value of $\frac{\Delta I}{I}$ is called the Weber fraction (Moore, 2013, p. 144) and it can also be expressed in decibels as a change in level, ΔL, as:

$$\Delta L = 10 \log_{10} \frac{\Delta I}{I} \ .$$

If sound intensity is doubled, then the JND is also doubled, that is, it increases by the same percentage. Nevertheless, the JND relatively to sound intensity depends on frequency (Pedroso de Lima, 2012, pp. 288–289). So, at low and midrange frequencies, the JND levels are of about 1 dB for soft sounds with sound pressure levels around 30 dB to 40 dB. Sinusoidal sounds or broadband noise with spoken speech sound pressure levels of about 60 dB at the ear inputs present JND levels between 0.3 dB and 1.0 dB. For sounds with higher frequencies and higher sound pressure levels, the JND levels are under 0.5 dB, as shown in figure 2.10. As a general rule of thumb the JND in sound level is about 1 dB (Howard & Angus 2001, p. 88; Pedroso de Lima 2012, p. 164).

For sine waves and narrow-band noise, it has been demonstrated, first by Robert R. Riesz in 1928 and later by many others, that Weber's Law does not hold, that is, that sound intensity discrimination improves at higher intensities (Moore, 2013, pp. 144, 150–151). This decrease in the Weber fraction is known as the *near miss* of the Weber's Law.

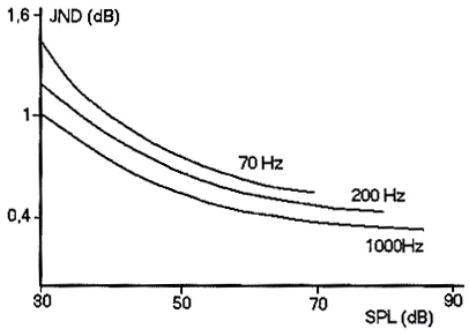


Figure 2.10: Just Noticeable Difference in sound pressure level for three different frequencies (Pedroso de Lima, 2012, p. 290).

The relationship between sensation and stimulus, which is introduced by Weber-Fechner's Law, can be described in terms of a differential equation (Pedroso de Lima, 2012, p. 289):

$$ds = k \times \frac{dI}{I} \ ,$$

where ds is the differential change of the sensation, dI is the differential increase in sound intensity I and k is an estimated constant. The global sensation is then obtained by integrating the above equation:

$$\int_{s_0}^{s} ds = k \times \int_{I_0}^{I} \frac{1}{I} \, dI \Leftrightarrow s - s_0 = k \times \ln \frac{I}{I_0} \ ,$$

where s is the sensation after the change in intensity, s_0 is the minimum sensation before the change in intensity, k is an estimated constant, I is the sound intensity after its change, I_0 is the original intensity before its change, and $\ln$ is the natural logarithm. Thus, whereas sound intensity is increased linearly, the sensation increases logarithmically (Pedroso de Lima, 2012, p. 290). In this case, the sensation related with intensity is called loudness.

a) Loudness is defined as that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud (Moore, 2013, p. 133), as a perceptual attribute that decodes how loud an auditory object is perceived (Pulkki, 2001b, p. 6, quoting Moore), or as a subjective quantity which allows us to rate a sound as strong or weak (Pedroso de Lima, 2012, pp. 159). As a subjective quantity, loudness cannot be measured directly, because the human ear presents sensitivity differences and because this sensitivity of hearing can vary considerably from person to person (Howard & Angus 2001, p. 82; Pedroso de Lima 2012, p. 293), that is, a sound with a higher sound pressure level than another may sound lower than the other. Thus, at 1 kHz, considering that sound pressure level and sound intensity level are approximately equivalent in free field, as already mentioned in section 2.1.3.1, the average human ear with normal hearing can sense sound intensities from about 10^{-12} $\mathrm{W\ m^{-2}}$ ($L_I = 0$ dB) to approximately 1 $\mathrm{W\ m^{-2}}$ ($L_I = 120$ dB), or, as a pressure sensitive organ, the ear can detect sound pressures from about 20 $\times$ 10^{-6} Pa ($L_p = 0$ dB) to more or less 20 Pa ($L_p = 120$ dB), which are respectively considered as the minimum and maximum limits, known as thresholds of perceptibility (Blauert, 1997, p. 16), between which the ear can perceive sounds without any

risk of injury, discomfort or pain for that frequency. Nevertheless, the ear's sensitivity varies with frequency and these limits are not constant over the audible frequency range, being less pronounced in the upper limit when compared with the variations of the lower limit, fundamentally due to the anatomy and physiology of the ear (Pedroso de Lima, 2012, p. 292). As a consequence, the variations for sinusoidal stimuli between 16 Hz and 20 kHz of the lower limit are defined, as an average, by a curve called absolute threshold of hearing, which represents the minimum sound pressure level of a single sinusoidal sound with a duration of more than 200 ms that is detected in free field by 50% of 18 to 25 year old young people with normal hearing (Pedroso de Lima, 2012, p. 293) or where the probability of an auditory event to be present or not present is equal to 50% (Blauert, 1997, p. 16) (see figure 2.11).

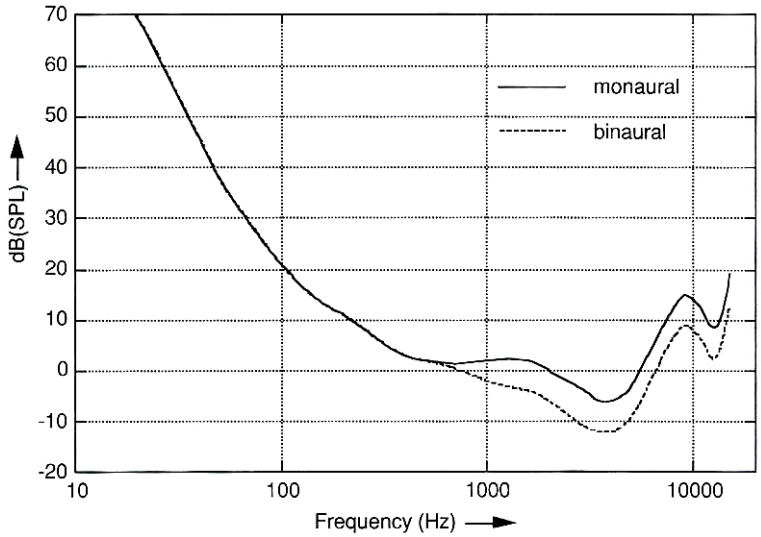


Figure 2.11: Absolute threshold of hearing as a function of frequency (Howard & Angus, 2001, p. 87).

This curve shows that the sensibility of the human ear for monaural listening to a sinusoidal sound presented frontally to the listener is greatest around 3.4 kHz, of about 10^{-5} Pa ($L_p \approx -6$ dB), due to a main resonance of the auditory canal in the outer ear at that frequency (see section 2.4), and that for binaural listening the sensibility increases further between 3 dB and 6 dB (Howard & Angus, 2001, pp. 80, 83, 86–87). So, higher sound pressure levels are more bearable for lower and higher frequencies than for the midrange. The curve for the upper limit is usually called threshold of pain.

The relationship between the measured sound pressure level and the perceived loudness of any audible sinusoidal sound is shown by the equal-loudness contours. Nevertheless, their shapes vary and depend on the method used to measure them (Moore, 2013, p. 134), an issue which is beyond the scope of this thesis. These contours were originally determined experimentally in 1933 by Harvey Fletcher (September 11, 1884 - July 23, 1981)[11] and Wilden Andrew Munson (April 6, 1902 - June 15, 1982)[12] (Howard & Angus, 2001, p. 83), therefore known as the Fletcher-Munson curves, where a 1 kHz sinusoidal sound was fixed in level and listeners had to adjust the level of different frequency sinusoidal test sounds in order to obtain a loudness match. The latest revised ISO standard equal-loudness contours for sounds presented binaurally from the

[11] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Harvey_Fletcher`
[12] Retrieved 02/01/2016, from `http://www.findagrave.com/cgi-bin/fg.cgi?page=gr&GRid=53148084`

frontal direction (ISO 226, 2003) are displayed in figure 2.12.

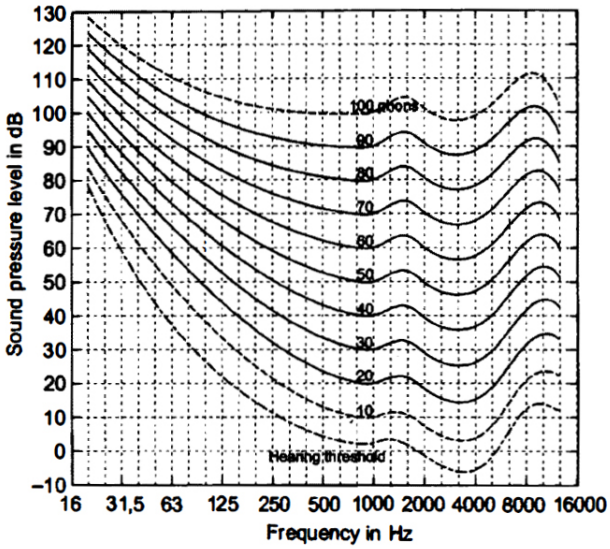


Figure 2.12: ISO 226:2003 standard equal-loudness contours (Moore, 2013, p. 135).

The loudness of sinusoidal sounds, as a function of frequency and sound pressure levels, is given by the *phon* scale (Howard & Angus, 2001, p. 84), where the number of phon of a sinusoidal sound is determined by the sound pressure level of a sound of 1 kHz which is perceived with equal loudness (Pedroso de Lima, 2012, p. 294).

The loudness of complex sounds, that is, sounds composed of more than one frequency, as already defined in section 2.1.3.1, is perceived in a slightly different way when compared to that of single sinusoidal sounds. In the inner ear (see section 2.4), each partial of an incoming complex sound leads to a displacement of the basilar membrane in the cochlea at a particular place or position, which is the basis of the place analysis of sound by the hearing system (Howard & Angus, 2001, p. 73–74). Nevertheless, the distinction that we can or cannot make between two partials presented simultaneously with close frequencies and similar amplitudes depends on the absolute separation or not of the basilar membrane displacements caused by each of the two partials. If the frequency difference between both partials is less than approximately 12.5 Hz, then a phenomenon called *beats* is normally perceived by the majority of listeners, which consists of the perception of a fused sound, whose frequency is equal to the arithmetic mean of both frequencies (Henrique, 2007, pp. 218–219), and which enables the sensation of a number of regular amplitude variations per second equal to the frequency difference, with an amplitude variation between the sum and the difference of both individual amplitudes. In addition, if the frequency difference is a little bit greater than about 15 Hz, then a rough but still fused sound is heard, and this roughness continues to be felt even when the sounds are already starting to separate themselves from each other. From a certain even higher frequency difference, the partials are heard separately and the rough sensation becomes smooth. It should be noted here that there is no exact frequency difference at which these changes in perception occur for every listener, and thus the frequency difference between the sinusoidal sounds at the point where a listener's perception changes from rough and separate to smooth and separate is known as the *critical bandwidth* (Howard & Angus, 2001, pp. 75, 228). Thus, two sinusoidal sounds are perceived as separate sounds only when their frequency difference is greater than the

critical bandwidth. The human ear has about 24 critical bands within the range of audible frequencies, whose bandwidth increases with frequency (Henrique, 2007, p. 877) (cf. item *b)* in this sub-chapter).

Complex sounds are therefore divided by the basilar membrane into frequency bands based on critical bands, where the loudness of a sound within a critical band is independent of the number of partials as long as their total intensity is constant (Howard & Angus, 2001, p. 89). However, when the partials of the complex sound spread out over more than one critical band, the brain appears to add their individual responses together, a process called *loudness summation*, leading to an increase of the perceived loudness of the sound even though the total intensity does not change (Moore, 2013, pp. 140–143) (see figure 2.13).

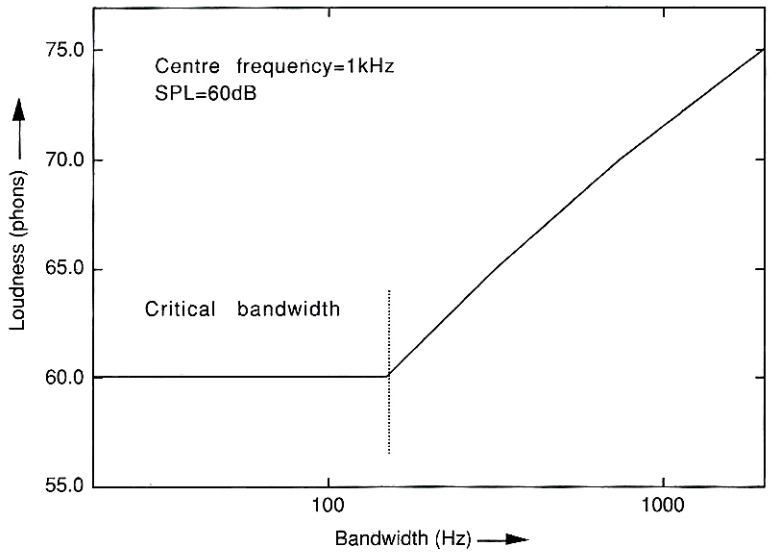


Figure 2.13: Loudness perception of a complex sound at a constant intensity level as a function of its bandwidth (Howard & Angus, 2001, p. 90).

Moreover, the human ear system has the capacity to filter or extract specific information from a complex sound, such that for example a particular conversation among multiple conversations taking place simultaneously with the same level in a reverberant environment (Henrique, 2007, p. 874) (see section 2.3.2) can be attended to. This process is known as the *cocktail party effect* (see section 2.5.6) and in this case the perception of loudness tends to be based on what was attended to.

Other circumstances can affect loudness, such as the exposure to excessive sound pressure levels above 90 dB, which can cause the loss of hearing acuity and a decrease of the auditory sensibility through auditory adaptation and auditory fatigue (see section 2.5.4.1.2). Auditory adaptation is defined as a reduction in the apparent magnitude of loudness during the first minutes of presentation, followed by a period of time in which the apparent magnitude remains roughly constant (Moore, 2013, p. 152). Auditory fatigue has to do with a temporary threshold shift (TTS) or with a permanent threshold shift (PTS) which can occur as a consequence of the exposure to these high level sounds (Henrique, 2007, p. 850). In the former case, there is a temporary hearing loss reflected in a temporary change of the threshold of hearing for a certain frequency, which usually disappears after a few hours or sometimes days. In the latter situation, tinnitus is the consequence, that is, the listener hears permanently buzzes or hisses, even in the absence of sound in the ear input. The loss of

hearing acuity consists of the increase of the critical bandwidths, which endangers the separation of different partials of a complex sound, reducing for example the ability of the hearing system to understand speech, or decreasing its selective capacity (Howard & Angus, 2001, p. 92). However, predictability can alter loudness perception, too, that is, it is quite different if we know that we are going to hear a strong and sudden sound or not (Henrique, 2007, p. 870).

A decrease of the auditory sensibility is also felt with age, essentially at higher frequencies, a phenomenon which is more marked for men than for women (Howard & Angus, 2001, p. 80).

Loudness can be changed by *masking* sounds as well, where each partial of a complex sound can be more or less difficult or even impossible to perceive due to the presence of another partial called *masker* (Howard & Angus, 2001, p. 231). The threshold of a partial can be shifted when the masker is simultaneously present or the partial is simply ignored when it is completely masked by another partial. Henrique (2007, p. 880) states that sinusoidal sounds of close frequencies are more easily masked than distant frequencies and that a sinusoidal frequency masks other partials of higher frequencies in a more efficient way than those of lower frequencies. In addition, a partial with higher intensity can mask a larger range of frequencies. Nevertheless, non-simultaneous masking, known as forward masking or backward masking, can also occur. In the first case, if one sound stops 20 to 30 milliseconds before another starts, the first one can mask the other, because the hearing cells which are stimulated earlier do not have the same sensibility as others still at rest. Relatively to backward masking, a sound can mask another, if it starts up to 10 milliseconds after the first stops.

As an attempt to derive scales relating the physical magnitudes of sounds to their subjective loudness, by asking subjects either to estimate the magnitude of sound levels according to their perceived loudness (magnitude estimation method) or to adjust the level of a test sound until it has a specified loudness (magnitude production method), Stanley Smith Stevens (November 4, 1906 - January 18, 1973)[13], an American psychologist, proposed in 1957 that the relationship between sound intensity (stimulus) and loudness (sensation) for pure tones can be expressed by:

$$S = k \times I^{0.3} \Leftrightarrow S = k \times p^{0.6} \ ,$$

where S stands for loudness, k is a constant of proportionality which depends on the subject and which governs the size of units, I is the sound intensity, and p is the sound pressure. Thus, loudness of a certain sound is proportional to its intensity raised to the power of 0.3 (Pedroso de Lima 2012, pp. 291–292; Moore 2013, p. 137), or to its pressure raised to the power of 0.6, since intensity is proportional to the square of pressure (see section 2.1.3.1). Furthermore, if intensity is increased by a factor of 10, which corresponds to an increase of 10 dB in sound intensity level or in sound pressure level (in free field conditions for a single pressure wave from the sound source), then loudness is approximately doubled or increased by a factor of about 2 (Pedroso de Lima 2012, p. 294; Moore 2013, p. 137):

$$10^{0.3} \approx 2 \ .$$

The unit of loudness suggested by Stevens is the *sone*, which is arbitrarily defined as the loudness of a 1 kHz sinusoidal sound presented binaurally from a frontal direction in free field at a sound pressure level of

[13] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Stanley_Smith_Stevens`

Table 2.1: Phon conversion to sone at 1 kHz

Phon	40	50	60	70	80	90	100	110	120
Sone	1	2	4	8	16	32	64	128	256

40 dB (Moore, 2013, p. 137), which in turn corresponds to 40 phon for any frequency along the earlier 40 phon Fletcher-Munson equal-loudness curve (Pedroso de Lima, 2012, p. 295), nowadays the 40 phon ISO 226:2003 equal-loudness curve (see figure 2.12), a level below which Stevens' power law does not hold (see figure 2.14). So, a sinusoidal sound of 1 kHz at a sound pressure level of 50 dB compared to another at 40 dB is considered to be about twice as loud, that is, to have a loudness of 2 sone. When comparing the same sinusoidal sound of 1 kHz at a sound pressure level of 60 dB with another at 40 dB, the former is said to have 4 sone or to be more or less four times as loud as the latter, and so on (see table 2.1). Nevertheless, if a 1 kHz sound at a level of 60 dB is compared with another at 50 dB, that is, with a difference of 10 dB, then it is again judged to be approximately twice as loud and it has a loudness of 2 sone. For frequencies other than 1 kHz, the measurements in sone need to be calibrated according to the frequency response of the ear (Pedroso de Lima, 2012, p. 295).

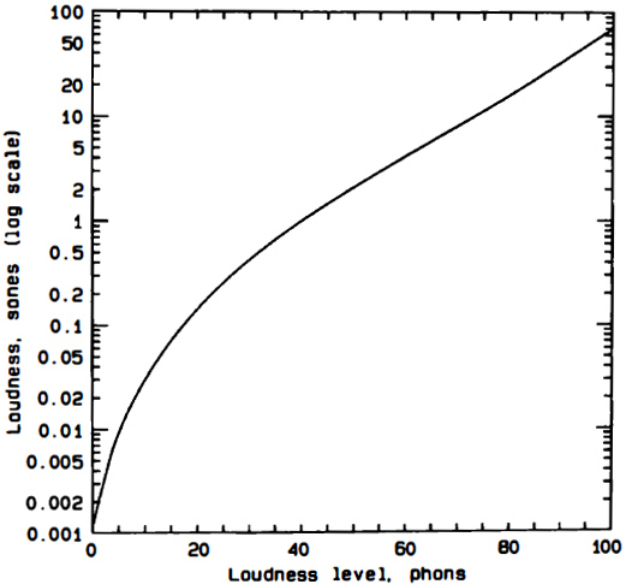


Figure 2.14: Relationship between loudness in sone and loudness level in phon for a 1 kHz sinusoidal sound presented binaurally in free field (Moore, 2013, p. 138).

Measurements of the loudness of complex sounds can be made with the help of sound level meters, which in this situation take the shapes of equal-loudness contours into consideration. Thus, the sound pressure level is weighted at each measurable frequency according to the shapes of the equal-loudness contours as accurately as possible, followed by their addition from the lowest to the highest frequency (Stark, 2002, p. 61; Moore, 2013, p. 136). However, different weightings, represented by corresponding weighting curves, such as the A, B, and C-weighting curves in figure 2.15, are used depending on the overall sound pressure level and on the type of sound. This way of measuring loudness gives therefore an approximate notion of the actual hearing response, because it is based on statistical averages.

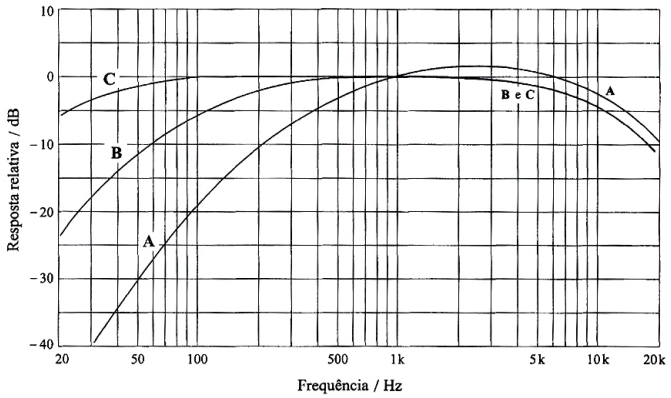


Figure 2.15: A, B, and C-weighting curves (Henrique, 2007, p. 250).

The A-weighting curve is based on the 40-phon equal-loudness contour, which means that the measured loudness value, given in the unit dBA, of a sound with a global low sound pressure level between 20 dB and 55 dB, already reflects the reduction that the A-weighting curve introduces in the result due essentially to the actual little sensibility of the ear to low-frequency components at low levels. The B-weighting curve is based on the 70-phon equal-loudness contour, and was thought to be used with sounds with global sound pressure levels between 55 dB and 85 dB, where the loudness is measured in the unit dBB. The C-weighting curve is based on the 100-phon equal-loudness contour, and is used to measure the loudness, given in the unit dBC, of sounds with global sound pressure levels of 85 dB to 140 dB, where the actual contribution of all frequencies to the total loudness is almost identical (Stark, 2002, pp. 61–62; Henrique, 2007, p. 250). Nevertheless, the A-weighting is recommended in measurements of loudness of any sound at any sound pressure level, in order to maintain the consistency between measurements, although actual hearing response changes with sound pressure level (Stark, 2002, p. 62; Howard & Angus, 2001, p. 85).

b) Pitch of a sound is another listener dependent attribute that cannot be measured directly and which represents the auditory sensation that allows us to sort sounds from low to high frequencies (Henrique 2007, p. 862; Pedroso de Lima 2012, pp. 159, 299), as it is also stated in the American National Standard Acoustical Terminology ANSI S1.1-1994 (Plack & Oxenham, 2005, p. 1). Although in the case of a sinusoidal sound it is tightly associated with its frequency, such that the higher the frequency, the higher the pitch, in the case of a complex sound a subject is asked to match the pitch of a sinusoidal sound to the pitch of a complex sound, so that the frequency of the sinusoid is then taken as a measure of the pitch of the complex sound (Moore, 2013, pp. 3–4). In addition, humans perceive the sounds of the musical notes of a scale on a keyboard as repeating once every twelve keys, which leads to the recognition that pitch has two dimensions that are respectively referred to as *pitch chroma* and *pitch height* (Warren et al., 2003, p. 10038) (see figure 2.16).

Whereas pitch chroma is used in tracking the information conveyed by a specific sound source (see section 2.2), pitch height is used in the segregation of sources, that is, it allows a listener to perceive that one source is higher than another (Warren et al., 2003, p. 10038).

However, there are other factors beyond frequency that contribute to define the pitch of a sound, such as duration, frequency spectrum, envelope, intensity (see section 2.1.3.1), the presence of other sounds, the age,

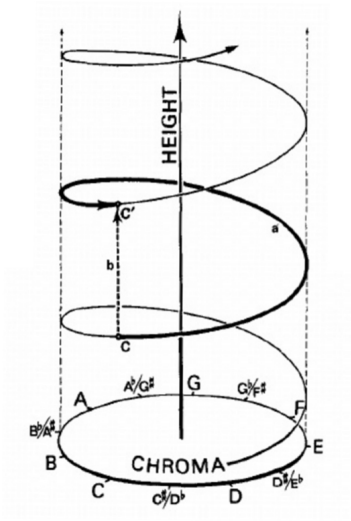


Figure 2.16: Representation of pitch chroma and of pitch height by a simple regular helix (Shepard, 1982, p. 308).

and musical training.

There are audible and inaudible sounds. With regard to the former, the average human hearing system can perceive sounds in the frequency range from around 16 Hz to approximately 20 kHz (Arau 1999, p. 12; Henrique 2007, p. 167), as already referred to in section 2.1.2. In the latter case, which is beyond the scope of this thesis, inaudible sounds fall into two categories, namely infrasounds and ultrasounds. While infrasounds are sounds with frequencies lower than 16 Hz, ultrasounds are sounds with frequencies higher than 20 kHz.

The ratio between the highest and lowest audible frequency, or acoustic interval i_a, is equal to:

$$i_a = \frac{f_{high}}{f_{low}} \Leftrightarrow i_a = \frac{20000}{16} = 1250 \ .$$

This corresponds to almost 10 octaves, where an octave is an acoustic interval between 2 frequencies, in which a frequency has twice the value of the other, represented by a ratio of 2 to 1 (Henrique, 2007, p. 926):

$$i_a = \left(\frac{2}{1}\right)^n \Leftrightarrow 1250 = 2^n \Leftrightarrow \log_{10} 1250 = \log_{10} 2^n \Leftrightarrow \log_{10} 1250 = n \times \log_{10} 2 \Leftrightarrow n = \frac{\log_{10} 1250}{\log_{10} 2} \Leftrightarrow n \approx 10.29 \ ,$$

where n is the number of octaves.

Nevertheless, the upper audible limit of 20 kHz gradually reduces to about 8 kHz due to age, a process called *presbyacusis* (Howard & Angus, 2001, p. 80) or *presbycusis* (Rossing 1990, p. 66; Howard & Angus 2001, p. 80), which is more marked for men than for women and that accompanies the decrease of the auditory sensibility with age, as referred to before in this section.

For steady sinusoidal sounds with slightly different frequencies that are presented one after the other, to avoid beats, at sound pressure levels around 60 to 70 dB, the JND relatively to frequency is smallest at low frequencies and increases monotonically with frequency, that is, it is equal to about 1 Hz at a frequency of 500 Hz, approximately 2 Hz at a frequency of 1 kHz, more or less 4 Hz at a frequency of 2 kHz, and so on (Moore, 2013, pp. 205–206) (see figure 2.17).

The relationship between the pitch of a sinusoidal sound and its frequency can also be described in terms

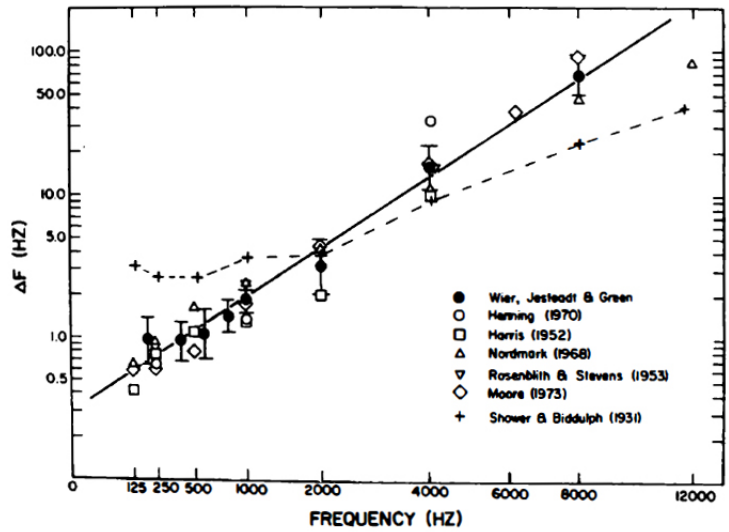


Figure 2.17: Just Noticeable Difference relatively to frequency (Moore, 2013, p. 205).

of a differential equation, identical to that applied before to the relation between loudness and intensity (Pedroso de Lima, 2012, p. 291):

$$dh = k \times \frac{df}{f} \ ,$$

where dh is the differential change of pitch, df is the differential increase in frequency f of the sound, and k is an estimated constant. The global sensation of pitch is then obtained by integrating the above equation:

$$\int_{h_0}^{h} dh = k \times \int_{f_0}^{f} \frac{1}{f} \, df \Leftrightarrow h - h_0 = k \times \ln \frac{f}{f_0} \ ,$$

where h is the sensation of pitch after the change in frequency, h_0 is the minimum sensation of pitch before the change in frequency, k is an estimated constant, f is the frequency of the sound after its change, f_0 is the original frequency before its change, and $\ln$ is the natural logarithm. Thus, a ratio of frequencies matches a difference in pitch or an interval in musical terms.

In this context, there have been attempts to derive scales relating the physical magnitudes of sounds to their subjective pitch in a similar way as described before in this section for loudness, by asking subjects to estimate the magnitude of sound frequencies according to their perceived pitch (magnitude estimation method) or to adjust the frequency of a test sound until its pitch gave the impression of being twice or half that of another frequency determined by the researcher (Moore, 2013, p. 215). The resulting *mel* scale, proposed by Stanley Smith Stevens, John Volkmann and Edwin B. Newman in 1937, and whose name was taken from the root of the word *melody*, has a reference point where the pitch value of a sinusoidal sound of 1000 Hz at a sound pressure level of 40 dB is equal to 1000 mel. Nevertheless, pitch in mel seems to be related to the *Bark* scale, proposed by the German acoustics scientist Karl Eberhard Zwicker (January 15, 1924; November 22, 1990)[14] in 1961, where 1 Bark unit, named after the German physicist Heinrich Georg Barkhausen (December 2, 1881 – February 20, 1956)[15], is equal to the bandwidth of a critical band in the inner ear (see section 2.4). This

[14] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Eberhard_Zwicker`
[15] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Heinrich_Barkhausen`

scale varies therefore from 1 to 24, since there are about 24 critical bands (Pedroso de Lima, 2012, p. 300) (cf. item *a)* in this sub-chapter).

For sinusoidal sounds with frequencies below approximately 2 kHz, pitch tends to decrease if level is increased. On the other hand, pitch increases with increasing level above more or less 4 kHz. However, in the frequency range of 1 to 2 kHz, changes in pitch due to a level increase are of the order of 1%, whereas for lower and higher frequencies changes can be of up to 5% (Moore, 2013, p. 213). This phenomenon began to be studied in 1935 by Stevens and is therefore known as Stevens' effect (Henrique, 2007, p. 863).

In addition, from the musical point of view, it seems that the sense of melody or musical interval is evoked only with a sequence of sinusoidal sounds with frequencies below about 5 kHz, having pitch chroma, although differences in frequency can be heard above this frequency (Moore, 2013, pp. 213, 241).

When the ear is confronted with complex sounds with a fundamental frequency up to about 200 Hz containing many harmonics, perception of pitch tends to be dominated by the lower harmonics, usually up to the fifth harmonic (Moore 2013, pp. 216, 232), that is, the pitch that will be perceived is the lowest common factor in these components, the fundamental frequency, even if it is weak or missing. The number of dominant harmonics decreases as the fundamental frequency increases. This phenomenon, where the fundamental frequency is missing, results in a slight change in the timbre of the complex sound, and is called *virtual pitch*, because the pitch does not correspond to any partial in the complex sound (Rossing, 1990, pp. 114–115).

Although in a lesser extent than for periodic sounds, it is also possible to perceive pitch in some complex sounds with non-harmonic spectrum. This is due to a greater concentration of energy in a certain audible frequency region (Howard & Angus, 2001, p. 127).

In addition, in the presence of at least two simultaneously produced sinusoidal frequencies, other sounds can be heard as a result of their combination. One of these resulting sounds can have its frequency equal to the difference between two of the presented frequencies. As already explained before in item *a)* of this section, if this difference is less than approximately 12.5 Hz, then beats will be heard. However, other combinations are possible (Howard & Angus, 2001, pp. 228–229):

$$f_{(n)} = f_l - [n\ (f_h - f_l)] = f_l - [n\ f_d]\ ,$$

where $f_{(n)}$ is the frequency of the nth combination sound, n is an integer number starting at 1, f_l and f_h are respectively the lowest and the highest frequencies of both sinusoidal sounds, and f_d is the frequency difference between f_h and f_l. The combination sounds are always below f_l and listeners cannot hear them all.

c) Whereas loudness is related to sound intensity level or sound pressure level, and pitch is associated with frequency, timbre is defined by the American Standards Association in the *American Standards Acoustical Terminology ASA Z24.1-1951* as that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar (ASA, 1951, p. 25; Howard & Angus, 2001, p. 210, quoting the same definition, but from the 1960's version). Thus, timbre is the most difficult sound attribute to define quantitatively, because it depends on subjective correlations of all attributes apart loudness and pitch.

Furthermore, the frequency spectrum and amplitude envelope, that is, the quantity variation of partials, the relations between frequency components during the production of sound, the sound pressure level variation of

each frequency component throughout the sound, the moment at which each of the partials is produced and listened to, and the changes during the transient periods of the sound, essentially the attack, and, in a lesser extent, the release, have a great influence on timbre. The attack is important because any colouration of the direct sound by the first reflection in the local environment usually occurs after the initial part of the sound has been heard uncoloured. Another reason for its importance is that listeners can dependably perceive the timbre of sounds during the sustain time portion after the attack (Howard & Angus, 2001, pp. 214–215). If the attack and the release portions are removed from sounds produced by musical instruments, listeners cannot reliably identify them.

Timbre is therefore a word that is used for convenience to gather all these attributes of sound that can not be set objectively. Thus, it is not possible to establish a direct correspondence between timbre and any quantity, as in pitch or loudness. Nevertheless, efforts were made in order to find methods that could track timbre variations of sounds produced by musical instruments, based on their harmonic changes throughout, such as the *tristimulus diagram* described by Howard F. Pollard and Erik V. Jansson in 1982 (Rossing, 1990, p. 133; Howard & Angus, 2001, p. 224). This is a triangular graph where the energy relationship between 1) the fundamental frequency, 2) the second, third and fourth harmonics that are resolved by the critical bands (mid frequencies), and 3) the fifth and above harmonics that are not resolved by the critical bands (high frequencies), is plotted as a line, where one end corresponds to the beginning of the sound and the other end represents the approximate moment where the sustain portion of the sound starts.

d) As a subjective attribute, duration of a sound is the sensation that physical time produces on human beings. It is also referred to as psychological time and it varies from person to person and even for the same person depending on circumstances (Henrique, 2007, p. 169), such as the state of mind and health of the human being: time seems to pass very slowly when a person is more tired physically and/or psychologically, and/or does not develop too much body or mental activity; time seems to pass more quickly when the body and/or mind are in good shape, but are faced with a situation that leads them to work above the normal rate.

Duration also changes the perception of loudness, just as frequency and amplitude do. Whereas the perceived level of a sinusoidal sound does not change if its duration is longer than about 200 milliseconds, for shorter durations the perceived level decreases (see figure 2.18). However, the loudness level is perceived more by the sound level averaged over 200 milliseconds than by short amplitude peaks when these sounds vary in amplitude (Howard & Angus, 2001, pp. 88–89).

In complex sounds, the perception of loudness is likewise affected by duration in an analogous manner as in sinusoidal sounds. Therefore, if complex sounds have a duration longer than about 200 milliseconds, then the level does not change, that is, constant energy leads to constant loudness (Moore, 2013, p. 143). For shorter durations than that, the perceived level also decreases (Howard & Angus, 2001, pp. 90–91).

In order to determine the smallest detectable increase in duration, that is, the JND relatively to duration, Δt, two successive sounds with different durations have been typically presented to listeners, where the shortest has a duration represented by t. According to Moore (2013, p. 197), all studies he addressed show that if t is longer than 10 milliseconds, then Δt increases with t. Furthermore, Δt is independent of the spectral characteristics of the sounds, but increases at low sound levels.

The effect of duration on pitch in terms of the number of cycles needed for a definite distinct pitch to be

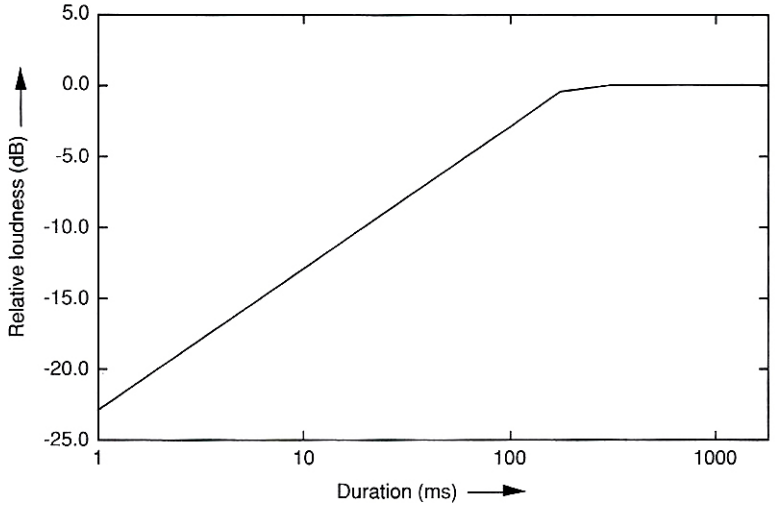


Figure 2.18: Loudness perception of a sinusoidal sound as a function of its duration (Howard & Angus, 2001, p. 89).

perceived for a given fundamental frequency is shown in figure 2.19 (Howard & Angus, 2001, p. 136).

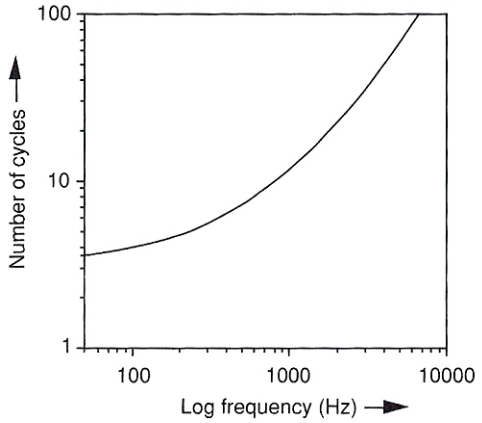


Figure 2.19: The effect of duration on pitch (Howard & Angus, 2001, p. 136).

Phase (see section 2.1.3.1) affects the time domain shape or waveform of a periodic complex sound, although the frequency domain may remain unchanged, with the amplitudes of the partials unaltered (see figure 2.20).

In many cases, the resulting sound is perceived as sounding the same, because the human ear is less sensitive to the phase of individual frequencies than to relative amplitudes (Howard & Angus, 2001, p. 54). Therefore, the insensitivity of the ear to static phase causes a particular waveform corresponding to only one timbre, whereas a specific timbre can correspond to an infinite number of waveforms (Pedroso de Lima, 2012, p. 160). Nevertheless, in non-periodic complex sounds, timbre depends slightly on phase. This effect on timbre is more pronounced when the change in phase takes place at a regular rate, as in second-order beats, which occur between frequency components of complex sounds other than the fundamental frequencies.

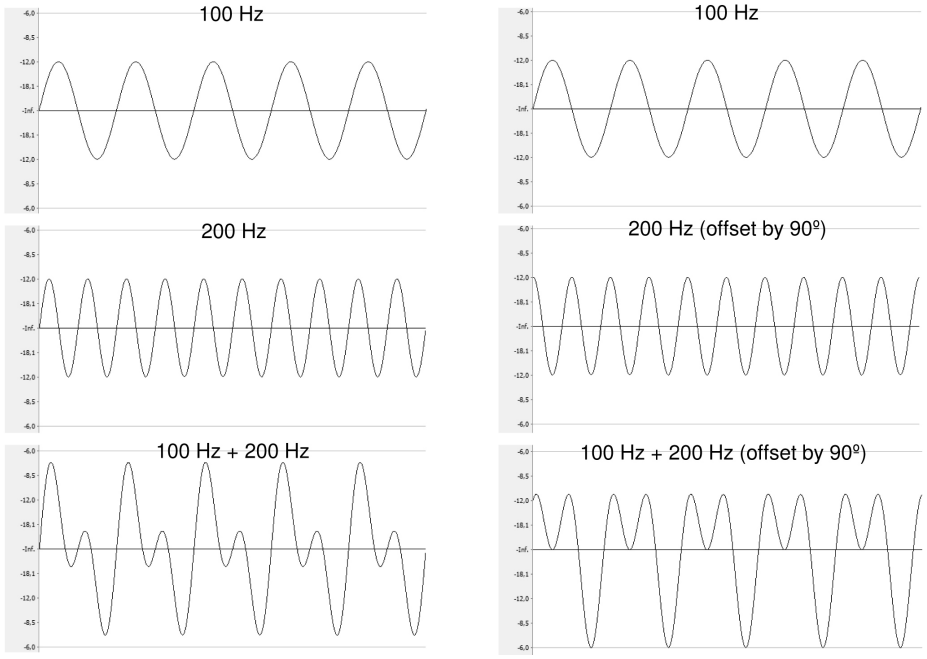


Figure 2.20: Two frequencies with the same amplitude are added together: in phase (on the left) and differing in phase by 90º (on the right).

2.1.4 Auditory Test Signals

Signals are functions of one or more independent variables which typically contain information about the behaviour and characteristics of certain physical phenomena (Lourtie, 2007, p. 3). An independent variable is a factor that is chosen to be manipulated or altered in order to obtain a result or outcome whose variation is being studied (Field, 2009, p. 7).

Any signal that can be represented as an amplitude that varies over time has a corresponding representation in the frequency domain. Thus, signals which may be used in auditory experiments, that is, auditory test signals, can be described as sound pressure p that varies as a function of time t, that is $p(t)$. Almost any function $x(t)$, where x represents sound pressure p, velocity v, voltage U, or any other variable, and whose independent variable is time t, can be decomposed into series of elementary signals, that is, very short impulses (Blauert, 1997, pp. 22–23). These functions are continuous if the independent variable time t is a real number, or discrete if the independent variable time n is an integer number. The necessity of digitally processing the continuous signals leads to a class of discrete signals which result from the sampling of continuous signals.

In a continuous time domain, the above mentioned elementary signal is the unit impulse of Dirac $\delta(t)$ (see figure 2.21), which is a fictional concept, that is, a null function for all t, except for $t = 0$, where its amplitude is infinite while the area under it remains constant and equals 1 (Lourtie, 2007, p. 14).

As the energy of a signal characterizes its size and as a short impulse has its energy concentrated at a definite point in time but distributed evenly over all frequencies, in practice, an impulse or click, such as a rectangular one, with a sufficiently even distribution of energy density (energy per bandwidth or energy per width of a range of frequencies), in the audible frequency range from 16 Hz to 20 kHz (see section 2.1.2), is attained if its duration is less than about 25 $\mu\mathrm{s}$ (Blauert, 1997, pp. 23-24) (see figure 2.22).

So the continuous signal $x(t)$ as a function of unit impulses of Dirac may be represented by the superposi-

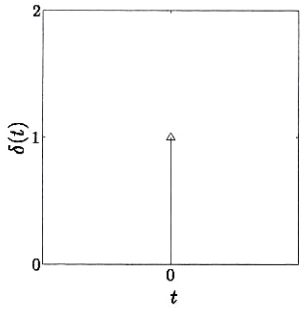


Figure 2.21: Unit impulse of Dirac in theory (Lourtie, 2007, p. 14).

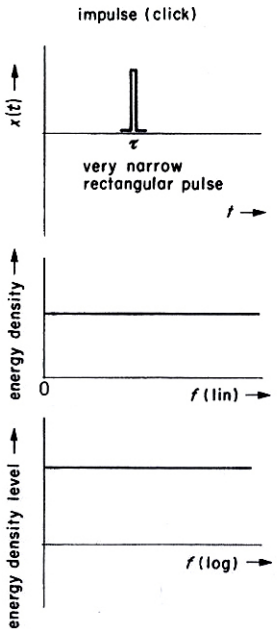


Figure 2.22: Unit impulse of Dirac in practice (Blauert, 1997, p. 23).

tion integral:

$$x(t) = \int_{-\infty}^{+\infty} x(\tau)\ \delta(t-\tau)\ d\tau\ ,$$

where $\delta(t-\tau)$ represents the unit impulse of Dirac at time t (Blauert, 1997, p. 23).

In a discrete time domain, the elementary signal is called unit impulse $\delta(n)$ (see figure 2.23). Therefore, any discrete signal $x(n)$ as a function of unit impulses may be represented by the superposition integral:

$$x(n) = \sum_{k=-\infty}^{+\infty} x(k)\ \delta(n-k)\ ,$$

where $\delta(n-k)$ represents the unit impulse (Lourtie, 2007, p. 20).

The continuous function $x(t)$ can also be decomposed into series of individual frequencies, that is, sinusoidal time functions, by the Fourier integral:

$$x(t) = \int_{-\infty}^{+\infty} \underline{X}(f)\ e^{j2\pi ft}\ df\ ,$$

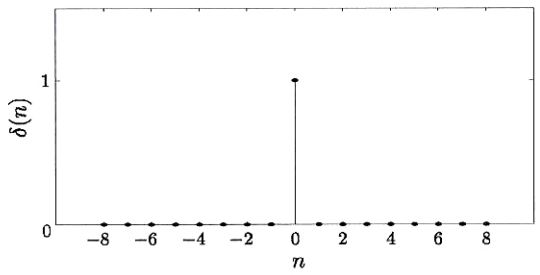


Figure 2.23: Unit impulse in a discrete time domain (Lourtie, 2007, p. 19).

where $\underline{X}(f)$ is the complex Fourier spectrum of the signal and $e^{j2\pi ft}$ represents a sinusoidal signal of frequency f and amplitude 1 (Blauert, 1997, p. 24).

For its part, a sinusoidal signal or pure tone, as it is also called, has an infinite duration and as such infinite energy. As a consequence, "its energy density cannot be defined meaningfully" (Blauert, 1997, p. 24). To solve this problem, power density or power per bandwidth, that is, energy transfer per unit time per width of a range of frequencies, measured by means of a bandpass filter, is determined instead (see figure 2.24).

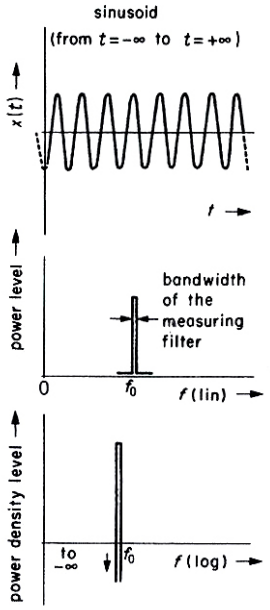


Figure 2.24: Sinusoidal signal (Blauert, 1997, p. 23).

"... the more concentrated the energy is in time, the greater its bandwidth in the frequency domain" (Blauert, 1997, p. 24). So, if energy concentration is necessary in a certain system at one point in the time domain, that is, if a large bandwidth in the frequency domain is desired, then impulses are applied. On the other hand, sinusoidal test signals are used when energy or power concentration is necessary at a certain point in the frequency domain: "... the more concentrated the energy is in the frequency domain, the more indefinite the signal in the time domain" (Blauert, 1997, p. 24).

Another test signal that is also used in auditory experiments is the Gaussian tone burst due to its optimum compromise between energies in the time and in the frequency domains (see figure 2.25).

Due to the limited energy that a system might admit and produce from a single impulse, this impulse can

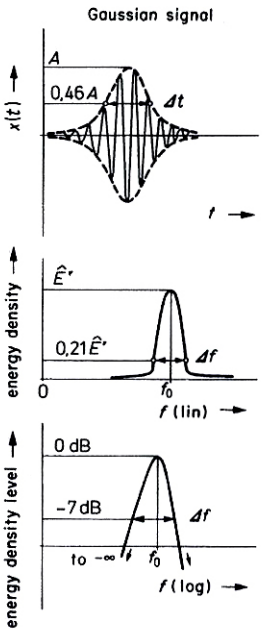


Figure 2.25: Gaussian Tone Burst (Blauert, 1997, p. 23).

be substituted by a series of short impulses of any convenient duration with equally probable mathematical sign and temporal spacing between each pair, that is, by white noise (Blauert, 1997, p. 25) (see figure 2.26). White noise, so called by analogy with white light (Pedroso de Lima, 2012, p. 82), has equal power density in any band of a given constant frequency bandwidth and a Gaussian distribution (Blauert, 1997, pp. 25–26), that is, it has a flat power spectral density which is directly proportional to $1/f^0$, where all frequencies have the same power in a characteristic random distribution without any specific trend to self-similarity (Pareyon, 2011, p. 241). For example, in white noise, the power is the same in the frequency interval between 100 Hz and 140 Hz as in the frequency interval between 5000 Hz and 5040 Hz.

The human auditory system analyses sound signals more or less by means of constant relative bandwidth, that is, the ratio of the bandwidth to the centre frequency is constant (Blauert, 1997, pp. 26–27). For that reason, pink noise is used in auditory experiments and in audio engineering as a reference signal, since pink noise has the same power in bands that are proportionally wide (see figure 2.26). For example, in pink noise, the power is equal in the frequency interval between 100 Hz and 200 Hz as in the frequency interval between 1000 Hz and 2000 Hz, where the doubling of frequency is called interval of an octave, as already referred to in section 2.1.3.2. Its name results from the fact that it is an intermediate case between white noise and red noise (see figure 2.26), the latter being most commonly called Brownian noise (Pedroso de Lima, 2012, p. 82). Red noise has a large predominance of low frequencies as is the case of red in the visible frequency range, that is, the power spectral density is directly proportional to $1/f^2$, decreasing 6 dB per octave (Pareyon, 2011, p. 241). Compared with white noise, the spectral power density of pink noise decreases 3 dB per octave, which is the same as saying that its power spectral density is proportional to $1/f$. Therefore, pink noise is often called $1/f$ *noise* (Wikipedia, 2015a; Pareyon, 2011, p. 239).

Normalized pink noise has usually a Crest Factor (CF) value or peak to average ratio of about 4, that is, a value which is the result of the division of its highest peak or maximum amplitude value (see section 2.1.3.1) by

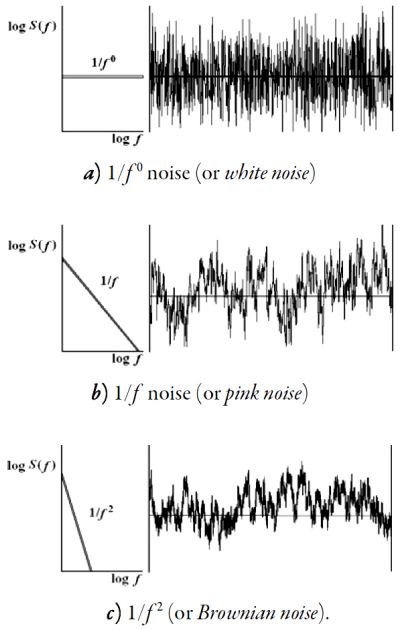


Figure 2.26: White, pink and Brownian noise (Pareyon, 2011, p. 240).

its Root Mean Square (RMS) value. The RMS value is an average of signals with positive and negative values, such as sinusoidal, white noise, pink noise, Brownian noise, speech, and music signals[16]. In logarithmic units (decibel), the CF level of normalized pink noise is equal to $20 \times \log_{10} \mathrm{CF} = 20 \times \log_{10} 4 \approx 12.04\ \mathrm{dB}$, which corresponds to the level difference between the highest peak or maximum amplitude value in logarithmic units (decibel) and its RMS value, in logarithmic units (decibel) as well. Anyway, the CF depends on the signals' waveform, varying normally from 1 to about 10 or more. Whereas in periodic signals it is constant, such as in a sinusoidal signal, in which it is always equal to $\sqrt{2}$ (a level of about 3 dB), in non-periodic signals, such as noise, speech, and music, the CF varies usually to higher values.

Summing up, in auditory experiments, white noise, pink noise, and other white noise derived random functions of any desired bandwidth and power density are used to simulate speech and music (Blauert, 1997, p. 25–27). According to Stern et al. (2005, pp. 2, 5, 9), clicks, sinusoidal signals, bandpass noise, broadband noise, and amplitude-modulated tones have been used in most binaural (see section 2.3.4) listening studies until fairly recently, instead of more stimulating and applicable signals such as speech or music. In this sense, Power et al. (2013, p. 3) refer that pink noise and speech have been used extensively in localization tests, for example by James L. Barbour (in 2004), Judith Liebetrau et al. (in 2007), Masahiko Naoe et al. (in 2008), and Florian Keiler and Johann-Markus Batke (in 2010). Power et al. also add that Liebetrau et al. report that speech shall be used in localization tests, because our ears are very sensible to it. Frank et al. (2008, p. 3) explain that pink noise is used due to its large bandwidth, which offers many localization cues.

[16] Visit `https://meyersound.adobeconnect.com/_a838360253/p3ua521culo/?launcher=false&fcsContent=true&pbMode=normal`, available on 29/10/2015, for a more in detail explanation about the terms *crest factor* and *RMS*.

2.2 Sound Sources

Whereas Choi (2000, p. 145) refers that any musical instrument can be considered as a sound source, Pierce (2014, p. 86) goes further and mentions that whatever produces an acoustic wave is called a sound source. Nevertheless, sound sources can be specified, so that we can distinguish between physical and perceptual sound sources, which will be described in the two following sections.

2.2.1 Physical Sound Sources

A physical sound source is a real object that is able to vibrate and that gives rise to mechanical wave phenomena, where pressure variations are generated from it to the surrounding media (Pedroso de Lima, 2012, p. 81). Apart from the mechanical sound sources found in Nature, the human being has developed artificial sound sources with many different purposes, such as loudspeakers and headphones (Pedroso de Lima, 2012, p. 82), which are transducers that convert some sort of energy to sound (William A. Kuperman, 2014, p. 163), that is, loudspeakers and headphones are output transducers or sources, as opposed to input transducers or receivers, such as microphones, which essentially convert sound to an electrical signal.

Physical sound sources can be divided into point sources, line sources and flat sources (Rossing, 1990, p. 42). In theory, it is believed that, in a homogeneous three-dimensional free field, a point source or a source that is spherically symmetric, having a very small radius, produces spherical waves that propagate radially and isotropically, or uniformly, from its centre, that is, sound is propagated in all directions or omnidirectionally (P. Brown, 2008, p. 36; Pedroso de Lima, 2012, pp. 79, 86; Attenborough, 2014, p. 119). Sound sources that produce spherical waves are also "... called spherical radiators of the zeroth order, elementary radiators, or pulsating spheres" (Blauert, 1997, p. 28). In these conditions, sound intensity I is inversely proportional to the square of distance r and sound pressure p is inversely proportional to the distance r (Blauert, 1997, p. 28; Pedroso de Lima, 2012, p. 139):

$$I \propto \frac{1}{r^2} \text{ and } p \propto \frac{1}{r} .$$

Thus, if the distance from the point source is doubled, sound intensity decreases by a factor of four, as mentioned in section 2.1.3.1, and sound pressure decreases by a factor of two. Both sound intensity level L_I and sound pressure level L_p decrease therefore by about 6 dB when distance from the point source is doubled.

In theory, a line source or a source with cylindrical symmetry radiates cylindrical waves (Rossing, 1990, p. 42). In this case, sound intensity I is inversely proportional to the distance r and sound pressure p is inversely proportional to the square root of distance (Pedroso de Lima, 2012, p. 139; William A. Kuperman, 2014, p. 163):

$$I \propto \frac{1}{r} \text{ and } p \propto \frac{1}{\sqrt{r}} .$$

Therefore, if the distance from the line source is doubled, sound intensity level and sound pressure level decrease by approximately 3 dB.

A flat source radiates plane waves (Rossing, 1990, p. 42), which do not spread out with increasing distance. In this case, there is theoretically no decrease in sound intensity level and sound pressure level with distance (Malham, 1998, p. 167).

In practice, although actual sound sources are not truly point sources, line sources, or flat sources, they are similar to one of these geometries (Rossing, 1990, p. 42). Thus, radiation of sound, such as a spherical sound wave, produced by real sound sources, even as simple as they can be to resemble a point source, in a free field or in a space where one or more objects exist, quickly becomes very complex with regard to its original shape, to the frequency spectrum, and to the timbre, due to reflections, absorptions, diffractions by objects, by the air, and by the surfaces of the sources themselves (Malham, 1998, p. 167). For example, the sound field resulting from the superposition of the original sound field of any source that emits sound towards a reflecting planar wall with that of its virtual mirror image (Blauert, 1997, p. 30) is completely different from the original one. Furthermore, if the wavelengths of sounds (see section 2.1.3.1) are larger than the physical size of the surfaces which radiate them, usually at very low frequencies, then the wave resembles a spherical wave as if sounds had been radiated by a point source (Gough, 2014, p. 572). This is the case of a loudspeaker in a sealed box with the longest edge of about 30 cm, for frequencies up to around 100 Hz (Blauert, 1997, p. 30). In this case, the inverse square relationship between intensity and distance from the source holds for its direct sound.

As already mentioned in section 2.1.3.1, typical sound sources favour a certain direction while radiating sound, essentially because the wavelength becomes comparable with the size of the radiating sources, and loudspeakers are not an exception. Thus, loudspeakers have an imaginary main centre axis, referenced as 0º and starting at the acoustical centre, and that is perpendicular to their front, on which the highest sound pressure level, produced by them at a given frequency or range of frequencies they were designed for, can usually be measured. The acoustical centre "... is that point from which inverse square radiation appears to originate" (Eargle & Foreman, 2002, p. 190) or "... the apparent point in space from which the sound emits" (D. Davis & Davis, 1997, p. 105), and it is not to be confused with the physical centre of the loudspeaker, which does not usually coincide with the acoustical centre. Furthermore, loudspeakers have coverage angles which are assigned to a particular plane of radiation, usually the horizontal and the vertical plane. Thus, the coverage angle $C_{\angle}$ of ideal loudspeakers or, in practice, systems where the maximum output is in the main centre axis, assigned to a given plane of radiation at a given frequency or range of frequencies, is defined as that angle, which is formed by the coverage angle edges or secondary axes that are found by moving off-axis in an arc on either side from the main centre axis in that plane, whose relative level of 0 dB is used as a reference, until the response at that frequency or range of frequencies has dropped 6 dB (D. Davis & Davis, 1997, p. 104; McCarthy, 2007, pp. 55, 484) (see figure 2.27).

In addition, a loudspeaker directional response over frequency can be described by a beamwidth plot, which shows the -6 dB points of the coverage angle edges over frequency (1/3rd octave or 1 octave resolution) without the off-axis response, that is, it shows "... the coverage angle trends over the full range of the speaker in a single chart" (McCarthy, 2007, pp. 57, 483) (see figure 2.28).

In free field conditions, if a loudspeaker behaves as a point source and is directional, then the level of the direct sound at any given distance r:

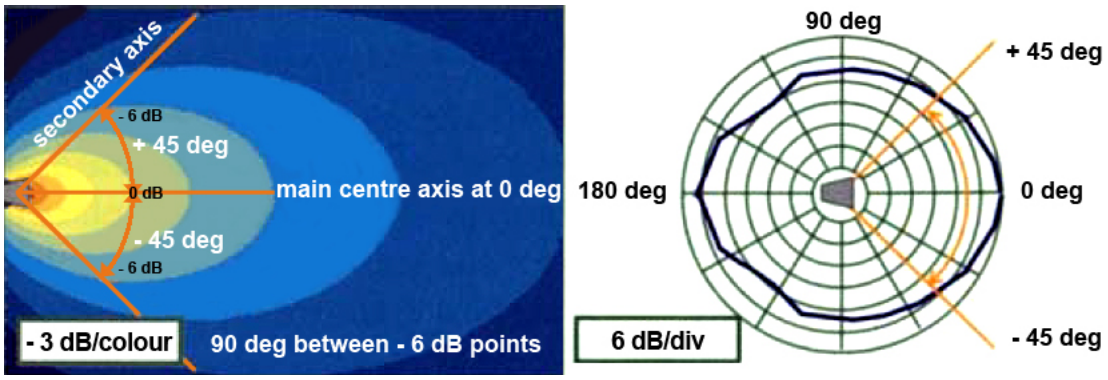


Figure 2.27: Coverage angle of 90 degrees assigned to a particular plane of radiation, shown on an equal level or isobaric contour (left) and on a polar pattern (right) of a loudspeaker (adapted from McCarthy, 2007, p. 56).

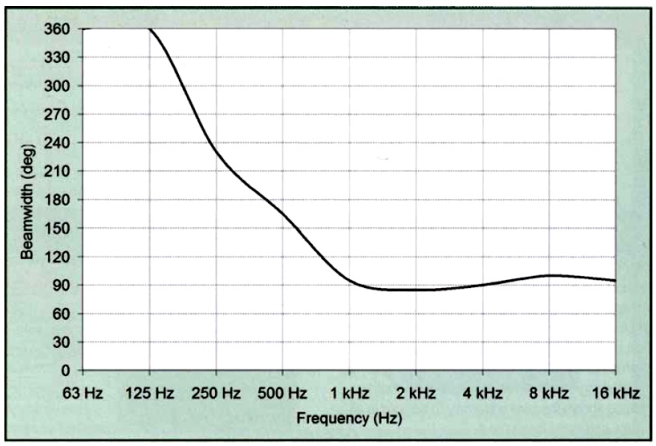


Figure 2.28: Beamwidth versus frequency of a loudspeaker with a nominal coverage angle of 90 degrees (McCarthy, 2007, p. 57).

$$L_I = 10 \times \log_{10} \frac{I}{I_{ref}} = 10 \times \log_{10} \left(\frac{\frac{W}{4\pi r^2}}{\frac{10^{-12}}{1}} \right) = 10 \times \log_{10} \left(\frac{W}{10^{-12}} \times \frac{1}{4\pi r^2} \right)$$

$$= 10 \times \log_{10} \frac{W}{10^{-12}} + 10 \times \log_{10} \frac{1}{4\pi r^2} = L_W + 10 \times \log_{10} 1 - 10 \times \log_{10} 4\pi r^2$$

$$= L_W - 10 \times \log_{10} r^2 - 10 \times \log_{10} 4\pi \Leftrightarrow L_I = L_W - 20 \times \log_{10} r - 10 \times \log_{10} 4\pi$$

is modified by the inclusion of the directivity index D_I (Attenborough, 2014, p. 119) or the directivity factor Q (D. Davis & Davis, 1997, p. 173), mentioned in section 2.1.3.1:

$$L_I \approx L_p \approx L_W + D_I - 20 \times \log_{10} r - 10 \times \log_{10} 4\pi$$

$$= L_W + 10 \times \log_{10} Q - 20 \times \log_{10} r - 10 \times \log_{10} 4\pi = L_W + 10 \times \log_{10} \frac{Q}{4\pi r^2} \, .$$

However, Chamness (1994, p. 4) points out that the use of conventional definitions of Q and $C_\angle$ can result in errors in Q and beamwidth calculations for devices that are non-ideal or non-well-behaved systems, because there are loudspeakers that do not have the highest sound pressure level at their centre axis of the

main direction. Thus, he suggests that the definition of Q, as the ratio of sound pressure squared, at some fixed distance and specified direction, to the mean sound pressure squared at the same distance averaged over all directions from the transducer (see equation in section 2.1.3.1), can be modified to be Q_{beam}, as the ratio of the mean sound pressure squared averaged over the coverage angle, at some fixed distance and specified direction, to the mean sound pressure squared at the same distance averaged over all directions from the transducer. Then, the efficiency with which $C_{\angle}$ matches Q_{beam}, or the Directivity Figure of Merit (DFM) of the loudspeaker (D. Davis & Davis, 1997, p. 594), is then expressed by dividing the calculated Q_{beam} by Q of an ideal device (which has a perfectly flat response in the coverage angle, and no output beyond that), and by multiplying it by 100%, making the result independent of the coverage angle (Chamness, 1994, p. 5):

$$\mathrm{DFM} = \frac{Q_{beam}}{ideal\ Q} \times 100\ \% \ .$$

According to McCarthy (2007, p. 488), it is also possible to set up a loudspeaker array configuration acting as a point source, by the way the most popular array type (McCarthy, 2007, p. 81), in which the axial orientation is outward from the front of the elements, thereby creating a virtual source which has a common point behind the elements.

Sound radiators that emit waves from a line rather than a point and which are approximately cylindrical in shape have been successfully constructed. Nowadays, most commercially available line source designs are made up of closely spaced discrete loudspeakers or loudspeaker systems, which are referred to as line arrays (P. Brown, 2008, pp. 37–38). Although the relationship between sound pressure level and distance from the line source is both frequency and line length dependent, the level change with increasing distance follows approximately the respective theoretical case presented before, that is, a more or less 3 dB decrease when the distance is doubled. McCarthy (2007, p. 115) refers that in a line array the axial orientation is identical (McCarthy, 2007, p. 486), where the only concern is the number of elements and their spacing.

If the radiating surfaces are much larger than the wavelength of the sound produced, that is, considered as large flat sources, then the wave resembles more a plane wave (Malham, 1998, p. 167).

Nevertheless, most sources appear to be point sources when the listener is at a sufficient distance from them (Attenborough, 2014, p. 119), although the sound field becomes more and more similar to a plane wave with distance (Blauert, 1997, pp. 29–30).

When considering musical instruments or other physical sound sources, such as loudspeakers or loudspeaker arrays, their geometry and the vibrational characteristics of the excited modes determine the directional properties of the radiated sound, which can be described by treating instruments or other sound sources as a superposition of monopole, dipole, quadrupole and higher-order multi-pole acoustic sources (Gough, 2014, p. 572) (see figure 2.29). Thus, a monopole source is equivalent to a point source, and most musical instruments or other sound sources behave like it at low frequencies. A dipole source can be represented by two similar monopole sources, 180 degrees out-of-phase with each other (see section 2.1.3.1), and very close together (Pierce, 2014, p. 90; Gough, 2014, p. 573). A quadrupole source in turn can be formed by two identical but oppositely directed dipoles brought very close together.

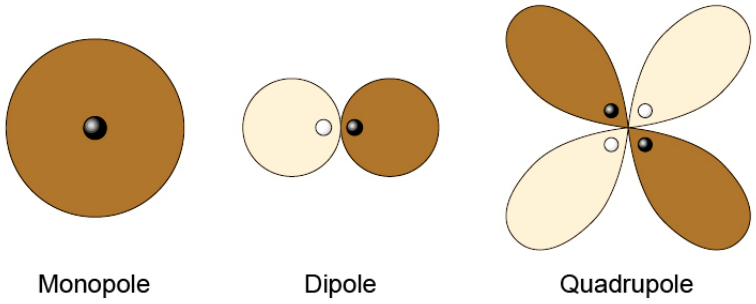


Figure 2.29: Monopole, dipole and quadrupole radiation patterns (adapted from Gough, 2014, p. 573).

2.2.2 Perceptual Sound Sources

A virtual, phantom or perceptual sound source is defined by Pulkki (2001b, p. 8) and Pulkki & Karjalainen (2001, p. 739) as an auditory object (see section 2.1.2), which is perceived in a location that does not necessarily coincide with any of the physical sound sources that produce it. In other words, loudspeakers can be used to present phantom sound sources in locations where no physical sound source or loudspeaker is otherwise present (Marentakis et al., 2008).

In this respect, Blauert (1997, p. 203) states that, when two sound sources radiate coherent sounds (see section 2.1.1), one auditory event can be perceived 1) at a position that depends on the locations of the two physical sound sources and the sounds radiated by them, or 2) at a position dominated by the place of, and by the sounds emitted by only one of, the two physical sound sources. Another possibility is that two auditory events are eventually perceived instead of one, in which case 3) the location of one auditory event depends more or less on the position of one physical sound source and on the sounds it produces, and the position of the other depends on the other physical sound source and on its sounds. In addition, 1) takes place when the levels and times of arrival of the two emitted sounds are slightly different (time difference equal to or less than about 1 millisecond), but heard as a single fused sound, leading the auditory system to the interpretation of the input sounds at both ears more or less as if they were the result of a perceptual sound source (Blauert, 1997, p. 204; A. D. Brown et al., 2015, p. 2). This phenomenon was identified by Hans Warncke in 1941 and defined as *summing localization* (see sections 2.3.4 and 2.5.5). In 2), if arrivals of the two radiated sounds at the ears are separated by a time delay of more than about 1 millisecond, but below the listener's echo threshold of about 30 milliseconds (Howard & Angus, 2001, p. 105), where the two sounds are still heard as fused, then the location of the auditory event is defined in most cases only by the location of, and by the sounds produced by, the physical sound source whose signal arrives first. This effect was called by the German electrical engineer and acoustician Lothar Cremer (August 16, 1905 - October 16, 1990)[17] in 1948 as the *law of the first wavefront*, although it is also known as *Haas effect* or *precedence effect* (Moore, 2013, p. 267) (see section 2.5.5), a term which often comprises summing localization and the law of the first wavefront in the Anglo-American literature (Blauert, 1997, p. 204). A wavefront is a surface on which the waves are in phase at every point (Kane & Sternheim, 1988, p. 551) (see section 2.1.3.1). In 3), two auditory events appear due to an excess of delay of more than approximately 30 milliseconds between the two radiated sounds, the second auditory event being

[17] Retrieved 02/01/2016, from `https://de.wikipedia.org/wiki/Lothar_Cremer`

the echo of the first one.

Nevertheless, the auditory cues of phantom sound sources do not usually match the cues of real sources. Due to some deviations in the phantom sound source cues in different frequency bands and between different cues, phantom sound sources are perceived as dispersed and cannot be produced in certain directions, which makes it very difficult to produce point-like virtual sources (Pulkki, 2001b, pp. 1, 8–9, 21). The perceived spatial spread is also dependent on the number of loudspeakers that are used to produce them (Pulkki, 1999, p. 1) (see section 2.3.4). As a result, the perceived size of the phantom sound source is usually bigger than that of the actual source. If two loudspeakers are close together, one auditory event appears localized diffusely (Blauert, 1997, p. 246). Sometimes, the perceptual sound sources are so diffuse that their locations are not identifiable, or they appear inside the head. As Hammershøi (2009, p. 5) mentions, nowadays most sound experiments are controlled by computers and therefore sounds are not perceived as authentic or ecologically valid any more.

2.2.3 Static Sound Sources

A static sound source can be any single physical sound producing object that is immobilized or stationary, or a phantom sound source which is produced by a single fixed loudspeaker or which is static among loudspeakers that produce it. In the former case, the perceived fixed position of the source is represented by the sound producing object itself, even if it is not visible to the listener (Hammershøi, 2009, p. 5).

As mentioned in 2.2.2, it is very difficult to produce point-like virtual sound sources. Nevertheless, if each single loudspeaker among many others is used and considered as a point-source instrument to which one sound, different from the others, is routed, the localization of the virtual sound source produced by it is more accurate than that of a sound dependent on any panning method (Thigpen, 2009) (see section 2.3.4). Furthermore, listeners anywhere within hearing distance can localize the sound at the fixed position of each loudspeaker, and in this case the loudspeakers do not have to be placed at equal distances and symmetrically around the audience, nor is it necessary for them to be of the same type, quality or size (Thigpen, 2009).

2.2.4 Moving Sound Sources

Moore (2013, p. 281) states that the human hearing system is relatively insensitive to the motion of sound sources. This is due to the fact that it requires about 300 milliseconds for the auditory localization cues to bring about perceived motion (Marentakis & McAdams, 2013, p. 2). Furthermore, the perception of auditory motion is firstly dominated by intensity cues presented to the human hearing system by a source that moves towards or away from a listener (see section 2.5.3). It is then secondly dominated by inter-aural cues (see section 2.5.4.1), and then finally by cues of the Doppler effect, a phenomenon which will be explained in the next paragraph. Still before that, it should be noted that this motion perception relationship is valid for moving sound speeds of up to 10 meters per second, and it is only at around 50 meters per second that the Doppler effect becomes more prevalent, which explains why this effect is frequently not taken into consideration in the simulation of motion in virtual applications and psychoacoustic experiments (Marentakis & McAdams, 2013, pp. 2–3). One exception is reported by Zelli (2009), according to whom the most important contribution to the simulation

of sound movement by the American composer, musician, inventor, and professor John M. Chowning (born August 22, 1934)[18] is considered to be precisely the inclusion of the Doppler effect in his 1972's work *Turenas*. Another exception is described by Rumsey (2008, pp. 642–643), in which Christos Tsakostas and Andreas Floros included the Doppler effect within the algorithm of a moving-source simulator presented in 2007 at the 123rd Audio Engineering Society Convention.

Thus, a moving physical sound source, such as the whistle of a moving train listened to at a certain distance, radiates sound that becomes louder and higher in pitch when it moves towards an observer, and softer and lower in pitch when it moves away (Henrique, 2007, p. 235; Moore, 2013, p. 28; Xiong, 2013). The original frequency of the sound source is only perceived as such by the observer during the transition between these two moments. This process, where the observed frequency of a wave depends on the relative speed of the source and the observer, is called *Doppler effect*, in honour of the Austrian mathematician and physicist Christian Andreas Doppler (November 29, 1803 - March 17, 1853)[19], who discovered it in 1842 (Wikipedia, 2016a). When the observer and the physical sound source approach each other, even if either of them is motionless having a null speed, then:

$$f_o = f_s \times \frac{c_{\text{air}} + v_o}{c_{\text{air}} - v_s} \, ,$$

where f_o is the perceived frequency by the observer, f_s is the original frequency of the physical sound source, c_{air} is the speed of sound in the air, v_o is the speed of the observer and v_s is the speed of the physical sound source (Henrique, 2007, pp. 237–238) (see section 2.1.3.1). When the observer and the sound source move away from each other, even if either of them is motionless having a null speed, then:

$$f_o = f_s \times \frac{c_{\text{air}} - v_o}{c_{\text{air}} + v_s} \, .$$

This effect can also be used by the human hearing system as a cue for the perception of distance changes (Xiong, 2013) (see section 2.5.3).

With regard to perceptual sound sources in motion, the physical sound sources are usually immobilized, and some kind of sound reproduction system, such as those described in section 2.3.4, is therefore used. According to Marentakis & McAdams (2013, p. 2), auditory motion perception is more accurate for frontal incidence, for horizontal movements, and for broadband sounds (see section 2.5). In addition, they mention that this process is improved if the radiated sounds have short transient times and if their frequency spectrum varies significantly over time. However, the perception of rotational motion ceases to be robust if the speed of rotation rises above approximately 2 rotations per second, and it gets worse in the presence of a distracting sound.

The Minimum Audible Angle (MAA) is defined as the smallest perceivable angular sound displacement from a given starting position, and the Minimum Audible Movement Angle (MAMA) is defined as the angular distance a moving sound has to traverse before its movement is perceived by a listener (Marentakis et al., 2008, p. 2; Moore, 2013, p. 276) (see section 2.5). Whereas the MAA is greater with lateral or elevated sounds, MAMA grows linearly with the speed of a moving sound. Thus, MAMA is equal to about 5º for a rate of

[18] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/John_Chowning`
[19] Retrieved 02/01/2016, from `https://en.wikipedia.org/wiki/Christian_Doppler`

movement of 15º per second, and equal to approximately 21º when the rate increases to about 90º per second (Moore, 2013, p. 276). In addition, Gröhn (2002) studied the localization of a moving virtual sound source in a spatially immersive virtual environment, and the effect of a distracting auditory stimulus, using a Vector-Based Amplitude Panning (VBAP) loudspeaker system for reproduction of sound sources (see section 2.3.4.8). As a result, Gröhn found out that the localization error of moving virtual sound sources is higher than that of static virtual sound sources, that this error is greater in a virtual room than in an anechoic chamber, and that the azimuth or horizontal localization error also increases with the presence of a distracting stimulus, although with a great variation between subjects.

Furthermore, Marentakis et al. (2008) studied the effects of the visibility of a performer's gestures on the identification of virtual sound trajectories in the concert hall. Thus, the trajectories were presented to the listeners sitting in the hall in three different ways: 1) audio only by an eight loudspeakers spatialization system, called Spat, which is an add-on to MAX/MSP system (see section 3.7), 2) audiovisual with the on-stage performer hearing sound spatialized out in the hall, and 3) audiovisual with the performer hearing sound over headphones in a binaural rendering of the spatialization in the hall from the ideal listening position. As a conclusion, the authors discovered that the identification of the trajectories by the listeners improves when there is visual feedback provided by the performer and when the performer receives binaural spatial audio feedback. Therefore, Marentakis & McAdams (2013, p. 1) have found out that visual cues from gesture control of spatialization affect the auditory movement or sound path perception.

More examples among many others of systems in which moving sound sources are used are presented in Harada et al. (1992) and Marshall et al. (2006). In the former case, the authors use a DataGlove system, which measures the bending of a finger and the respective hand position, and an algorithm that calculates the centre of gravity of six point masses related to the power of each of the six loudspeakers' outputs, according to the finger bending and hand position, so that it is possible to control sound movement in a three dimensional sound space with a gesture. The authors mentioned in the second place describe the development of a system in which adjustable parameters, such as the rotation, orientation of radiation patterns, stereo spread of a 2 channel sound source, ballistic curves, boomerang curves, pendular movements, and artificial life algorithms, used in a Virtual Microphone Control (ViMiC) spatialization system (see section 2.3.4.10), enable the movements of sound sources in a performance set-up for small ensemble.

Zelli (2009) states that, as one of the most important features of spatial music, sound movement refers both to inner and real space sound structures, which, in transition from one position to another, go through an audible change in at least one of their musical dimensions: timbre, rhythm, dynamic, or spatial parameters. In addition, he also mentions that in that context electronic simulation of sound movement is the only way to move sound around the auditory horizontal plane, and that, according to John Chowning, movement is, together with reverberation (see section 2.3.2), distance, and direction angle, one of the parameters necessary to produce a virtual space in a quadraphonic system (see section 2.3.4.5). With regard to Ambisonics (cf. section 2.3.4.7), Zelli (2009) refers that the movement of sound, particularly the circular movement, is very simple to achieve in that system.

However, Thigpen (2009) argues that it is also possible to create the illusion of a moving object using rapid successions of similar sounds in different loudspeakers that are used and considered as point-source

instruments, as described in section 2.2.3. The sound will give the impression to swing from one loudspeaker to another, and the movement will appear to be continuous if the succession of sounds is fast enough, or if the sounds overlap. Under these conditions, perceptual segregation of sounds can be a result, because two sounds radiated by two different loudspeakers are much more easily separated by the human hearing system than when they are emitted by the same loudspeaker. Instead of perceiving a single mass of frequencies, the brain perceives separate sound sources by using various cues, one of which is spatial location related with the cocktail party effect (see sections 2.1.3.2 and 2.5.6). Nevertheless, if the time difference between two sounds is in the range of about 5 to 30 milliseconds, one sound can be heard closest to the position of the physical source that reaches the ears first (law of the first wavefront - see sections 2.2.2 and 2.5.5), rather than two separate sounds or a single sound moving. Furthermore, applying various time differences to different frequency ranges of the sound can lead to the perception of them as moving or segregating in independent manners.

2.3 Sound Spatialization

Although there are some documented and known attempts that were made by composers over centuries prior to the electroacoustic and electronic era, such as the Flemish composer of the Renaissance Adrian Willaert (circa 1490 - December 7, 1562)[20], the Italian composer and organist of the late Renaissance Andrea Gabrieli (circa 1532/1533 - August 30, 1585)[21], the Italian composer and organist Giovanni Gabrieli (circa 1554/1557 - August 12, 1612)[22], the Italian composer, gambist, singer, and Roman Catholic priest Claudio Monteverdi (May 15, 1567 (baptized) - November 29, 1643)[23], the Austrian Classical composer Wolfgang Amadeus Mozart (January 27, 1756 - December 5, 1791)[24], the French Romantic composer Hector Berlioz (December 11, 1803 - March 8, 1869)[25], the Italian composer Giuseppe Verdi (October 10, 1813 - January 27, 1901)[26], and the Austrian late-Romantic composer Gustav Mahler (July 7, 1860 - May 18, 1911)[27], in order to apply spatial techniques in occidental music composition (Roads et al., 1996, p. 452; Gibbs, 2007, p. 21), it is essentially in the last decades, with the invention of the loudspeaker, that the art of sound spatialization has been a subject of great interest to many people like musicians, music composers, sound engineers, and artists, as well as to the scientific community.

As a result, several examples of sound spatialization are, nowadays, available in the film industry (Holman, 2000, pp. 12–24; Odowichuk, 2012, p. 27), in Electroacoustic (Hollerweger, 2006, pp. 25–26) and Electronic Music (Roads et al., 1996, pp. 452–454; Gibbs, 2007, pp. 134–135), in the video games industry (Castellanos, 2006, pp. 2–7; Odowichuk, 2012, p. 27), and at universities and research centres (Castellanos, 2006, pp. 2–7).

Accordingly, Marentakis & McAdams (2013, p. 1) point out the French composer Edgar Varèse (December

[20] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Adrian_Willaert
[21] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Andrea_Gabrieli
[22] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Giovanni_Gabrieli
[23] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Claudio_Monteverdi
[24] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Wolfgang_Amadeus_Mozart
[25] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Hector_Berlioz
[26] Retrieved 31/01/2016, from https://pt.wikipedia.org/wiki/Giuseppe_Verdi
[27] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Gustav_Mahler

22, 1883 - November 6, 1965)[28], the Greek-French composer, music theorist, and architect-engineer Iannis Xenakis (May 29, 1922 - February 4, 2001)[29], the French composer, conductor, writer, and pianist Pierre Boulez (March 26, 1925 - January 5, 2016)[30], the German composer Karlheinz Stockhausen (August 22, 1928 - December 5, 2007)[31], the American composer, musician, inventor, and professor John M. Chowning (to whom has already been made reference in section 2.2.4), the American composer Roger Reynolds (born July 18, 1934)[32], and the Canadian composer Barry Truax (born 1947)[33], as composers who manipulated or have been manipulating sounds in space, in the 20th and 21st centuries.

As an example, Schacher (2007, p. 358) states that in 1972 Chowning used sound events in his piece *Turenas* that were sent around the listener and space describing carefully planned lissajou trajectories, and Zelli (2009) reports that the English composer Trevor Wishart (born October 11, 1946)[34] includes sound movement as a main component in his compositional work. However, the use of spatial audio trajectories in composition as common practice does not guarantee by itself their identification by listeners in real environments, because of "... the relative inefficiency of the [human] auditory system in processing spatial information...", and due to "... the fact that most spatial audio systems are designed for the center of the listening area" (Marentakis et al., 2008, p. 1) (see sections 2.3.4 and 2.6.3).

2.3.1 Spatialization

The spatial relationship between musical performers has always been an integral part of performance practice, and antiphonal performance, where a call and response or alternate style of singing performed by a choir divided into two interacting parts is used, is extremely ancient, dating back to the biblical times (Zvonar, 2005b). Furthermore, 16th century composers at the Basilica San Marco in Venice, such as Adrian Willaert, Andrea Gabrieli, Giovani Gabrieli and others (see section 2.3), have contributed to the earliest published works in which spatial antiphony is used as a compositional technique (Roads et al., 1996, p. 452). However, from the late Baroque period to the beginning of the Romantic period there seems to have been little interest in spatial antiphony. Since then until the electronic era there are several examples of spatial placement being used for special theatrical effects (Zvonar, 2005b), but in the 1950's the spatial dimension in music is rediscovered with the separation of the physical sound sources, most noticeably instruments (Zelli, 2009), that is, the locations of performers are used as a manner to articulate contrasting layers of musical activity (Zvonar, 2005b).

In this respect, Zvonar (2005a) presents four different sorts of spatialization techniques, philosophies, or formal approaches, which have been implemented individually or combined in any way until today: 1) live performance or diffusion of sound; 2) environmental multichannel soundscape; 3) classic studio multitrack tape composition; and 4) automated location control.

Electroacoustic works that can be considered as typical of case 1) are works composed by the French com-

[28] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Edgard_Var%C3%A8se
[29] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Iannis_Xenakis
[30] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Pierre_Boulez
[31] Retrieved 31/01/2016, from https://en.wikipedia.org/wiki/Karlheinz_Stockhausen
[32] Retrieved 02/01/2016, from https://en.wikipedia.org/wiki/Roger_Reynolds
[33] Retrieved 02/01/2016, from https://en.wikipedia.org/wiki/Barry_Truax
[34] Retrieved 02/01/2016, from https://en.wikipedia.org/wiki/Trevor_Wishart

posers Pierre Schaeffer (August 14, 1910 - August 19, 1995)[35] and Pierre Henry (born 9 December 1927)[36]. Also as an engineer at the Radiodiffusion-Télévision Française (RTF), broadcaster, musicologist, acoustician, and writer, Schaeffer developed in the 1940's a genre of electroacoustic music that he called *Musique Concrète* (*Concrete Music*), which is based on recordings of sounds produced in Nature or directly or indirectly by the human being. In the 1950's, both Schaeffer and his collaborator Henry often used multiple non-synchronized mono tape decks, and a four-channel loudspeaker system, comprised of two loudspeakers placed at the front left and right of the audience, one at the rear, and another in the centre high above the audience, to reproduce their works (Zvonar, 2005a). In 1951, Schaeffer invented a mechanism called the *potentiomètre d'espace* (Zvonar, 2005a; Marshall et al., 2006, p. 360; Marshall et al., 2009, p. 228; Odowichuk, 2012, p. 28; Marentakis & McAdams, 2013, p. 1) (see figure 2.30), so that the position of one of five tape signals was controlled either to the left or right, above, or behind the audience, according to the spatial movement of a performer's hand on stage, holding a small induction coil that interacted with other four coils around the performer, while each of the other four tape signals were routed to a single one of the four loudspeakers (see also section 2.6.3). It is considered as the first sound spatialization control system (Odowichuk, 2012, p. 28).

Figure 2.30: Potenciomètre d'espace: Pierre Henry (left) and Pierre Schaeffer (right) (retrieved 04/02/2016, and adapted from `https://upload.wikimedia.org/wikipedia/en/2/2f/Pierre_Henry_in_concert%2C_1952.jpg` and `http://deaddouble.blogspot.pt/2014_04_01_archive.html`, respectively).

With regard to case 2), in 1952, the first work ever composed for eight-channel surround sound was created by the American composer, music theorist, writer, and artist John Cage (September 5, 1912 - August 12, 1992)[37]. This work was named *Williams Mix* and prepared for eight non-synchronized mono tapes, whose signals were routed to eight equally-spaced loudspeakers surrounding the audience (Zvonar, 2005a).

According to Roads et al. (1996, p. 453), *Kontakte*, composed in 1960 by the German composer Karlheinz Stockhausen, is considered as the first purely electronic composition for multitrack tape, that is, with only electronically generated signals, using a four-channel Telefunken T9 tape recorder and a four-channel loudspeaker reproduction system, and can be regarded as representative of case 3), although Stockhausen already employed a multitrack tape recorder in his 1956's *Gesang der Jünglinge*, for electronic sounds and a recorded voice of a boy soprano (Zvonar, 2005a), a composition that is often seen as the first successful combination of

[35] Retrieved 02/02/2016, from `https://en.wikipedia.org/wiki/Pierre_Schaeffer`
[36] Retrieved 02/02/2016, from `https://en.wikipedia.org/wiki/Pierre_Henry`
[37] Retrieved 02/02/2016, from `https://en.wikipedia.org/wiki/John_Cage`

the strictly electronically generated music with Concrete Music.

In 1958, the French composer Edgar Varèse presented *Poème Électronique* as part of a multimedia environment at the Philips Pavilion at the Brussels World's Fair, considered as representative of case 4), in which the three-track tape composition was synchronized with visual effects by a complex multitrack sprocketed tape system (Zvonar, 2005a,b).

To gather more in detail information about diverse environments or places where these various spatialization techniques were applied and by whom, further reading of Roads et al. (1996, pp. 452–454) and of Zvonar's (2005a) article is recommended.

Thus, in Concrete Music and Electronic Music, space is considered as the fifth independent attribute in music, beyond loudness, pitch, timbre, and duration, the four subjective attributes proposed by Rossing (1990, pp. 63 and 80) as being often used to describe musical sound, and presented in section 2.1.3.2.

Chowning's (1971) research on the simulation of the Doppler effect (see section 2.2.4) and of local and global reverberation effects for moving sounds (see section 2.3.2), at the Stanford University *Center for Computer Research in Music and Acoustics* (CCRMA), represents the turning point between the intuitive and empirical spatialization work developed in the 1950's and 1960's and the succeeding scientific approaches implemented in the computer music systems' design, in which spatialization algorithms could begin to be employed by composers in increasingly smaller and affordable computers (Zvonar, 2005a). Nevertheless, the spatialization technique of live performance or diffusion of sound introduced by Schaeffer and Henry in the 1950's developed at the same time to other levels, so that composers created orchestras of loudspeakers using special manually-controlled diffusion consoles. *L'Apocalypse de Jean*, composed in 1968 by Henry, is an example of the common practice distribution of a two-track source recording through a large number of loudspeakers of various sizes and tonal qualities (Zvonar, 2005a).

According to Zelli (2009), the American modernist composer Charles Ives (October 20, 1874 - May 19, 1954)[38], the French composer Edgard Varèse (see section 2.3), and the Canadian-born American composer Henry Brant (September 15, 1913 - April 26, 2008)[39] created the theory behind sound spatialization based on the principles of psychoacoustical perception. In this sense, Odowichuk (2012, p. 28) states that nowadays spatialization requires a physical model that incorporates psychoacoustics. Consequently, sound spatialization can be carried out in a physical space or even in a perceptual or virtual space.

2.3.2 Physical Space

Physical sounds can be produced in free field conditions, that is, in which there are no reflections but only direct sound, or in reflective spaces. The former can be experienced as close as possible in an anechoic chamber (see figure 2.31), or outdoors if the weather conditions are favourable and if one is suspended off the ground and far away from buildings (Rumsey, 2001, p. 2).

In an anechoic chamber, sound energy produced inside by a physical sound source and reaching its walls, covered with absorbing material, is absorbed to a maximum and is virtually not reflected back. Thus, considering that sound pressure level attenuation by the air depends on temperature, humidity, and pollution, and

[38] Retrieved 01/02/2016, from `https://en.wikipedia.org/wiki/Charles_Ives`
[39] Retrieved 01/02/2016, from `https://en.wikipedia.org/wiki/Henry_Brant`

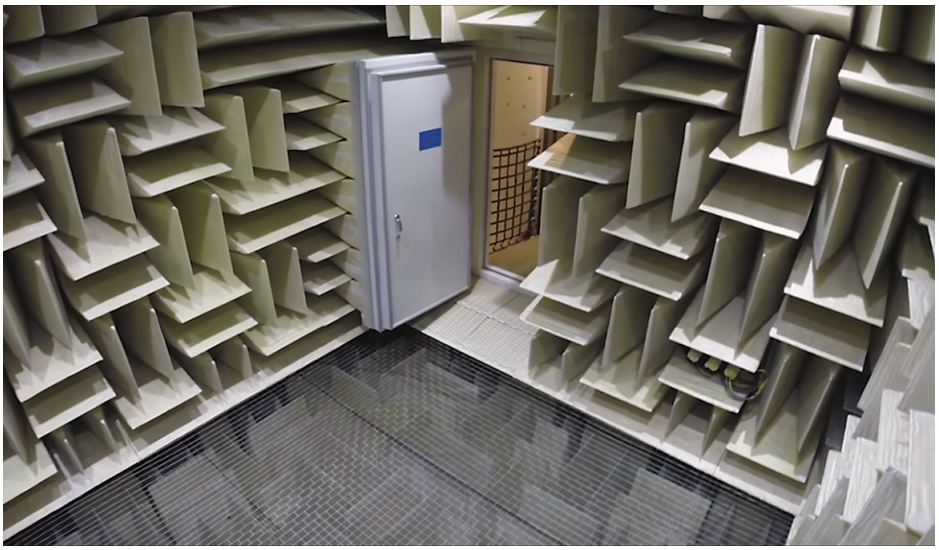


Figure 2.31: Microsoft's Eckel Anechoic Chamber at its Redmond, Washington, Audio Lab, USA, and currently the quietest place on Earth (Retrieved 07/02/2016, from `http://mms.businesswire.com/media/20151015006693/en/491446/5/Microsoft_Eckel_Chamber_highres.jpg?download=1`).

that its energy dissipates in the form of heat (Henrique, 2007, p. 228), the sound pressure level in free field conditions for omnidirectional sources drops about 6 dB per distance doubling, as already mentioned in section 2.1.3.1.

Absorption of sound energy is a phenomenon which is caused by the air, but also by other diverse materials. The sound absorption coefficient α defines therefore the amount of energy, or power, which is removed from the sound by a given area of absorbing material when it strikes it (Howard & Angus, 2001, p. 253), varying with frequency, that is, it is a quantity which compares the absorbed sound intensity I_a with the incident sound intensity I_i, and is so given by:

$$\alpha = \frac{I_a}{I_i} \Leftrightarrow \alpha = \frac{I_i - I_r}{I_i} \ ,$$

where I_r is the reflected sound intensity. Whereas α = 1 means that all the energy has been absorbed, α = 0 means that all the energy has been reflected (Henrique, 2007, pp. 229, 766).

In reflective spaces, direct sound and reverberant sound coexist, so that a percentage of the radiated direct sound produced by a physical sound source is absorbed by the air and by the surfaces within the space, and the remaining percentage is reflected back into it. After a short time, numerous reflections build up at different times to create a complex sound field, formed by the direct sound field and by the reverberant sound field, whose sound pressure level does not drop as rapidly as one moves away from the physical sound source. As a consequence, a relatively unchanging level of diffuse sound throughout the space can be measured (Rumsey, 2001, p. 5). However, the direct sound pressure level drops with distance in the same manner as in free field conditions, and at some distance, known as the critical distance D_c or room radius, it is equal to the sound pressure level of the reflected sound. This distance depends on the level of the reflected sound and on the reverberation time of the space, which is defined as the time it takes for a sound to drop by 60 dB below the sound source's original level, after the sound source ceases (Henrique, 2007, p. 772), usually represented by RT_{60}:

$$D_c = 0.141 \times \sqrt{R \times Q} \Leftrightarrow D_c = 0.141 \times \sqrt{\frac{S \times \bar{\alpha}}{1 - \bar{\alpha}} \times Q}\,,$$

where R is called the room constant, Q is the directivity factor of the physical sound source (see sections 2.1.3.1 and 2.2.1), S is the total surface area of the room in square metres, and $\bar{\alpha}$ is the average absorption coefficient of the room (Rumsey, 2001, pp. 5–6). In reflective spaces, the resulting sound is therefore much different from the sound signal originally produced by a physical sound source (Pulkki, 2001b, p. 3).

In a physical space, sound spatialization can be achieved on the one hand by producing sounds by means of physical sound sources, such as musical instruments or multichannel sound systems consisting of multiple loudspeakers distributed over a place, assuming that each one is an independent sound source or even an instrument (Hollerweger, 2006, pp. 25–26; Roads et al., 1996, pp. 452–454), that is, creating a loudspeaker orchestra placed within a performance space controlled by a diffusion mixing desk operated by a trained performer (Malham, 1998, p. 171). In this case, if there are n physical sound sources, then at least $2 \times n$ sound paths have to be considered for a normal listener hearing with the two ears (Blauert, 1997, p. 201).

On the other hand, sound spatialization can also be accomplished by using headphones, or by using loudspeakers that interact with each other in a physical space as if they were invisible, creating phantom sound sources (Hollerweger, 2006, pp. 25–26) (see section 2.2.2), or imaginary environments. Nevertheless, from the human hearing system perspective the result will be rather part of a perceptual space than a physical space (cf. section 2.3.3).

In auditory experiments concerning spatial hearing, a head-related reference system of spherical coordinates, whose origin lies halfway between the upper margins of the entrances to the two ear auditory canals (see section 2.4), is usually used (Blauert, 1997, p. 14) (see figure 2.32). Thus, the directions of physical or perceptual sound sources can be indicated by the horizontal angle or azimuth φ, and by the elevation angle δ (Pulkki, 2001b, p. 4).

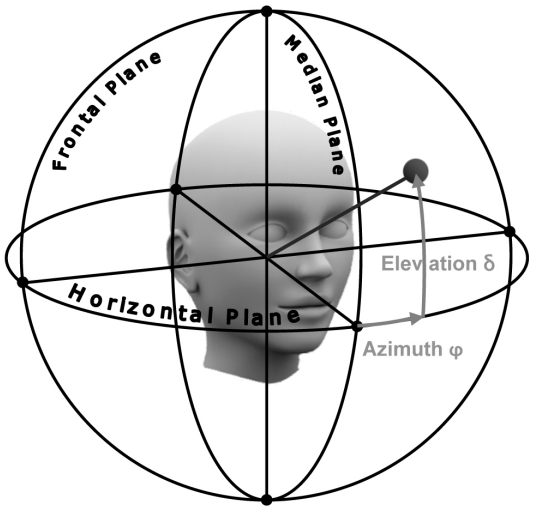


Figure 2.32: Horizontal, frontal and median planes in a head-related spherical coordinates system (adapted from `http://www.secondpicture.com/tutorials/3d/3d_modeling_of_a_human_head_3ds_max_01.html`).

In this system, the horizontal plane is considered to be parallel to the surface of the earth and defines the internationally agreed anatomical position of the skull, in which it passes through the lower margins of the eye

orbits and the upper margins of the auditory canals (Blauert, 1997, p. 14) when a subject is standing erect and facing forward in an ordinary way (Wikipedia, 2015e). In this plane, all points are at the same height from both ears (Pulkki, 2001b, p. 4).

The frontal plane is defined as a plane that intersects the upper margins of both auditory canals and which makes a right angle with the horizontal plane (Blauert, 1997, p. 14).

The median plane is defined as a plane that makes right angles with the horizontal and with the frontal planes (Blauert, 1997, p. 14), dividing the head into left and right sides. In this plane, each point is equidistant from both ears (Pulkki, 2001b, p. 4).

2.3.3 Perceptual Space

Blauert (1997, p. viii) refers that the three basic aspects of spatial hearing are the physical, psychophysical and psychological aspects, and that the most important physical parameters of spatial hearing are the acoustic signals presented to the two ears (Blauert, 1997, pp. xiii, 51, 201; statement also stressed by Pulkki, 2001b, p. 3). Spatial perception of sound is therefore dependent on the listeners' hearing system, on the way it receives the information, and on the way it interprets the information it receives.

Thus, auditory or perceptual space is defined by Blauert (1997, p. 4) as being comprised of all possible positions of auditory events or perceptual sounds (see section 2.1.2), which become more accurate with age. The position of an auditory event is most often the same as that of the physical sound source, but the position of a physical sound source is not always the same as that of the auditory event. For example, if the listener's ears receive only the reflection of a sound, reflected on a plane surface larger than its wavelength (see section 2.1.3.1), then he or she will be before a virtual sound source which is in a different position than that of the actual physical sound source.

In a perceptual or virtual space it is feasible to create imaginary environments or phantom sound sources by using loudspeakers that interact with each other as if they were not there (Gibson, 1997, pp. 8–14; Hollerweger, 2006, pp. 25–26). Other possibilities are the use of *binaural* systems, that is, systems that are thought for headphone listening, the use of *transaural* systems, so named by Duane H. Cooper and Jerald L. Bauck in 1989 to designate systems that use the same approach as binaural systems but that are modified so that loudspeakers can be used, or the use of a loudspeaker system designed to produce a precomposed sound field in a wider area (Malham, 1998, p. 171) (see also section 2.3.4).

The term *3D sound* is normally used in situations where auditory cues, that a listener uses to determine the location of a sound, are artificially created or recreated, such as in non-real environments generated by binaural systems (Audio Products Division of National Semiconductor Analog Products Group, n.d.). In this sense, interactive or passive 3D sound is possible. In the former case, the position and the attributes of the sound being localized depend on which way the head is moving. In the latter case, the position and the attributes of the sound being localized do not depend on the movement of the listener's head, so that a statically frontal positioned sound always stays in that position regardless of whether the head is moving or not. In both situations, since sounds are monophonic in Nature and not generated in a binaural form, in 3D sound, monophonic signals are used initially, and then binaural Head-Related Transfer Functions (HRTFs) (see

section 2.4.1) are calculated mathematically. Since HRTFs are discrete measurements for a given location, calculations have to be done for other locations in between by means of interpolations, so that it is possible to position a monophonic signal to almost any direction for a desired virtual source direction (Pulkki, 2001b, p. 16). Furthermore, this process has to be put into practice for each sound that is spatialized.

Odowichuk (2012, p. 27) suggests that in the creation of realistic auditory environments, where immersive sound experience can be a requirement, several features of each sound source should be controllable, such as their sound level, distance, and apparent motion. In order to achieve a perceptual or virtual space that responds to sounds as in real life, early reflections and reverberation (see section 2.3.2) should also be taken into consideration (Zelli, 2009; Odowichuk, 2012, p. 27). If early reflections are produced within about 35 milliseconds, then the human hearing system combines them into a single perceptual whole (Wikipedia, 2014).

According to Malham (1998, p. 169), the human hearing system has evolved under conditions in which it had and still has to deal with potentially ambiguous or inconsistent information, which results from its interpretation of a number of cues (see section 2.5 and more specifically section 2.5.4) gathered around from mostly complex sound fields, where the perceived direction and distance of a sound source can be interpreted as non coincident with the actual direction and distance. In addition, Malham (1998, p. 169) argues that it seems that these cues are analysed and sorted according to their ambiguity, and that this is the reason why it is possible for us to accept simplified artificial sound fields produced by sound reproduction systems developed by the human being, as those presented in section 2.3.4.

2.3.4 Sound Reproduction Systems

An important objective of sound reproduction is for many people, such as musicians, artists, music composers, sound engineers, and producers, on the one hand, to get results as close as possible to the original artistic performance sound and aesthetics, and to achieve as much as spatial fidelity as possible, which has never been totally accomplished until today, essentially due to technical, acoustical, and practical reasons in the process of recording and playback (Streicher & Everest, 1998, p. 1.1). On the other hand, another feasible aim can be to modify the original sound along the process of recording and playback for maximum emotional impact, because it is known beforehand that it is impossible to replicate the original (Linkwitz, 2015).

Thus, in either case, five different forms of sound reproduction can be specified: 1) monaural, 2) binaural, 3) monophonic, 4) stereophonic, and 5) biphonic sound reproduction (Streicher & Everest, 1998, p. 1.1) (see figure 2.33).

In 1), whereas *mono* means single, the suffix *aural* refers to the ear. Therefore, *monaural* sound reproduction is carried out via a single channel using one or even two headphones, earphones or in-ear devices driven by a common signal (Streicher & Everest, 1998, pp. 1.1, 1.2). The difference between these three types of devices is that a headphone has a large pad that can completely surround each ear in order to maximally attenuate outside noise (circumaural headphone), or a large pad that simply presses against the ear, allowing external noise to be attenuated in a lesser extent (supra-aural headphone); an earphone is a very small device that is adjusted and held in the outer ear at the entrance of the ear canal, allowing environmental noise to be also heard from the outside (considered as an intra-aural device); and an in-ear device is a very small de-

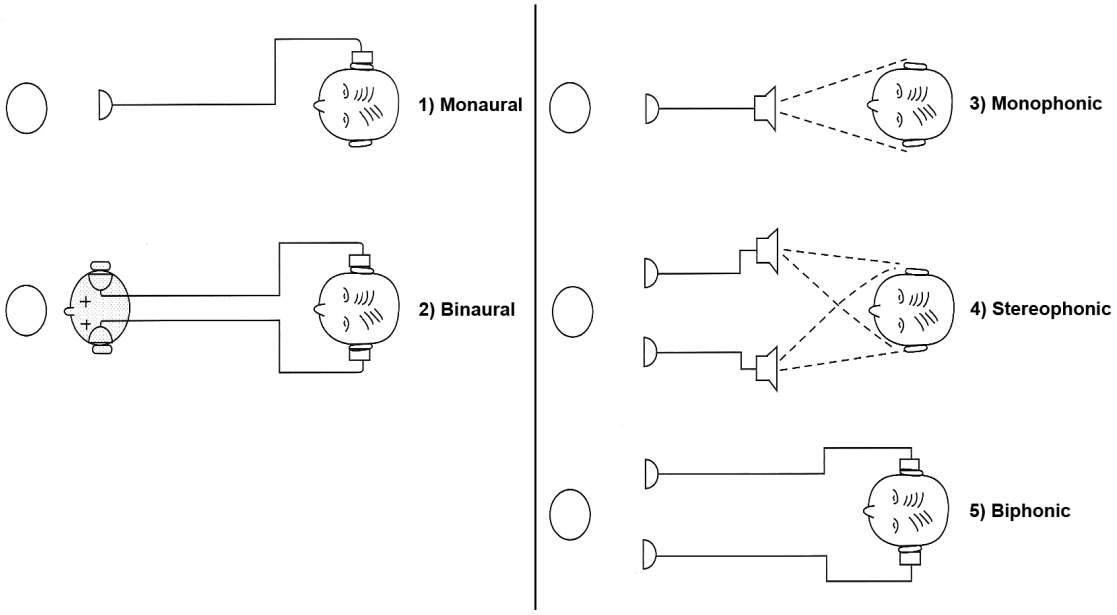


Figure 2.33: Five different forms of sound reproduction (adapted from Streicher & Everest, 1998, p. 1.2).

vice that is inserted into the ear canal, sealing it and maximally attenuating noise from the outside (intra-aural device) (Wikipedia, 2016b).

In 2), *bi* means two, and once again *aural* refers to the ear. Thus, *binaural* sound reproduction implicates the use of two headphones, earphones, or in-ear devices that are driven by two different channel signals, which are the result of sound picked up by two microphones set into both ears of an artificial head, known as *dummy head*, so that the effects caused by a head and outer ears to sound can be simulated (see section 2.4.1).

In case 3), *mono* means single, and the suffix *phonic* designates sound. Accordingly, *monophonic* reproduction is performed through a single channel as in a monaural system, but now one or more loudspeakers are connected to this single channel rather than headphones, earphones, or in-ear devices (see also section 2.3.4.1).

In 4), the Greek word *stereo* means solid (Malham, 1998, p. 173) and it is related to the three spatial dimensions of width, depth, and height, and therefore *stereophonic* reproduction uses two or more individual channels feeding two or more separate loudspeakers in order to reproduce the spatial dimensions of a performance (see sections 2.3.4.3 up to 2.3.4.11).

In 5), *biphonic* reproduction means that two headphones, earphones or in-ear devices are used to listen to recordings that were thought for the common stereophonic reproduction system with two loudspeakers. Biphonic reproduction has become frequent in recent years with the use of portable stereo players.

Whereas in situations 1), 2), and 5) a signal sent to an ear is isolated from the other ear and the natural effect of the outer ear is not taken into account because of the use of a headphone, earphone, or in-ear device, in 3) and 4) a signal can be picked up by both ears and with direct influence of the outer ears when using a loudspeaker.

Headphones have been used in almost all major psychoacoustical studies involving directional hearing, because measurements can be made easily (Malham, 1998, p. 170–171), and in most studies of binaural hearing the presentation of single sources through headphones has been made (Stern et al., 2005, p. 5) (see

also section 2.1.4). In this respect, Blauert (1997, p. 94) adds that in 1905 the German philosopher and psychologist Carl Stumpf (April 21, 1848 - December 25, 1936)[40] introduced three different ways of presenting test signals to a subject over headphones: 1) monotic presentation, 2) diotic presentation, and 3) dichotic presentation. In the first case, only one headphone receives the signal. In the second case, the same signal is sent to both headphones. In the latter case, a different signal is sent to each headphone. Thus, whereas monotic and diotic presentations can be performed in monaural sound reproduction, dichotic presentation is possible in binaural sound and in biphonic reproductions.

It should be noted here that a distinction has to be made between the terms *lateralization of sound sources* and *localization of sound sources* (Stern et al., 2005, p. 5) (see section 2.5). The former has to do with the perception of the apparent lateral position of sound sources within or near the head using headphones, as can be observed in monaural sound and biphonic reproductions, where the natural effect of the external ears and the use of Head-Related Transfer Functions (HRTFs) (see section 2.4.1) to simulate the effect of the outer ears are not taken into account (Blauert, 1997, pp. 185–187, regarding experiments carried out in 1948 by Herbert Klensch). In this sense, Streicher & Everest (1998, p. 3.9) refer that whereas coherent sounds (see section 2.1.1) seem to be perceived in the middle of the head, incoherent sounds are perceived near the ears. Localization of sound sources in turn deals with the perception of the apparent direction and distance of sound sources outside the head (Stern et al., 2005, p. 5), which can be achieved via headphones with the use of HRTFs, as is the case in binaural sound reproduction, or through loudspeakers with the direct influence of the outer ears, as is the case in monophonic and stereophonic reproduction.

Neukom & Schacher (2008) define panning as a technique of the positioning of a single or monophonic source within a stereophonic image. When using two or more loudspeakers, two different panning techniques can therefore be distinguished: amplitude panning and time panning.

The former is the most often used (Pulkki, 2001b, p. 11; Pulkki & Karjalainen, 2001, pp. 739, 740), and, according to the English electronics engineer Alan Dower Blumlein (June 29, 1903 - June 7, 1942)[41], amplitude panning is a technique in which the amplitudes of a sound being reproduced by two or more loudspeakers, equidistant from a listener, lead the listener's ears to note low frequency phase differences (see section 2.1.3.1) and high frequency amplitude differences, so that the angular direction from which the sound arrives can be determined (Blumlein et al., 1931, pp. 9–10), that is, it is a technique in which the same sound signal $x(t)$ is applied to two or more loudspeakers equidistant from a listener with appropriate amplitudes enabling the perception of a virtual, phantom or perceptual sound source (Pulkki, 1999, p. W99-1; Pulkki & Karjalainen, 2001, p. 739, 740; Pulkki, 2001a, p. 754; Neukom & Schacher, 2008) (see section 2.2.2):

$$x_i(t) = g_i \times x(t) \ , \quad i = 1, 2, ..., N \ ,$$

where $x_i(t)$ represents the signal that is applied to loudspeaker i as a function of time t, g_i is the gain factor of the channel feeding loudspeaker i, and N is the total number of loudspeakers. Thus, the sound signals from each loudspeaker reach both ears, being summed at the ear canals, a phenomenon known as *summing localization*, as already mentioned in section 2.2.2 (see also section 2.5.5).

[40] Retrieved 22/02/2016, from `https://en.wikipedia.org/wiki/Carl_Stumpf`
[41] Retrieved 07/03/2016, from `https://en.wikipedia.org/wiki/Alan_Blumlein`

In addition, amplitude panning is most often used in a two-loudspeaker stereophonic system configuration (see section 2.3.4.3), that is, in a system in which the main centre axes (see section 2.2.1) of the two loudspeakers are ideally 60 degrees apart from each other and pointing symmetrically to the listener (Pulkki, 1999, p. W99-1).

Nevertheless, amplitude panning is also used in systems where more than two loudspeakers are needed (Pulkki, 2001b, p. 1). Since in amplitude panning it is usual that only the nearest loudspeakers to the perceptual sound source are activated (Neukom & Schacher, 2008), in sound reproduction systems with more than two independent channels and more than two loudspeakers, amplitude panning methods can be used where pairs or triplets of loudspeakers are usually fed at a time, producing a satisfactory quality of static perceptual sound sources in relatively large listening areas (Pulkki, 1999, p. W99-1–W99-3). When using pairs of loudspeakers, all of them are installed in the same plane as the listener, typically in the horizontal plane (two-dimensional configuration) (see section 2.3.2), and the perceptual source is produced by two adjacent loudspeakers, or by a single loudspeaker if the panning is adjusted in such a way that the angular direction of the perceptual source is coincident with that of the loudspeaker, becoming rather a physical sound source, which enables the best possible directional quality of amplitude panning or its minimum directional dispersion (Pulkki, 1999, p. W99-3 ; Pulkki, 2001b, p. 13) (see also sections 2.2.1 and 2.2.2). For that reason, and because the localization of a perceptual sound source seems to be increasingly controlled by the most frontal loudspeaker as the loudspeaker pair becomes more lateral relatively to the human head-related reference system, so that the perceived position of the perceptual source is no longer in the middle between both loudspeakers when their levels are equal, as Gunther Theile and Georg Plenge found out in 1977 (Pulkki, 2001b, p. 25; Frank, 2013, p. 25), the use of loudspeakers at lateral directions is highly recommended (Pulkki, 2001b, p. 28). A quadraphonic system (see section 2.3.4.5), for example, does not produce stable perceptual sound sources in lateral directions, because there are no loudspeakers in those positions.

According to Pulkki (1999, p. W99-2), when the perceptual sound source is between two loudspeakers, its directional dispersion or spatial width, and the corresponding localization blur (see section 2.5) increase when compared to panning on loudspeakers, that is, relatively to the situation when the angular direction of the perceptual source is coincident with that of a loudspeaker (cf. section 2.5). In addition, its perceived localization can be incorrect, and it can suffer changes in timbre. Furthermore, the directional dispersion varies with the direction of panning (Pulkki, 2001b, p. 13) (see section 2.3.4.3), and the localization accuracy inside and outside the best listening position, or sweet-spot as it is also called, is better if the angle between adjacent loudspeakers is as small as possible (Pulkki, 2001b, pp. 28–29). Thus, according to Pulkki (2001b, p. 29), the quality of perceptual sound sources is relatively good for a large listening area in an eight loudspeaker sound reproduction system where the angle between adjacent loudspeakers is of 45º.

In this respect, Frank (2013, pp. 81–82) argues that the Vector-Based Amplitude Panning (VBAP) system (see section 2.3.4.8) produces the greatest timbre changes between panning on and panning between loudspeakers in an eight or sixteen horizontal circular equidistant loudspeaker configuration, relatively to other systems, such as Multiple Direction Amplitude Panning (MDAP) (see section 2.3.4.9) and Ambisonics max r_{E} weighting (see section 2.3.4.7), because in this configuration VBAP alternately uses 1 or 2 loudspeakers. In VBAP, this effect and the changes of a perceptual sound source's spatial width will be most perceptible if the

perceptual sound source is set in motion around a listener. In contrast, in Ambisonics and in other matrixing sound reproduction systems, since the same signal is present in all loudspeakers, although with different amplitudes, which can cause a perceptual sound source to become worse in quality, the spatial width of the perceptual sound source is almost kept constant, and the number of loudspeakers responsible for its perception is not dependent on panning direction (Pulkki, 1999, p. W99-1) (cf. section 2.3.4.3). When only the horizontal plane is used in Ambisonics, the system can also be designated as a pantophonic system (Fellgett, 1974, p. 537; Streicher & Everest, 1998, p. 13.15, quoting Fellgett (1974)) (see also section 2.3.4.7).

In the situation where triplets are used, as can be the case in VBAP (see section 2.3.4.8), not all loudspeakers are in the same plane as the listener (three-dimensional configuration), and the perceptual source is always produced by maximally three adjacent loudspeakers at a time, which form a triangle set, even when the number of channels and loudspeakers is greater than three (Pulkki, 1999, p. W99-3). Such a system which involves loudspeakers placed at different heights and depths in order to give the impression that sound can be produced from all directions, suggesting a full-sphere reproduction, is nowadays also designated as a periphonic system (Castellanos, 2006, p. 20), although this term was initially used by Fellgett (1974, p. 537) in relation to Ambisonics (see section 2.3.4.7).

In any case, the directional quality generally degrades more outside the best listening position, because the perceptual source is then located closer to the loudspeaker that produces it due to the precedence effect (Pulkki, 2001b, p. 28) (see section 2.5.5). For that reason, in most sound reproduction systems the best listening position is at their centre (Marentakis et al., 2008, p. ICAD08-1). However, investigations have been made to analyse the effects and improve the systems outside the best listening position, such as in Stitt et al. (2013), where it was found out that a higher order in Ambisonics increased the size of the listening area, including off-centre listening positions.

With respect to time panning, Pulkki (2001b, p. 15) refers that it is a technique in which a constant delay between 0 and a maximum of about 1 millisecond is applied to one of two loudspeakers in a two-channel stereophonic system, so that the perceived virtual or perceptual sound source moves or appears closer to the loudspeaker which produces the earliest sound signal. Furthermore, Pulkki states that this technique does not produce steady perceptual sources, because the perceived direction of perceptual sources is dependent on frequency, and that this technique is, for this reason, rarely used in positioning sound sources within a stereophonic image, except in creating special effects.

A hybrid technique of time and amplitude panning is not recommended either, because unstable perceptual sources are produced as well when too few loudspeakers are used and the perceptual sources are far away, and when the listener is outside the best listening area, due to the precedence effect (Pulkki, 2001b, p. 15).

Lastly, spatialization techniques involving no panning at all are also possible (Thigpen, 2009). As already mentioned in section 2.2.3, each sound is associated with its own loudspeaker, which is then used and considered as a point-source instrument, the localization of which is more accurate than that of a sound dependent on any panning method. Furthermore, the localization is independent of the listening position, the loudspeakers do not need to be placed at equal distances and symmetrically around the listener, and they do not need to be from the same size, the same quality or type. If the illusion of a moving sound object is to be created, then this can be more or less accomplished by using rapid successions of similar sounds in different loudspeakers, as

already described in section 2.2.4.

Odowichuk (2012, p. 28) mentions that the dominant commercial sound reproduction systems are the 5.1 Surround and 8.1 Surround systems (see section 2.3.4.6), which shows that nowadays two channels and a stereo pair are no longer sufficient. In this sense, Grigoriou & Floros (2010, pp. 430, 432) refer that 3D reproduction is traditionally made with the use of multiple channels, such as in 5.1 Surround systems or in Wave Field Synthesis (see section 2.3.4.11), and that modelling methods, such as binaural processing and Ambisonics (see section 2.3.4.7), have been developed over the past few years in order to recreate the sound space as accurately as possible in three dimensions.

Being initially considered a luxury and only available to a restricted social class of occidental society, the success of portable music during the 20th century increased the binaural production and simulation of spatial characteristics through HRTFs until today (Streicher & Everest, 1998, pp. 1.2–1.3; Zelli, 2009). As a consequence, a new form of music composition evolved: the headphone composition. On the other hand, from the moment when the human performer was replaced by loudspeakers, composition began to deal increasingly with the placement of loudspeakers and with spatial location of the sound sources (Zvonar, 2005a).

2.3.4.1 Mono

As already mentioned in section 2.3.4, *mono* means single. Thus, the mono reproduction system uses only one single channel to transmit a signal (which can be a mixture of many other signals) from a source to a listener, as is the case in monaural and monophonic systems (cf. section 2.3.4). Furthermore, if such a signal is reproduced over one or more loudspeakers, such as in single channel clusters, where two or more loudspeakers are close together, in mono split clusters, or in mono distributed loudspeaker systems, fed by this unique channel, then a loss of the spatial complexity of the original sounds is observed (Blauert, 1997, p. 248). When only one loudspeaker is used, then a well defined auditory event or perceptual sound source is usually perceived by the human auditory system, being independent of the best listening position (M. F. Davis, 2014, p. 812). When more than one loudspeaker is used, linear distortions of the single input signal can be introduced by comb-filtering effects, which can result from the summation of two coherent or nearly coherent monophonic output sounds (see section 2.1.1) combined out of time (Brunner et al., 2007, p. 1; McCarthy, 2007, pp. 78–79).

It should be recalled here that the ear input signals are only coherent if the source is placed in the median plane (cf. sections 2.1.1 and 2.3.2). If a sound source is displaced laterally, the degree of coherence of the ear input signals can decrease more than 50%, while the localization blur and the perceptual sound source width increase simultaneously (Blauert, 1997, p. 250) (see sections 2.3.4 and 2.5).

2.3.4.2 Pseudo-Stereophony

A pseudo-stereophonic system is achieved by deriving two partially coherent signals (see section 2.1.1) from an original one and by reproducing them over two loudspeakers (Blauert, 1997, p. 248). Due to its characteristics, the spatial distribution of the reproduced auditory events or perceptual sound sources is independent of that of the original sound. Blauert (1997, pp. 248–250) and Streicher & Everest (1998, pp. 11.1–11.7)

present some versions of pseudo-stereophonic systems which were successively proposed: a) by Wilhelm H. Janovsky (in 1948) – pseudo-stereo by frequency response differences, so that low frequencies are reproduced for the most part by one loudspeaker (low-pass filter applied to the mono signal) and high frequencies are largely reproduced by the other loudspeaker (high-pass filter applied to the mono signal); b) by Holger Lauridsen (in 1954) – pseudo-stereo by two complementary comb filters, resulting from the sum of the direct mono signal with its delayed signal sent to one loudspeaker and from the difference between them to another loudspeaker; c) by Holger Lauridsen (in 1954), G. R. Schodder (in 1956), Holger Lauridsen and Franz Schlegel (in 1956) – pseudo-stereo by delaying, attenuating and recombining the original mono signal in phase at one loudspeaker and out-of-phase at the other (see section 2.1.3.1); d) by F. Enkl (in 1958) – pseudo-stereo by using two filters feeding a single loudspeaker each, whose characteristics are continuously altered by changes in the original mono signal; e) by Manfred R. Schroeder (in 1958), J. P. A. Lochner and W. de V. Keet (in 1960) – pseudo-stereo by using a loudspeaker which produces a monophonic signal inside a reverberation chamber (see section 2.3.2), picked up by two microphones, whose increasing spacing raises the degree of incoherence, each one feeding a single loudspeaker outside the chamber, or pseudo-stereo by using a reverberation plate or spring; and f) by Manfred R. Schroeder (in 1961) – pseudo-stereo by phase shifting, instead of delay applied by Lauridsen in 1954, using all-pass electronic circuits, which let all frequencies pass through and that only modify the phase relationship between various frequencies (Ballou, 1987, p. 585).

2.3.4.3 Stereophony

Recalling what has already been said before in section 2.3.4, the Greek word *stereo* means solid (Malham, 1998, p. 173) and it is related to the three spatial dimensions of width, depth, and height. Therefore, the most elementary stereophonic reproduction system uses two individual channels feeding two ideally 60º separate loudspeakers pointing symmetrically to, and placed in the horizontal plane with an elevation of 0º in front of, a listener (Pulkki & Karjalainen, 2001, p. 740; Pulkki, 2001b, p. 11), whose signals are partially or completely incoherent (see section 2.1.1), in order to reproduce the spatial dimensions and characteristics of a performance as well as possible (Blauert, 1997, p. 250) (see figure 2.34).

Nevertheless, this system seems to cover only a limited perceptual space (Meares, 1973, p. 7), whose boundaries are usually defined as follows if the listener is in the best listening position and if he or she is capable of imagining it, which mostly depends on his or her outer ears shape (which can cause phase cancellations - see section 2.1.3.1) or that depends on his or her consciousness of it: left and right boundaries are approximately defined by the positions of the loudspeakers themselves, depending on the place where they are installed and on the listener's imagination skills; back and front boundaries depend on the size of the loudspeakers and on previous auditory experience of the listener, so that perceptual sound sources normally seem to be maximal only at a short distance behind or in front of the loudspeakers; finally, the top and bottom boundaries seem to be as high as the top of the loudspeakers and as low as the floor (Gibson, 1997, pp. 8–14). The localization accuracy in this perceptual space becomes worse if the angle between both loudspeakers is greater than 60º (Pulkki & Karjalainen, 2001, p. 740).

Thus, amplitude differences or time differences can be used to produce a stereo image (Malham, 1998, p. 173). The former case plays an important role in the so-called intensity stereo recording techniques, mostly

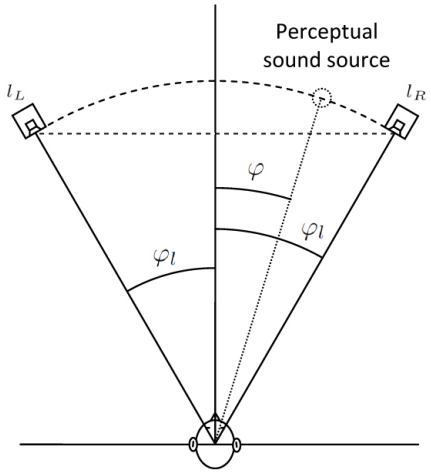


Figure 2.34: Most elementary stereophonic reproduction system.

intended for loudspeaker reproduction, where usually pairs of coincident microphones are used, that is, where microphones are placed as close together as possible in order to minimize arrival time differences of sounds (Streicher & Everest, 1998, p. 7.1). Amplitude differences also play a significant part in the already presented amplitude panning technique (section 2.3.4) and in the panning laws described in the next paragraphs. Time differences in turn are mostly found in widely spaced microphone and in near-coincident microphone stereo recording techniques, the latter based essentially on the natural spacing between the two human ears, of which the binaural pick-up system involving artificial heads makes part. Nevertheless, time differences are rarely used in synthetically positioning sound sources within a stereophonic image, except in creating special effects, as already mentioned in section 2.3.4.

Some panning laws have been developed for stereophonic reproduction systems, such as the sine law and the tangent law, which estimate the perceived horizontal angle or azimuth φ (see section 2.3.2) of a perceptual sound source between the two loudspeakers from their gain factors (Pulkki & Karjalainen, 2001, pp. 740–741; Pulkki, 2001a, p. 755; Pulkki, 2001b, pp. 11–13). Furthermore, the estimated direction is also known as panning direction or panning angle.

The stereophonic sine law was proposed by Bauer (1961), assuming that the listener's head is pointing forward, that the elevation of the loudspeakers relatively to the ears is 0°, and that the path followed by sound from the loudspeakers to the two ears is carried out in a straight line, so that inter-aural time differences exist (see section 2.5.4.1.1), but without considering the presence and the shape of the head:

$$\frac{\sin\varphi_e}{\sin\varphi_l} = \frac{g_{l_L} - g_{l_R}}{g_{l_L} + g_{l_R}} \Leftrightarrow \sin\varphi_e = \frac{g_{l_L} - g_{l_R}}{g_{l_L} + g_{l_R}} \times \sin\varphi_l \ ,$$

where φ_e is the estimated panning angle of the perceived azimuth φ of a perceptual sound source, φ_l is the angle between the main centre axis of the loudspeakers (see section 2.2.1) and the median plane (see section 2.3.2), and g_{l_L} and g_{l_R} are the gains of the signals applied to the left l_L and right l_R loudspeakers, respectively (Bennett et al., 1985, p. 315) (see figure 2.34).

In addition, the sine law is valid only at low frequencies (Pulkki & Karjalainen, 2001, p. 741), below about

700 Hz (Malham, 1998, p. 173), 600 Hz (Pulkki, 1999, p. W99-2), or 500 Hz (Pulkki, 2001b, p. 12). Above 700 Hz the panning angle of the perceived horizontal angle or azimuth φ of a perceptual sound source between the two loudspeakers increases, although in this situation it can be virtually corrected if $(g_{l_L} - g_{l_R})$ is multiplied by 0.7 (Malham, 1998, p. 174).

The tangent law, in turn, was proposed by Bennett et al. (1985), considering the presence and the shape of the head, so that the path of sound from the left (right) loudspeaker to the right (left) ear is carried out taking into account the curvature of the head (Pulkki, 2001b, p. 12). This law is more precise than the sine law if the listener is facing towards the perceptual sound source, although limitations are identical to those of the sine law (Pulkki, 1999, p. W99-2):

$$\frac{\tan\varphi_e}{\tan\varphi_l} = \frac{g_{l_L} - g_{l_R}}{g_{l_L} + g_{l_R}} \Leftrightarrow \tan\varphi_e = \frac{g_{l_L} - g_{l_R}}{g_{l_L} + g_{l_R}} \times \tan\varphi_l \ ,$$

where φ_e is the estimated panning angle of the perceived azimuth φ of a perceptual sound source, φ_l is the angle between the main centre axis of the loudspeakers and the median plane, and g_{l_L} and g_{l_R} are the gains of the signals applied to the left l_L and right l_R loudspeakers, respectively (Bennett et al., 1985, p. 316) (see figure 2.34). This estimated panning angle matches exactly the direction that the Inter-aural Time Differences (ITD) (see section 2.5.4.1.1) of frequencies below 1.5 kHz suggest (Pulkki, 2001b, p. 36) (see section 2.5).

Thus, in an elementary stereophonic reproduction system, the panning angle matches the angle of the perceived perceptual sound source at most frequencies in a fairly consistent way (Pulkki & Karjalainen, 2001, p. 739). Nevertheless, according to Malham (1998, p. 174), to make it happen loudspeakers should not be more than 60º apart from each other, because then perceptual sound sources, which are already inherently less stable at smaller angles between loudspeakers, get even more unstable. In addition, the generation of perceptual sound sources and the localization capability get worse as the listener moves away from the best listening position or as the listener rotates the head until it is parallel to the loudspeakers.

In the early 1930's, a three-channel stereophonic reproduction system with left, centre, and right positioned loudspeakers was presented by Bell Labs engineers in the United States of America as a practical solution to represent a much more complex system consisting of a theoretically infinite number of front loudspeaker channels (Holman, 2000, p. 12). This system, also called 3-0 stereo, is still used nowadays as the main set-up in motion picture sound productions where dialogues are usually reproduced by the centre loudspeaker, which is a physical sound source that provides a stable image for every listener independently of the seating location (Streicher & Everest, 1998, p. 13.2).

In the 1940's and 1950's, rear channel loudspeakers were added first in cinemas to the main three-channel stereophonic set-up in order to be used for ambience and/or sound effects. A fourth monophonic channel was therefore introduced to feed several loudspeakers placed to the sides, or to the sides and at the back of the audience (Streicher & Everest, 1998, p. 13.4). This system has become to be called 3-1 stereo and was not specifically intended to enable 360º sound source localization. A variety of loudspeaker configurations with two or more rear channels based on this idea of ambience are known (see next sections).

Nevertheless, whereas multichannel recording was too expensive for the motion picture industry, the three-channel stereophonic reproduction system also proved to be quite expensive and impractical for the consumer, and so the elementary two-channel stereophonic system prevailed commercially from the 1950's to the 1970's

(Streicher & Everest, 1998, p. 13.2; Holman, 2000, p. 14).

2.3.4.4 Pseudo-Quadraphony

The pseudo-quadraphonic reproduction system was a relatively simple and not too expensive attempt to satisfy the demand for a home system that would transmit the sense of ambience coming from behind a listener as in an actual live performance, an aim that the stereophonic reproduction system cannot fulfil (see section 2.3.4.3).

Thus, a pseudo-quadraphonic system is achieved by simulating a four-channel system from two-channel stereophonic signals. It is also known as a 2-2-4 system, because the signals are originally picked up by two microphones, recorded and/or transmitted over two channels, and reproduced through four loudspeakers (Meares, 1973, p. 7). The ambient information that the difference between the two stereophonic signals holds is transmitted by two 180 degrees out-of-phase rear loudspeaker signals (see section 2.1.3.1) in order to increase the sense of ambience, where the rear left and the rear right loudspeakers are placed behind and slightly to the sides of the listener, and respectively fed with a left-channel minus right-channel signal, and a right-channel minus left-channel signal (Streicher & Everest, 1998, p. 13.3).

Thus, the two rear loudspeaker signals are liable to a smaller bandwidth as ambient sounds also tend to have less high frequency contents due to air absorption at higher frequencies (Meares, 1973, p. 8).

In this system, when a perceptual sound source moves towards left or right, it becomes diffuse and it moves too far to that side. A delay can be introduced in the rear signals in order to enhance the front signals and minimize this effect, but if a perceptual sound source is to be clearly located, then the pseudo-quadraphonic system is certainly not the recommended one (Meares, 1973, pp. 7–8).

2.3.4.5 Quadraphony

Seeking for a full surround home system, preferably compatible with, and playable on, existing two-channel stereo systems, several methods of encoding the surround contents, such as quadraphonic matrix and discrete systems, were developed in the 1970's. According to Streicher & Everest (1998, p. 13.6), the American Peter Scheiber (born in 1935)[42] was one of those responsible for the invention of matrix encoding. In any case, the loudspeakers, preferably of the same type, quality, and size, are placed evenly around the listener in the horizontal plane as shown in figure 2.35.

Thus, in quadraphonic matrix systems, also designated as 4-2-4 systems, four signals are encoded, matrixed or reduced into two signals so as to be recorded and/or transmitted via two-channel standard stereophonic devices, such as the phonograph disc (vinyl record) or the cassette tape, most probably already existing at the consumers' homes in the 1970's. Afterwards, the two signals are decoded, dematrixed or expanded back, by an extra matrix decoding device, to four channels and reproduced through four loudspeakers (Streicher & Everest, 1998, p. 13.5). However, as a result, the four signals are not, at the end of the process, identical to the original ones any more, that is, the complete separation of the four original signals is lost in this process, and in some cases the perceptual sound sources are not perceived at the original positions (Meares, 1973, p. 9).

[42] Retrieved 27/04/2016, from `https://en.wikipedia.org/wiki/Peter_Scheiber`

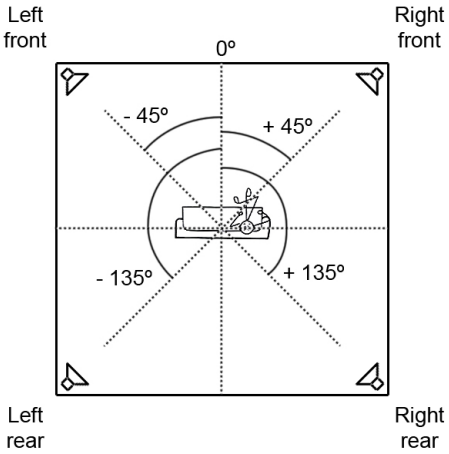


Figure 2.35: Quadraphonic reproduction system (adapted from Streicher & Everest, 1998, p. 13.5).

In a quadraphonic discrete reproduction system, also known as a 4-4-4 system, compatibility with stereo is avoided, and so four signals are recorded and/or transmitted over four channels, and reproduced through four loudspeakers (Meares, 1973, p. 8).

Nevertheless, in 1970 the British audio pioneer and mathematician Michael Anthony Gerzon (December 4, 1945 - May 6, 1996)[43] argues that satisfactory quadraphonic sound can be transmitted via three channels, although there is some loss of directional information, and that it has several advantages over four-channel quadraphony, such as a better recording performance on tape and the use of a half as expensive multitrack heads system (Gerzon, 1970).

Although in quadraphony amplitude panning methods can be used where pairs of loudspeakers are fed at a time (Pulkki, 2001b, p. 13), a quadraphonic system does not produce stable perceptual sound sources in lateral directions, because there are no loudspeakers in that positions (Pulkki, 2001b, p. 29) (see figure 2.35), as was already mentioned in section 2.3.4. Furthermore, the best listening position is that in which the listener is equidistant from all loudspeakers.

According to Streicher & Everest (1998, p. 13.7) and to Holman (2000, p. 14), the commercial failure of quadraphonic reproduction systems in the 1970's was due to technical problems, but also because of persistence and stubbornness of the audio industry in wanting to keep the main three incompatible with each other competing commercial matrix systems, which led the consumers to give them up: 1) SQ Matrix, developed by the North American Columbia Broadcasting System (CBS) Records, the Japanese Sony Corporation, and others; 2) QS Matrix, developed by the North American Warner Record Group, the Japanese Sansui, and others; and 3) CD-4 (Compatible Discrete 4 channel sound for phonograph vinyl records), developed by the Japan Victor Company (JVC).

2.3.4.6 Surround 5.1

After several years of technical improvements in cinema audio and cinema reproduction systems, in 1987, with the advent of digital audio, the 5.1-channel system, whose nomenclature was proposed by Holman (2000, p. 66), although the rigorous international standard nomenclature should be 3-2-1 (Rumsey, 2001, p. 88–89), was

[43] Retrieved 26/04/2016, from `https://en.wikipedia.org/wiki/Michael_Gerzon`

set as a standard both for cinema and for home applications by a subcommittee, called Digital Sound on Film, of the international standards organization, also known as Society of Motion Picture and Television Engineers (SMPTE). This system uses normally six independent signal channels, three of which are for frontal left (-30º), centre (0º) and right (+30º) loudspeakers, two for two rear/side left (-110º$\pm$10º) and right (+110º$\pm$10º) surround loudspeakers, all five at a seated ears height of 1.2 meters, and one dedicated low frequency effects (LFE) channel for a sub-woofer, which can be placed anywhere, where its frequency response is best at the listening position, as shown in figure 2.36. It is usually recommended that the surround loudspeakers are of the same type, quality, and size as the three front loudspeakers. If there are impediments in placing the five main loudspeakers at the same distance from the best listening position at the centre, then time delays can be inserted in the channels whose loudspeakers are closer to the centre, in order to synchronize their signals. The same applies to the sub-woofer. In addition, if it is not possible to place the two surround loudspeakers where they need to be placed, they can be exceptionally elevated up to 30º above the ears height, because our ears are approximately three times less sensitive to errors in elevation than to errors in the horizontal plane (Holman, 2000, pp. 44, 46).

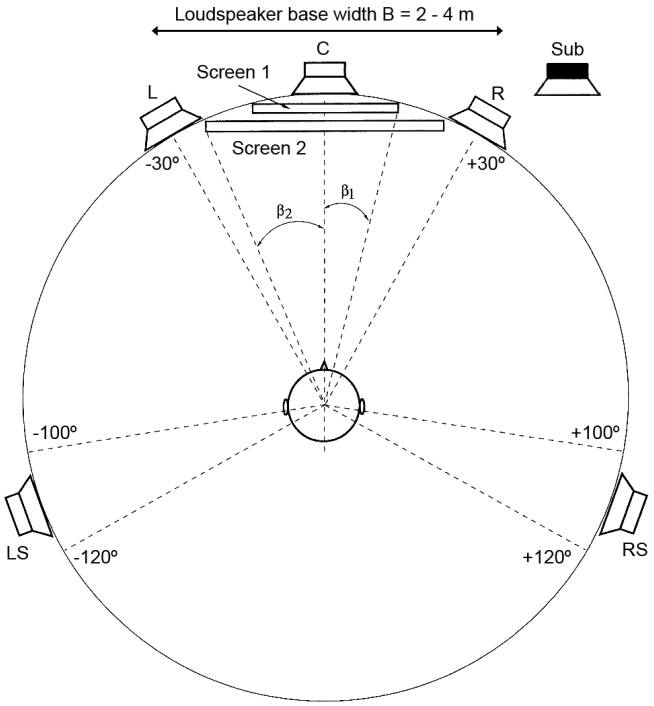


Figure 2.36: International Telecommunication Union (2012) ITU-R BS.775 5.1 reproduction system (adapted from Rumsey, 2001, p. 88, and Holman, 2000, p. 45).

This system was thought to be fundamentally a front three-channel stereophonic reproduction system with a rear/side two-channel system for stereo ambience, in order to satisfy, more convincingly than the previously presented systems, the demand for the sense of ambience coming from behind a listener, just like in an actual live performance, in which the sound is primarily coming from the front. This is the reason why it is also often called 3-2 stereo rather than 5-channel surround (Rumsey, 2001, p. 87). Just like the 3-1 system, 5.1

was not specifically intended to enable 360º sound source localization. In 5.1 surround, amplitude panning methods can also be used where pairs of loudspeakers are fed at a time (Pulkki, 2001b, p. 13), although conventional panning laws are not optimized for three loudspeakers. Furthermore, amplitude panning does not work effectively on the sides, because some parts of the spectrum are emphasized from the front and others from behind due to the different Head-Related Transfer Functions (HRTFs) (see section 2.4.1), which lead to a perception of two different spectrally perceptual sound sources (Holman, 2000, pp. 123, 209–210). However, whereas the rear/side positions of the loudspeakers in 5.1 allow for a better imaging for the lateral directions than the quadraphonic configuration, the localization of perceptual sound sources behind the listener is still unstable, because the angle between the rear/side loudspeakers is too large (Pulkki, 2001b, p. 29).

In cinema, the audio information of the LFE channel requires usually 10 dB more headroom or signal-handling capabilities than that of the five main channels, because this channel is thought specifically for sound effects in the lower range of the audio spectrum, such as explosions, which the main loudspeakers cannot reproduce accordingly. Thus, the LFE channel is, in this case, directly connected to the sub-woofer. In this lower range, the ear is less sensitive to absolute level, but more sensitive to changes in level than in the midrange, because the equal loudness contours are closer together (Holman, 2000, pp. 62, 67–68) (see figure 2.12 in section 2.1.3.2).

The LFE channel is also known as the 0.1 channel, although it should be called 0.005, because its analogue bandwidth is limited up to 120 Hz. In a digital system, according to the Shannon or Nyquist sampling theorem, an analogue continuous signal of finite bandwidth can be represented by a discrete sequence of samples, if it does not contain frequency components above one-half of the sampling frequency or sampling rate, which in turn corresponds to a certain number of samples per second (Smith, 1997, p. 40). Thus, the bandwidth of 120 Hz can be properly sampled with a 240 Hz sampling rate, which corresponds to 1/200 = 0.005 of the standard sampling rate of 48000 Hz of the main channels (Holman, 2000, p. 66):

$$240\ \mathrm{Hz} = \frac{48000\ \mathrm{Hz}}{200} = 48000\ \mathrm{Hz} \times 0.005\ .$$

In turn, according to the International Telecommunication Union (2012, p. 22) recommendation, a bass management system is used in home 5.1-channel surround systems, which opens up the possibility of combination or isolation of the signals which feed the loudspeakers, depending on the use of a sub-woofer or not. Nevertheless, it is suggested that the five main channels convey the full range contents and that the LFE channel carries only the extreme low frequency effects or enhancement information, as in cinema (International Telecommunication Union, 2012, p. 12). Thus, the integrity of the programme which is transmitted to the listener is not at risk if the LFE channel is not reproduced (Rumsey, 2001, p. 91). In this respect, if the sub-woofer is not used, the LFE signal can be added to the five main channels, and the combination can then be sent to the five main loudspeakers (International Telecommunication Union, 2012, p. 22).

In many home bass managed systems with limited bass response, the signals of the five main channels are high-pass filtered and sent to their respective loudspeakers, and the sub-woofer reproduces the low frequencies from the main five low-pass filtered channels summed with the LFE information (International Telecommunication Union, 2012, p. 22; Holman, 2000, pp. 43, 69–70), because most main loudspeakers used in home systems cannot reproduce the lowest frequencies from about 40 Hz down to approximately 20 Hz. The sep-

aration of the low and high frequency signals is carried out between 80 Hz and 160 Hz by crossover systems (Rumsey, 2001, p. 92). Furthermore, the difficult localization of low-frequency sounds in general and their practically insignificant perceivable stereophonic effect are also psychoacoustical reasons why this system can use a common bass sub-woofer, which has contributed to the spread of multichannel sound (Holman, 2000, pp. 52–53, 70, 207). However, Rumsey (2001, p. 92) stresses the fact that low frequency spaciousness turns to be limited if it is confined to a monophonic channel.

Many set-up variations of 5.1 have been developed where more surround channels and loudspeakers are used, such as 6.1, 7.1, 9.1, 10.1, 10.2 (Birkner, 2004, pp. 37–41), 11.1 and 22.2, whose descriptions are beyond the scope of this thesis.

2.3.4.7 Ambisonics

The sound reproduction systems described so far involve primarily loudspeakers placed in the same plane with a listener, that is, commonly in the horizontal plane (see section 2.3.2). Looking for better solutions for recording and reproducing the entire three-dimensional sound field, so that perceptual sound sources could be localized, Michael Gerzon (already mentioned in section 2.3.4.5), the British physicist Peter Fellgett (April 11, 1922 - November 15, 2008)[44], Peter Graham Craven, Geoffrey James Barton, and, in an independent manner, the American physicist Duane H. Cooper (1923 - April 4, 1995)[45], and Takeo Shiga (Malham, 1998, p. 175), of the Nippon Columbia Company, Ltd., developed a system in the 1970's, called Ambisonics because it is capable of reproducing ambience (Fellgett, 1974, p. 535), whose basic principle is to capture the actual acoustical signals of a performance in the best feasible way, so that the original sound field can afterwards be recreated as accurately as possible (Daniel, 2001, p. 101; Grigoriou & Floros, 2010, p. 431; Stitt et al., 2013, p. DAFX-1; Power et al., 2013, p. 1). In order to do so, the sound field at a specific point in space can be recreated only if the absolute sound pressure, referred to as *W*, and the three pressure gradients associated with the three orthogonal *x* (fore/aft), *y* (left/right), and *z* (vertical) reference axes, designated respectively as *X*, *Y*, and *Z*, at that point are defined and maintained (Streicher & Everest, 1998, p. 13.11; Daniel, 2001, p. 101). A pressure gradient describes the difference in pressure around a particular location in space, in this case in one of these reference axes (Streicher & Everest, 1998, p. 7.3).

Thus, theoretically, a microphone with an omnidirectional polar pattern can be used to capture the absolute sound pressure at that point in space, because it ideally responds equally and evenly to sound arriving from all directions at any frequency, although in practice the response and consequently the shape of the pattern usually change somehow with the angle at higher frequencies.

Three microphones with a bidirectional or figure-of-eight polar pattern can also be used to capture the pressure gradients on the three orthogonal axes at that point, since each responds to pressure differences on either side of its diaphragm with equal magnitude, but of opposite polarity (Streicher & Everest, 1998, p. 7.3; Eargle, 2001, pp. 61–63). In practice, however, it is not possible to position these four microphones at that point at the same time. As a solution, a special microphone system was developed, consisting of a sound field microphone and of processing of captured signals via dedicated hardware (Streicher & Everest, 1998, pp.

[44] Retrieved 13/05/2016, from `https://en.wikipedia.org/wiki/Peter_Fellgett`
[45] Retrieved 16/05/2016, from `https://en.wikipedia.org/wiki/Duane_H._Cooper`

13.12–13.13), or nowadays also via software installed on a computer (Malham, 1998, p. 175).

Therefore, the head of the sound field microphone is composed of four closely spaced sub-cardioid or cardioid polar pattern input transducers (cf. section 2.2.1), placed at the face centres of a regular tetrahedron, that is, a regular pyramid with four equilateral triangle faces (see figure 2.37).

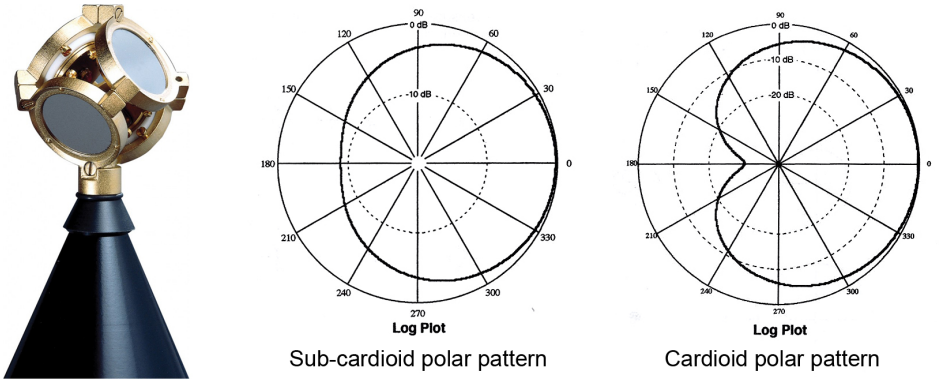


Figure 2.37: To the left: head of a sound field microphone, firstly developed by Calrec Audio Limited (retrieved 09/05/2016, from `http://www.musictech.net/2014/01/10mm-soundfield-microphones`). At the centre and to the right: sub-cardioid and cardioid polar patterns (adapted from Eargle, 2001, p. 84).

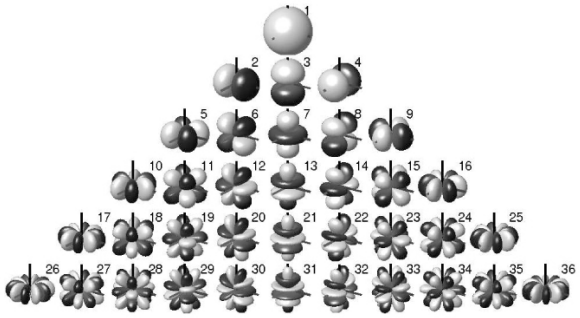
The signals of these four input transducers, known as A-format signals, should represent four equal regions of the sound field, but they do not, because the four input transducers do not coincide at a point in space. Thus, the exact representation of the absolute sound pressure *W*, and of the three pressure gradients *X*, *Y*, and *Z*, at one point in space is only possible by processing or encoding these four A-format signals into the so-called four B-format signals, which define the sound field at that point (Streicher & Everest, 1998, p. 13.13; Malham, 1998, p. 175; Daniel, 2001, pp. 102–103). The encoding formulas are derived from the solution of the three-dimensional wave equation, which describes sound waves mathematically, in spherical coordinates, where a point in space is represented by a radius *r*, azimuth angle φ, and elevation angle δ (Neukom & Schacher, 2008) (see also section 2.3.2). The four discrete B-format signals can therefore be recorded on multi-track systems, transmitted, or further processed into two-channel stereo compatible C-format signals, decoded and then reproduced over a wide variety of loudspeaker configurations as D-format signals, which can be derived from B- or C-format signals and adjusted depending on the selected loudspeaker configuration (Rumsey, 2001, pp. 111–115).

In Ambisonics, it is possible to re-create the horizontal sound field only, known as pantophonic reproduction, or to re-create the whole sound field including the vertical information as well, known as periphonic reproduction (see also section 2.3.4). Whereas pantophonic reproduction requires at least three loudspeakers to surround a listener in the horizontal plane (see section 2.3.2), in periphonic reproduction a minimum of four is needed. Nevertheless, a minimum of four and six loudspeakers is respectively recommended in practice (Fellgett, 1974, p. 537; Zelli, 2009).

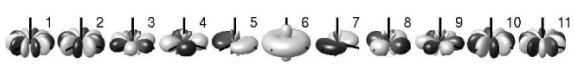
Several authors, such as Fellgett (1974, p. 536), Malham (1998, p. 175), Daniel (2001, p. 101), Birkner (2004, p. 32), and Grigoriou & Floros (2010, p. 431), designate *W*, *X*, *Y*, and *Z* as spherical harmonic components. Since non-redundant combinations of higher gradients can exist, these gradients can be represented by higher-order spherical harmonics. In theory, the sound field can be only recreated with an infinite number of

spherical harmonics. In practice, however, the necessary total finite number of independent signals, each representing a spherical harmonic, and the corresponding number of loudspeakers in periphonic reproduction is equal to $(n+1)^2$, where *n* indicates the finite integer order of the spherical harmonic, starting at zero. This leads to what can be designated as 0th-order, 1st-order, 2nd-order, nth-order Ambisonics. In turn, in pantophonic reproduction, the necessary total number of independent signals and respective number of loudspeakers is equal to $2n + 1$. In the latter case, circular harmonics, consisting of sine and cosine functions (Neukom & Schacher, 2008) which are part of the Fourier series (see sections 2.1.3.1 and 2.1.4), are used rather than spherical harmonics (Monro, 2000, p. 1). With higher orders, the spatial resolution can be improved (Power et al., 2013, p. 2), and the listening area can be increased, so that it can become larger than that of a normal quadraphonic system (Birkner, 2004, p. 32) (see section 2.3.4.5), but obviously more loudspeakers are needed (Stitt et al., 2013, p. DAFX-1; Power et al., 2013, p. 2). However, Ambisonics requires that all loudspeakers are of the same quality.

Thus, for 0th-order Ambisonics ($n = 0$), the only spherical harmonic of order zero is the absolute sound pressure *W*, which is reproduced as a monophonic sound without directional information (Fellgett, 1974, p. 536; Neukom & Schacher, 2008; Power et al., 2013, p. 2) (identified as 1 in figure 2.38(a)). Whereas for 1st-order Ambisonics ($n = 1$), in periphonic reproduction, a combination of one spherical harmonic of order zero (*W*) and three spherical harmonics of order one (*X*, *Y*, and *Z*) is used, respectively identified as 1, 4, 2, and 3 in figure 2.38(a), in pantophonic reproduction only three circular harmonics are reproduced (identified as 6, 7, and 5 in figure 2.38(b)). For 2nd-order Ambisonics ($n = 2$), signals are made up by spherical or circular harmonics of order zero, one, and two. This applies to other orders in a similar manner.



(a) N = 5 spherical harmonic basis



(b) N = 5 horizontal-only encoded

Figure 2.38: Spherical harmonics of order $n = 5$ in periphonic (a) and in pantophonic (b) reproduction (Nachbar et al., 2011, p. 6).

Although Ambisonics can be considered basically a microphoning technique (Pulkki, 2001b, p. 14), it can also be used to synthesize spatial audio, so that amplitude panning methods can be applied, in which a signal is sent to all loudspeakers, preferably equally placed around a listener (Grigoriou & Floros, 2010, pp. 431–432), but with different gain factors. In this sense, Power et al. (2013, p. 2) argue that this regular loudspeaker distribution is easily accomplished in the horizontal plane by spacing them equally, but that for elevated sources only the positions of the vertices of the Platonic solids (tetrahedron, hexahedron or cube, octahedron, dodeca-

hedron, and icosahedron - see Hollerweger, 2006, p. 69) are possible (Neukom & Schacher, 2008). Since the same signal is present in all loudspeakers (Pulkki, 2001b, p. 13), but with different amplitudes, this can cause a perceptual sound source to become worse in quality, although the spatial width of the perceptual sound source is almost kept constant (cf. section 2.3.4), which provides stable perceptual sound sources (Odowichuk, 2012, p. 28), and a sound field that remains the same if the listener rotates his head (Grigoriou & Floros, 2010, p. 432). Thus, the sum of the gains of all loudspeakers is equal to 1 (Neukom & Schacher, 2008), that is, 100 %. As the Ambisonics order increases, the absolute values of the gains are considerably lower on the opposite side of a panning direction (Pulkki, 2001b, p. 14), and higher sound source directionality is achieved (Malham, 1998, p. 176; Power et al., 2013, p. 3). In addition, higher order Ambisonics can also yield superior immersiveness and reproduction of moving sound sources (Thigpen, 2009; Penha & Oliveira, 2013, p. 660) (see section 2.2.4). In this sense, Zelli (2009) refers some musical examples that use Ambisonics, such as *Vox 1* (1982) by Trevor Wishart (to whom has already been made reference in section 2.3), *Pyrotechnics* (1996) by Ambrose Field, *What a Difference a Day Makes* (1997) by Tim Ward, and *Spherical Construction* (1997) by John Richard.

Power et al. (2013, p. 1) argue that it is possible to reduce the vertical order relatively to the horizontal order of Ambisonics for a particular localization resolution, because the human hearing system is less sensitive to changes in the vertical plane than in the horizontal plane. Furthermore, many studies support the idea that the vertical localization of the perceptual sound source depends on its frequency contents rather than on the actual physical sound source position, that is, if signals with frequencies above 7 kHz are produced by a physical sound source in front of a listener, then the perceptual sound source will be naturally localized at a vertically displaced position, described in terms of directional bands (Blauert, 1997, pp. 102, 112) (see sections 2.5 and 2.5.2).

Summing up, Ambisonics is a process which usually involves: 1) capturing the four A-format signals; 2) encoding these signals into B-format and recording them on a standard recording device, or transmitting them in C-format; 3) decoding the signals for reproduction purposes in D-format; and 4) reproducing the signals over loudspeakers placed in suitable positions (Fellgett, 1974, pp. 535–536; Streicher & Everest, 1998, p. 13.12).

Signals feeding loudspeakers that are opposite or far away from the original sound position, and signals which are out-of-phase relatively to others (see section 2.1.3.1), are errors which occur in ordinary or basic Ambisonics, when the theoretically infinite number of spherical or circular harmonics, respectively necessary for the description of a sound field in periphonic or in pantophonic reproduction, is limited in practice to a finite number by a truncation operation. However, these errors can be diminished by correcting or weighting gains according to the orders of spherical or circular harmonics, although the precision of directionality is affected to some extent, as is the case in the so-called in-phase correction, proposed by D. G. Malham in 1992 (Stitt et al., 2013, p. DAFX-1). In-phase correction was thought for circumstances where many listeners are not in the best listening position, so that all signals are set in phase, the gains of the most distant loudspeakers from the original sound position are maximally reduced, and the source position is therefore widened (Monro, 2000, p. 1; Neukom & Schacher, 2008; Frank et al., 2008). These principles are analogous to those in window design, which is used in finite impulse response (FIR) filter design, in signal processing (Smith, 1997, pp. 261–276, 285–296). Another possible way of correction is Ambisonics max r_{E} (maximum energy) weighting (see section

2.3.4), in which the energy of all loudspeakers is concentrated in the region of the reproduced sound source.

Neukom & Schacher (2008) present panning functions which are equivalent to the result of encoding and decoding basic and in-phase Ambisonics, and that can be used in applications, in which many perceptual moving sound sources have to be rendered in real-time, such as in Bisig et al. (2007) (cf. section 2.1.1), in order to simplify and speed up calculations. Whereas ambisonic encoding is only possible with integer orders, in Ambisonics Equivalent Panning (AEP), the exponent in the equivalent panning function, or order of the ambisonic resolution, can be any positive number, which opens up the possibility of mixing sounds with different orders. Furthermore, fewer loudspeakers are needed as this order increases, because the AEP function narrows increasingly with the order (Neukom & Schacher, 2008) (cf. appendix E.1).

Comparing the performance of 1st-, 3rd-, and 5th-order basic Ambisonics, max r_{E} weighting, and in-phase correction, in the localization of broadband pink noise (see section 2.1.4) providing many localization cues (see section 2.5), but with very limited head movement possibilities, in an almost circular loudspeaker configuration in the horizontal plane (see section 2.3.2), and therefore with and without an appropriate delay compensation for the loudspeaker positions, within a common reverberant room (broadband $\mathrm{RT}_{60} < 1\ \mathrm{s}$) (see section 2.3.2), Frank et al. (2008) found out that the 5th-order max r_{E} weighting without delay compensation is the best option, followed by basic Ambisonics, and lastly by in-phase correction, both for the central listening position and off-centre position, although localization is more accurate at the former position than at the latter. It should be noted here that the improvement of spatial resolution with higher orders is consistent with the fact that the localization performance is better with higher order Ambisonics, independently of the listening position.

Stitt et al. (2013, p. 1) use max r_{E} weighting without delay compensation in a 1st- and 3rd-order Ambisonics localization experiment involving broadband amplitude modulated white noise target and pointer sounds for non-centred seated listeners, with their heads kept as still as possible, comprised of a completely circular loudspeaker configuration in the horizontal plane, because it was shown by Frank et al. (2008), as mentioned above, that it performed better in localization for both the central listening position and, in this case, fundamentally the off-centre position. This study was intended to isolate the influence of sitting off-centre from that of the irregular loudspeaker configurations, and in a room with a very short reverberation time (broadband $\mathrm{T}_{30} < 0.095\ \mathrm{s}$, extrapolated as broadband $\mathrm{RT}_{60} = 2 \times \mathrm{T}_{30} < 0.19\ \mathrm{s}$). This experiment confirms that the listening area increases with the order of Ambisonics.

Subjective localization tests of broadband pink noise and speech signals, reproduced in a semi-anechoic chamber at different elevations and azimuths by 1st-, 2nd-, and 3rd-order max r_{E} weighting Ambisonics irregular loudspeaker configuration systems, are presented in Power et al. (2013). The perceived sound source positions were verbally reported by the listeners in spherical coordinates. The results show that the localization of elevated perceptual sound sources is more accurate in 3rd-order max r_{E} weighting, although higher orders are necessary to match the localization resolution of the human hearing system. In addition, statistically significant differences in localization accuracy occurred between the 1st- and 3rd-order systems for speech only, which means that the ideal order of Ambisonics in the reproduction of elevated sound sources depends on the desired localization accuracy and the frequency contents of the original sound signal.

2.3.4.8 Vector-Based Amplitude Panning (VBAP)

Vector-Based Amplitude Panning (VBAP) was developed by the Finnish researcher Ville Pulkki (Neukom & Schacher, 2008), and it uses amplitude panning methods (see section 2.3.4) in any two- or three-dimensional loudspeaker configurations, where pairs or triplets of loudspeakers are respectively and usually fed at a time in order to position perceptual sound sources (Pulkki, 1999, p. W99-3; Pulkki, 2001b, pp. i, 16, 31; Marentakis et al., 2008, p. ICAD08-2; Odowichuk, 2012, p. 28).

The tangent panning law is reformulated with vectors in two-dimensional VBAP, where pairs of loudspeakers are fed at a time (Pulkki, 1997, p. 457; Pulkki, 1999, p. W99-3; Pulkki, 2001b, pp. 16, 35; Pulkki & Karjalainen, 2001, p. 741) (see section 2.3.4.3). A *vector* is a quantity with direction and magnitude, which is represented by an arrow, indicating the direction of the quantity, with a length that is proportional to its magnitude (Tipler, 1999, p. 55). This new approach can be easily applied to situations where triplets of loudspeakers are fed at a time.

The direction of a perceptual sound source in three-dimensional VBAP is given by a Cartesian unit-length panning vector p, which is expressed as a linear combination of three unit-length vectors l_n, l_m, and l_k, each one defining the direction, starting from the listening position, of loudspeakers n, m, and k in a triplet, with the respective normalized loudspeaker gain factors g_n, g_m, and g_k, that is, with the adjustment of the gain values to a common scale (Pulkki, 1997, p. 460; Pulkki, 1999, p. W99-3; Pulkki, 2001b, pp. 16–17, 35):

$$p = g_n \times l_n + g_m \times l_m + g_k \times l_k \ .$$

Thus, the loudspeakers of a triplet, forming a triangle as small as possible in order to improve the quality of a perceptual sound source at the cost of smaller angles between loudspeakers (Pulkki, 2001b, pp. 28–29) (see also section 2.3.4), are not all in the same plane with the listener. In addition, if more than three loudspeakers are used and if a perceptual sound source crosses the side of this triangle, then another triplet assumes the control of the perceptual sound source, so that triangles do not overlap (Pulkki, 2001b, p. 18).

In two- and three-dimensional set-ups, the directional dispersion or spatial width of a perceptual sound source and the corresponding localization blur (see section 2.5) are dependent on the panning direction (cf. section 2.3.4.3), which means that the same signal can be either feeding a loudspeaker only, causing the smallest possible dispersion of the perceptual sound source when its angular direction is coincident with that of a loudspeaker (see section 2.3.4), or up to three loudspeakers at a time, leading to a higher dispersion (Pulkki, 1999, p. W99-3).

Furthermore, the VBAP system produces great timbre changes between panning on and panning between loudspeakers (Frank, 2013, pp. 81–82), an effect which, together with changes of the directional dispersion, will be most perceptible if the perceptual sound source is set in motion around a listener, as already referred to in section 2.3.4. However, in a two-dimensional VBAP with an eight-loudspeakers circular configuration, a more accurate localization of perceptual sound sources is accomplished, in comparison to lower-order Ambisonics (Marentakis & McAdams, 2013, p. 22:2) (see section 2.3.4.7).

In order to maintain the dispersion and the timbre of a perceptual sound source homogeneous, so that there is no dependence on the panning direction, Multiple Direction Amplitude Panning (MDAP) can be used (Pulkki, 2001b, p. 19) (see section 2.3.4.9). Thus, whereas other systems, such as Ambisonics, apply the

same signal to all loudspeakers with different gain factors, in MDAP the signal feeds only a subset of them, which leads to a better directional quality performance.

2.3.4.9 Multiple Direction Amplitude Panning (MDAP)

Multiple Direction Amplitude Panning (MDAP) was proposed by Pulkki (1999), and it is a solution to maintain the dispersion and the timbre of a perceptual sound source, reproduced by a pair or a triplet of loudspeakers, homogeneous, which Vector-Based Amplitude Panning (VBAP) cannot perform, as mentioned in sections 2.3.4 and 2.3.4.8. This is achieved by applying the same signal to more than one loudspeaker at a time, degrading the perceptual sound source's quality, essentially when its angular direction coincides with that of a loudspeaker, but leaving it identical in the case when the perceptual sound source lies between two or three loudspeakers (Pulkki, 1999, pp. W99-3, W99-4).

Thus, the loudspeaker gain factors are calculated for multiple panning directions around any panning direction, and the resulting gain factors of each loudspeaker are then added together and normalized (Pulkki, 2001b, pp. 18, 35). The dispersion of a perceptual sound source is therefore increased and made independent of its panning direction, where the dispersion angle is defined as the largest angle between the multiple panning directions, and the panning direction is defined by the average direction of the panning directions (Pulkki, 1999, p. W99-3). The signals are then reproduced simultaneously in multiple panning directions near the direction of the perceptual sound source, which is still perceived by a listener as a single sound source.

2.3.4.10 Virtual Microphone Control (ViMiC)

Virtual Microphone Control (ViMiC) is a virtual environment generated by computer, which was proposed by the psycho-acoustician, aural architect, and experimental musician Jonas Braasch[46], in 2005. Gains and delays between virtual microphones and virtual sound sources, which can be placed and moved in a virtual room, are computed, taking their distances and the axes orientations of the microphone polar or directivity patterns into account (Marshall et al., 2006, pp. 360–361). Direct sound, early reflections, reverberation, geometry, absorbing and reflecting properties of the virtual enclosed environment are also considered in the calculations. As a result, sound source propagation is simulated and loudspeakers in a real room are then fed with the estimated microphone signals (Marentakis et al., 2008, p. ICAD08-2).

2.3.4.11 Wave Field Synthesis (WFS)

Wave Field Synthesis (WFS) is a system, whose concept was proposed by A. J. Berkhout in 1988 and put into practice in 1993 at the Technical University Delft, based on Huygens' principle (Berkhout et al., 1993, p. 2767; de Vries & Boone, 1999, p. 15; Birkner, 2004, p. 43). The Huygens' principle was described by the Dutch mathematician and physicist Christian Huygens (April 14, 1629 - July 8, 1695)[47] in 1678, which states that any point of a primary wavefront (see section 2.2.2) can be considered as a source of small spherical secondary wavelets that propagate with a speed and frequency equal to those of the primary wave (Tipler, 1999, p.

[46] Retrieved 24/05/2016, from `http://www.aes.org/events/135/presenters/?ID=1162`
[47] Retrieved 25/05/2016, from `https://en.wikipedia.org/wiki/Christiaan_Huygens`

1040), so that the wavefront at a later time is the surface tangent to the secondary wavelets, also designated as their envelope (Kane & Sternheim, 1988, p. 551). This principle can be described mathematically by the Kirchhoff-Helmholtz integral, whose complete expression can be found in Berkhout et al. (1993, p. 2767). It states that if the sound pressure and its gradient (see section 2.3.4.7) are known on a surface *S*, comprised of virtual secondary sources generated by the primary sources, of a closed, source-free volume *V*, then the sound pressure can be calculated at any point within that volume. When the closed surface *S* is changed to an infinite or almost infinite plane, separating the primary source area from the listening area, the Rayleigh II integral can be applied (Berkhout et al., 1993, p. 2767).

Thus, a sound field, generated by a primary source, can be captured by an array consisting of a large number of microphones placed in the primary sound field, the signals of which can be recorded on a multitrack system. Therefore, the sound field can be reconstructed over a wide horizontal listening area, and not only at a point, by using hundreds of loudspeakers (Odowichuk, 2012, p. 28) that are placed very close to each other in a horizontal linear array, so that the shape of the wave front is unchanged in the direction along the linear array (Berkhout et al., 1993, p. 2770), the result of which is identical to that of a real sound event from the perception point of view. Furthermore, sound sources can be reproduced outside and inside the listener's area (de Vries & Boone, 1999, p. 17). The reproduced sound field is spatially and temporally accurate if the loudspeakers are maximally a half wavelength (see section 2.1.3.1) apart from each other. The highest frequency reproduction is therefore limited up to about 1200 Hz (Pulkki, 2001b, pp. 15–16; Marentakis et al., 2008, p. ICAD08-2), but is considered to be acceptable, because the important spatial hearing low frequency Inter-aural Time Differences' (ITD) cues (see section 2.5.4.1.1) are produced rather accurately up to that frequency. Nevertheless, reflections of the listening room have to be minimized in order to avoid distance, depth, and spatialization errors. To prevent that diffraction waves are generated near the edges of the finite number loudspeaker array, the levels of the loudspeakers near these edges are decreased, and the listening area becomes therefore slightly limited (Berkhout et al., 1993, p. 2771).

Summing up, in WFS, perceptual sound sources are localized in the same position and plane waves are localized in the same direction anywhere in a large listening area (Theile & Wittek, 2004, pp. 394–395). In addition, the amplitude of a perceptual sound source increases when a listener comes closer to it in the listening area, and the amplitude of a source in an infinite distance, represented by a plane wave, is not significantly altered anywhere in the listening area, although cylindrical rather than plane waves are generated by the horizontal loudspeaker configuration, which causes a decrease of 3 dB instead of a constant sound level per distance doubling (see section 2.2.1).

2.4 Human Ears and the Head

Human beings have a hearing system consisting of two ears, each one located respectively on the left and right sides of the head. In the following paragraphs, a broad overview of this system is presented. A more detailed description about this issue can be found in Blauert (1997), Henrique (2007), Pedroso de Lima (2012), and Moore (2013); this aspect goes beyond the scope of this thesis.

Thus, each ear is commonly divided in three parts (see figure 2.39): 1) the external or outer ear (auris

externa); 2) the middle ear (auris media); and 3) the inner ear (auris interna) (Blauert, 1997, p. 52; Henrique, 2007, p. 810; Moore, 2013, p. 23).

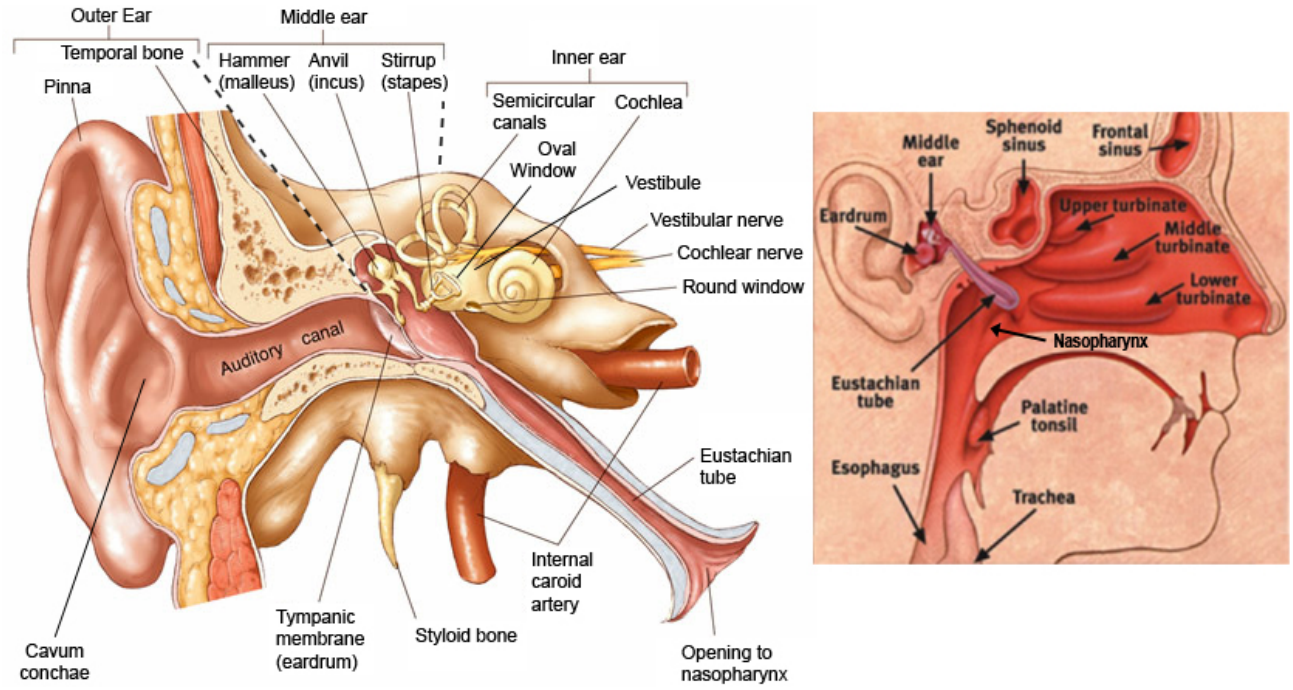


Figure 2.39: Human Ear (retrieved 06/06/2016, adapted from `http://www.academyhearing.ca/blog/news/Blog/2014/09/24/5:how-does-human-hearing-work` and `https://s-media-cache-ak0.pinimg.com/736x/8a/0d/bf/8a0dbf593f546b728e59f9d9ee93bb67.jpg`).

The external ear consists of the auricle or auricula (pinna) and the external auditory meatus (meatus acusticus externus) or auditory canal, which is closed inside by the eardrum or tympanic membrane (membrana tympani), separating the external ear from the middle ear. The middle ear, also called tympanic cavity (cavum tympani), consists of a space filled with air, where, besides other elements, the eardrum, the three smallest bones of a human's body or ossicles, known as hammer (malleus), anvil (incus), and stirrup (stapes), and the Eustachian tube (tuba auditiva) can be found. In turn, the inner ear is located in a complex shaped bone cavity, being thus protected against external disturbances and resonances, and is composed of the cochlea (containing the Organ of Corti), the vestibule (vestibulum), and the three semicircular canals (ductus semicirculares).

The pinna is located outside the head, behind and below its centre (see section 2.4.2), and surrounds the open end of the auditory canal, at an angle of about 25º to 45º relatively to the surface of the head (Blauert, 1997, p. 53), so that the top of the pinna is normally aligned with the outer corner of an eye (Magee, 2008, p. 95). The pinna consists of cartilage covered with skin, with differing irregular shape and size from person to person and slightly different between the left and the right ear, which alters or filters the spectral contents of sounds that reach the eardrum, in a location- and frequency-dependent manner, especially at higher frequencies, where the wavelengths λ (see section 2.1.3.1) are comparable with the size of the pinna (Stern et al., 2005, p. 3). It is essentially this direction-dependent filtering produced by the pinna that allows a person to perceive sounds outside the head (Moore, 2013, p. 280). This effect is important in the process of localization of sound sources in a free field (see sections 2.1.2 and 2.1.3.1), including localization of sound sources in the vertical plane, so that sounds coming from above and below can be distinguished (Odowichuk,

2012, p. 30), and any front-back ambiguities that can occur can be resolved (see section 2.5). The spatial cues provided by the pinna to the remaining hearing system of a person contribute to his or her unique individual Head-Related Transfer Functions (HRTFs) (see section 2.4.1), that is, the pinna codes spatial attributes of the sound field into temporal and spectral attributes (Blauert, 1997, p. 63). When a sound is received at incidence angles different from 0º (front) and 180º (back), then it reaches the most distant ear later and weaker (see section 2.5.4.1). The Inter-aural Time Differences (ITD) (see section 2.5.4.1.1) and the Inter-aural Level Differences (ILD) (see section 2.5.4.1.2), caused respectively by the difference of arrival times between the propagation paths to both ears, essentially for sounds containing frequencies below about 1.5 kHz, and by the shadowing effect of the head for sounds containing frequencies above approximately 1.5 kHz, where the head acts as a barrier because of the shorter wavelengths of these frequencies when compared with the size of the head, are therefore the main cues of sound source localization (Pulkki, 1999, p. W99-1; Pulkki & Karjalainen, 2001, p. 740). At an incidence angle of 90º or 270º (sides), the inter-aural attenuation is quite marked (Blauert, 1997, p. 72). In addition, the pinna also minimizes wind noises.

The central part of the pinna, where the concha, more precisely the cavum conchae, is located, leads to the auditory canal, which is an about 25 to 30 millimetres long curved tube with an approximately 7 to 8 millimetres variable diameter, closed by the eardrum (Blauert, 1997, p. 53; Henrique, 2007, p. 811). As already mentioned in section 2.1.3.2, the auditory canal has a main resonance at approximately 3.4 kHz. A second resonance due to its non-uniform shape is also present at about 13 kHz (Howard & Angus, 2001, p. 83) (see figure 2.12 in section 2.1.3.2). The eardrum is nearly circular and concave, resembling a loudspeaker cone, acting as a pressure-sensitive receiver (Blauert, 1997, p. 128). On average, it is approximately 0.1 millimetres thick, and positioned at an angle of 40º to 50º relatively to the axis of the auditory canal (Blauert, 1997, p. 53).

The Eustachian tube, named after the Italian anatomist Bartolomeo Eustachi (1500 or 1514 - August 27, 1574)[48], is an about 35 to 38 millimetres long and narrow duct, which connects the tympanic cavity to the nasopharynx at the back of the nasal cavity, where this tube is normally closed (Henrique, 2007, p. 815). Sometimes, it opens when a person swallows, yawns, or shouts, so that an identical air pressure on both sides of the eardrum is obtained, allowing it to vibrate properly, when pressure variations of sound signals are present in the auditory canal. These vibrations are then received by the hammer, which is attached to the top of the eardrum pulling it inwards and making it concave, and transmitted to the inner ear through the two other ossicles (Henrique, 2007, p. 814; Pedroso de Lima, 2012, p. 211). The stirrup is in turn attached to the oval window membrane in the inner ear, so that the cochlea can receive the incoming vibrations. A reflexive contraction of muscles in the middle ear due to sound pressure levels above about 80 dB, with maximum effect over 100 dB, known as *acoustical reflex*, causes a reduction of the auditory sensitivity (Blauert, 1997, pp. 54, 63). It should be stressed out here that another way of reception of sound in the inner ear, beyond that by the eardrum-ossicle path, is that by bone conduction through the temporal bone, which is usually absent when a person listens to his or her own captured or recorded voice through loudspeakers, being of secondary importance under normal conditions.

The cochlea fulfils an auditory function (see figure 2.40). It is an about 30 millimetres long tapered spirally-wound tube (Henrique, 2007, p. 817), which is divided longitudinally into three parts: the vestibular duct (scala

[48] Retrieved 10/06/2016, from `https://en.wikipedia.org/wiki/Bartolomeo_Eustachi`

vestibuli), the tympanic duct (scala tympani), and the cochlear duct (scala media) (Pedroso de Lima, 2012, p. 218). The Reissner's membrane, named after the German anatomist Ernst Reissner (September 24, 1824 - September 16, 1878)[49], separates the vestibular duct from the cochlear duct, and the basilar membrane separates the cochlear duct from the tympanic duct. The vestibular duct, originating at the base of the cochlea at the oval window membrane, and the tympanic duct, terminating on the border of the middle ear at the round window membrane, contain both a liquid, called perilymph, and are respectively located above and bellow the cochlear duct. At the apex of the cochlea, the vestibular and the tympanic ducts merge with each other through the helicotrema (Moore, 2013, p. 25). When vibrations enter the cochlea through the oval window membrane, the virtually incompressible perilymph is displaced accordingly, so that the oval window membrane and the round window membrane vibrate with opposite phases. The cochlear duct in turn contains a liquid, called endolymph, and the Organ of Corti, which lies on the basilar membrane along its length. The Organ of Corti, named after the Italian anatomist Alfonso Giacomo Gaspare Corti (June 22, 1822 - October 2, 1876)[50], consists of thousands of sensory cells, known as inner and outer hair cells, which can convert or encode mechanical sound vibrations into electrical impulses that are then sent to the brain via the auditory nerve for further processing. A pressure difference is therefore applied across the basilar membrane when the oval window membrane vibrates, which sets the basilar membrane in motion and stimulates the hair cells, causing audition to take place (Pedroso de Lima, 2012, p. 17; Moore, 2013, p. 25).

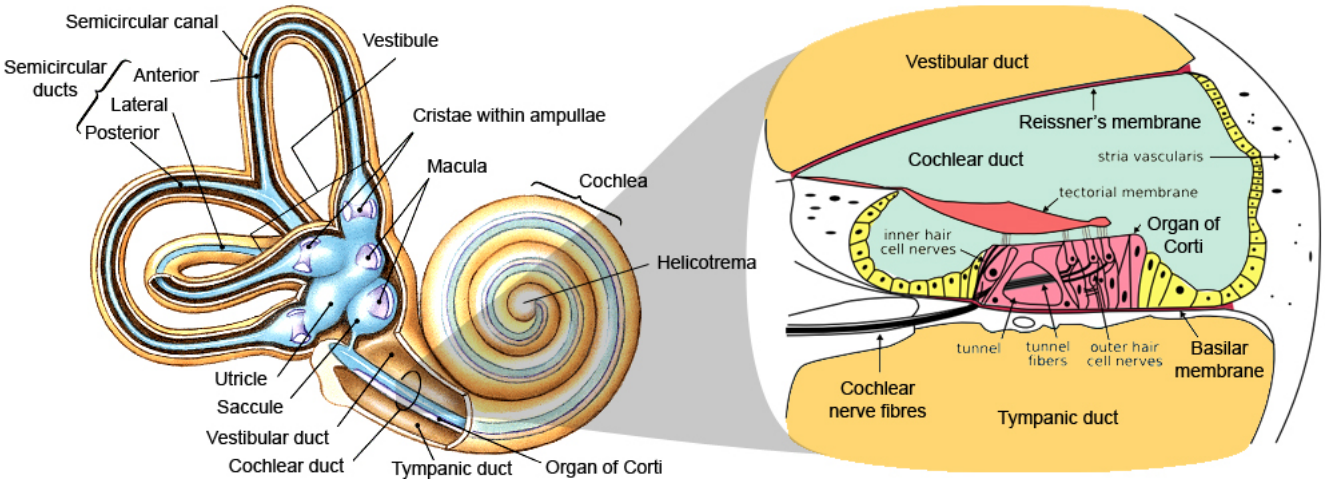


Figure 2.40: Inner Ear: the semicircular canals, the vestibule, and the cochlea (left); the vestibular duct, the cochlear duct, and the tympanic duct (right) (retrieved 06/06/2016, adapted from `http://www.spiralzoom.com/Science/SpiralsHumanBody/SpiralsHumanBody.html` and `https://upload.wikimedia.org/wikipedia/commons/0/0c/Cochlea-crosssection.png`).

The basilar membrane is, however, subject to a pattern of motion, whose amplitude peak position varies with the frequency of stimulation. Thus, whereas the displacement of the basilar membrane is maximum close to its base and almost non-existent throughout the remaining part of the membrane when high-frequency sounds enter the cochlea, the displacement of the basilar membrane is maximum close to the apex, but also present throughout the remaining membrane, when low frequencies are received (Moore, 2013, p. 25). For a sinusoidal signal or pure tone, the frequency that makes a certain place of the basilar membrane to reach

[49] Retrieved 10/06/2016, from `https://en.wikipedia.org/wiki/Ernst_Reissner`
[50] Retrieved 10/06/2016, from `https://en.wikipedia.org/wiki/Alfonso_Giacomo_Gaspare_Corti`

a maximum amplitude is called the Characteristic Frequency (CF) for that place (Moore, 2013, p. 26). For complex sounds, different patterns of motion of the basilar membrane are produced with many maxima, so that the highest maximum may not match the CF corresponding to the fundamental frequency (see section 2.1.3.1), although, in general, the perceived pitch still corresponds to this component (see section 2.1.3.2). The place representation of frequency along the basilar membrane gave rise to the now well established place theory of hearing (Moore, 2013, pp. 203–204). Since this organization is maintained in the auditory nerve and other auditory areas in the brain, it is known as tonotopic organization (Moore, 2013, p. 39). As a consequence, neurons with different CFs are also excited (Stern et al., 2005, p. 4).

The vestibule and the three semicircular canals accomplish essentially an orientation and body balance function, together with the visual and sensory systems, which is an issue that will not be covered in this thesis.

2.4.1 Head-Related Transfer Function

The only valuable information that reaches the eardrums is sound pressure variation of sound signals, which can be further processed by the brain in order to generate the perception of spatial images and to localize sound sources (see section 2.5). These sound pressure variations are the result of a sum of variable and fixed factors, such as: 1) the variable sound source spectrum as a function of direction, because of the directivity factor Q or directivity index D_I of the sound source (see sections 2.1.3.1 and 2.2.1); 2) the variable effects caused by the transmission path from a sound source to an ear due to distance and reflection losses (see section 2.1.3.1); 3) the variable pinna, shoulder, and torso reflections, the variable concha resonances, and the variable head diffraction, all these latter factors dependent on the angle of incidence of sound; and 4) the constant inherent auditory canal resonances, which are characteristic of the frequency response or transfer function of an auditory canal (Streicher & Everest, 1998, p. 6.3) (see figure 2.41).

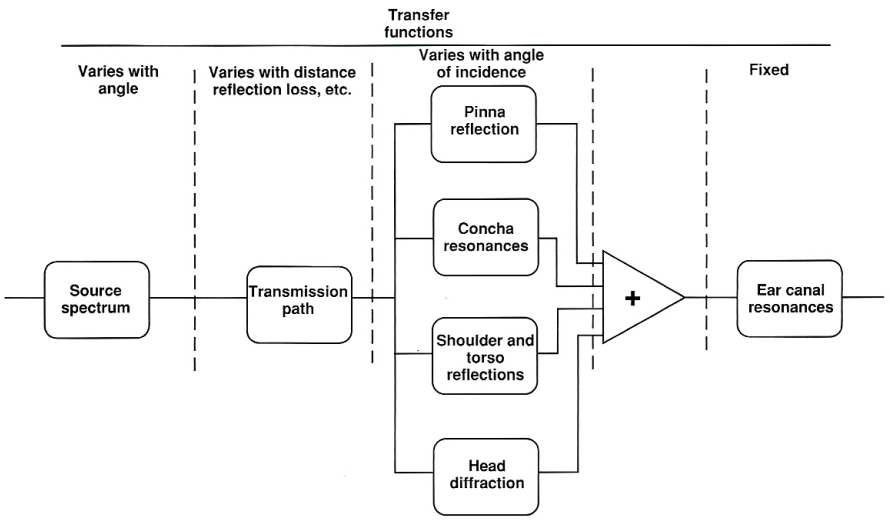


Figure 2.41: Transfer functions (Streicher & Everest, 1998, p. 6.3).

Therefore, a transfer function describes the alteration in the frequency domain that sound signals are subject to by passing through a certain system (Blauert, 1997, p. 373), and can be represented as a graph of the sound pressure and timing variations with frequency (Streicher & Everest, 1998, p. 6.1). In the time domain,

this modification is in turn described by an impulse response. Thus, a single Head-Related Transfer Function (HRTF) is a specific person's left or right ear far-field transfer function, which describes the alteration in the frequency domain that sound signals are subject to by the respective outer ear, before they reach the respective eardrum, measured from a certain point in the free field to a certain point in the auditory canal (Cheng & Wakefield, 2001, p. 233). In the time domain it is designated as Head-Related Impulse Response (HRIR) (Stern et al., 2005, p. 3).

Since the shape and size of the outer ears vary from person to person and slightly between the left and the right ear (see section 2.4), HRTFs also vary from person to person and between the left and the right ears, as can be shown when measuring HRTFs (Audio Products Division of National Semiconductor Analog Products Group, n.d.). In addition, HRTFs are different for every new sound source position, so that variations of the frequency response above 4 kHz as a function of azimuth and elevation (see section 2.3.2) can therefore be observed (Stern et al., 2005, p. 3).

As already referred to in section 2.3.4, HRTFs can be used with headphones in binaural sound reproduction, in order to enable the localization of sound sources outside the head by simulating the effect of the outer ears. Thus, HRTFs can be measured: 1) on an individual listener; 2) on a generalized dummy head or an anatomically realistic manikin, such as the Knowles Electronics Manikin for Acoustic Research (KEMAR) (Stern et al., 2005, p. 7); 3) on several listeners and then averaged; or 4) synthesized through mathematical calculations (Malham, 1998, p. 171).

In the first case, the results can be the best among all approaches, although problems may occur if the HRTFs stay fixed to the head when it moves, rather than to the surrounding environment, which can be solved by tracking the head. Thus, the HRTFs are updated accordingly with the head movement in order to maintain the absolute positions of sound sources in the surrounding environment constant (Kendall, 1995, p. 38). Furthermore, it is very difficult to measure every possible HRTF for each listener. For this reason, the second and third situations are mostly used, although front-back ambiguities can occur, fundamentally due to high-frequency differences between the generalized and individual HRTFs, which can be improved by tracking the head. In the fourth case, calculations are nowadays made with high-speed computers or digital signal processors, so that interpolations for locations between known HRTF locations are also performed for each distinct sound source. Databases of HRTFs can be provided by some institutions, such as that of the University of California, at Davis, in the USA (`http://interface.cipic.ucdavis.edu/sound/hrtf.html`, visited on 12/06/2016).

2.4.2 Cone of Confusion

A cone of confusion is defined theoretically as a set of locations on the left or on the right side of the head, where the difference of their distances to both ears is constant (Blauert, 1997, p. 179–180; Pulkki, 2001b, p. 4; Pulkki & Karjalainen, 2001, p. 740) (see figure 2.42).

At these locations, the Inter-aural Time Differences (ITD) (see section 2.5.4.1.1) and/or the Inter-aural Level Differences (ILD) (see section 2.5.4.1.2) are the same, with the result that a listener can describe the extent to which a sound is to his or her left or right side, but is not able to specify whether this sound is produced from

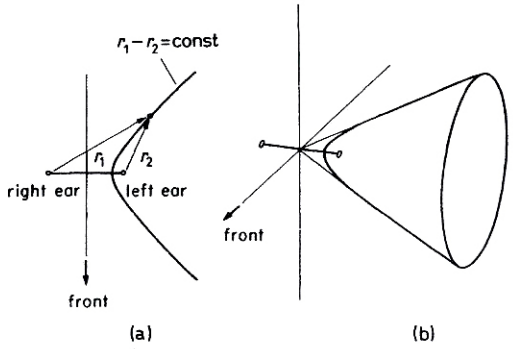


Figure 2.42: Cone of Confusion: seen as a hyperbola in the horizontal plane (a); seen as a cone in three dimensions (b) (Blauert, 1997, p. 179).

the front, back, top, bottom, or from any other direction defined by these locations on the edge of the base of the geometric form resembling a cone, whose axis of symmetry passes through the listener's ears with its apex at the centre of the head, at any given distance from the ears, unless the listener uses the spectral cues and moves the head relatively to the sound source, in order to avoid any of these ambiguities or confusions of localization at a given degree of lateralization (Hollerweger, 2006, p. 16).

Although the Austrian ethnomusicologist Erich Moritz von Hornbostel (February 25, 1877 - November 28, 1935)[51] and the Austro-Hungarian psychologist Max Wertheimer (April 15, 1880 - October 12, 1943)[52] already knew this phenomenon in 1920 (Cheng & Wakefield, 2001, p. 232) and used a simple model to calculate the ITD, where both ears were considered as points in free space, separated by 21 centimetres, without taking the head into account (Blauert, 1997, pp. 73, 179), the term *cone of confusion* was suggested by the American psychologist Robert Sessions Woodworth (October 17, 1869 - July 4, 1962)[53] and the American professor of psychology Harold H. Schlosberg (January 3, 1904 - August 5, 1964)[54] in 1954, assuming that the head had a spherical form (Duda et al., 1999, p. 965; Aaronson & Hartmann, 2014, p. 818). Nevertheless, the head is actually not spherical at all, but rather ellipsoidal, and the ears are not positioned across a diameter, but rather behind and below the centre of the head (Duda et al., 1999, p. 965) (see section 2.4). For these reasons, the ITD varies around a cone of confusion (Pulkki, 1999, p. W99-1; Castellanos, 2006, p. 15), so that the ITD is a function of elevation and azimuth, as was shown by Duda et al. (1999).

2.5 Sound Source Localization

Localization is defined by Blauert (1997, p. 37) and Hammershøi (2009, p. 3) as the law, rule, or process by which the location of a perceptual sound (see section 2.1.2) is related to a particular attribute or attributes of a physical sound (see sections 2.1.1 and 2.1.3.1), or of another event that is somehow correlated with the perceptual sound. In mathematical terms, it is defined as the function which relates the points of the physical space (see section 2.3.2) to those of the perceptual space (see section 2.3.3) (Blauert, 1997, p. 38).

[51] Retrieved 11/06/2016, from https://en.wikipedia.org/wiki/Erich_von_Hornbostel
[52] Retrieved 11/06/2016, from https://en.wikipedia.org/wiki/Max_Wertheimer
[53] Retrieved 11/06/2016, from https://en.wikipedia.org/wiki/Robert_S._Woodworth
[54] Retrieved 11/06/2016, from https://en.wikipedia.org/wiki/Harold_H._Schlosberg

Nevertheless, sound signals at the eardrums, important for a listener to generate the perception of spatial images and to localize sound sources (cf. section 2.4.1), have attributes that require only one ear (monaural cues; see section 2.5.4.2) or both ears (inter-aural cues or binaural cues; see section 2.5.4.1) to process them (Blauert, 1997, p. 93; Pedroso de Lima, 2012, p. 313).

As already mentioned in section 2.1.2, different perceptual sounds can be perceived simultaneously, even if there is only one physical sound source. Sound source localization, as it is called in Psychoacoustics, can be influenced by the position of the physical sound source and the type of signal it produces, and is therefore dependent on the following cues or factors (Roads et al., 1996, pp. 457–468; Blauert, 1997, p. 40; Rumsey, 2001, pp. 21–29, 33–36; Howard & Angus, 2001, pp. 96–105; Hollerweger, 2006, pp. 12–21): 1) Horizontal angle or azimuth φ in a spherical coordinate system (see sections 2.3.2 and 2.5.1); 2) Zenith (altitude), vertical angle, or elevation angle δ in a spherical coordinate system (see sections 2.3.2 and 2.5.2), in which the spectrum of a physical sound source can be modified due to the reflections of sound off the torso, the shoulders, and the pinnae (see section 2.4.1); and 3) Distance (for static sound sources; see section 2.2.3) or velocity (for moving sound sources, in which the Doppler Effect is noticed; see section 2.2.4) (see also section 2.5.3).

The azimuth cue, in turn, depends on three factors: 4) Inter-aural Time Difference (ITD), representing the difference between arrival times of a single wavefront from a single physical sound source to both ears, essentially consisting of low frequencies below about 1.5 kHz (see section 2.5.4.1.1); 5) Inter-aural Level Difference (ILD), Inter-aural Intensity Difference (IID), or Inter-aural Amplitude Difference (IAD), describing the difference in level, intensity, or amplitude of sounds containing frequencies above approximately 1.5 kHz, perceived by both ears as a result of the shadowing effect of the head for these sounds (see sections 2.4 and 2.5.4.1.2); and 6) Spectral cues derived from reflections of sound off the torso, the shoulders, and the pinnae (see section 2.4.1). Cues 4) and 5) are part of the duplex theory the English physicist John William Strutt (November 12, 1842 - June 30, 1919)[55], third Baron of Rayleigh, best known as Lord Rayleigh, proposed in 1907 (Pedroso de Lima, 2012, p. 313). In this respect, Pulkki (2001b, pp. 26, 36) and Pulkki & Karjalainen (2001, p. 750) found out that the most consistent cues of localization of perceptual sound sources in stereophonic listening are the ITD cues at low frequencies below about 1.5 kHz and also, to some extent, the ILD cues at higher frequencies above approximately 1.5 kHz. Furthermore, the ITD cues were found to be unreliable at frequencies above approximately 1.5 kHz and ILD cues were found to be unstable close to the frequency range from 700 Hz to 2 kHz, giving rise to localizations of perceptual sound sources at directions outside the area between loudspeakers (see section 2.3.4.3).

It must be noted here that there is a dissimilarity between the spatial perception resulting from the arrival of a single wavefront from a single physical sound source to both ears and the spatial perception resulting from two slightly delayed arrivals of an identical sound (two wavefronts) approaching both ears from different directions (see figure 2.43). The former is essentially related to the ITD cue and the latter to the so-called *law of the first wavefront*, *Haas effect*, or *precedence effect* (see section 2.5.5), as already referred to in section 2.2.2, according to which the sound arriving first at one ear is the one used by the brain to determine the location of the sound source (Rumsey, 2001, pp. 26-27; Castellanos, 2006, p. 13).

[55] Retrieved 14/06/2016, from `https://en.wikipedia.org/wiki/John_William_Strutt,_3rd_Baron_Rayleigh`

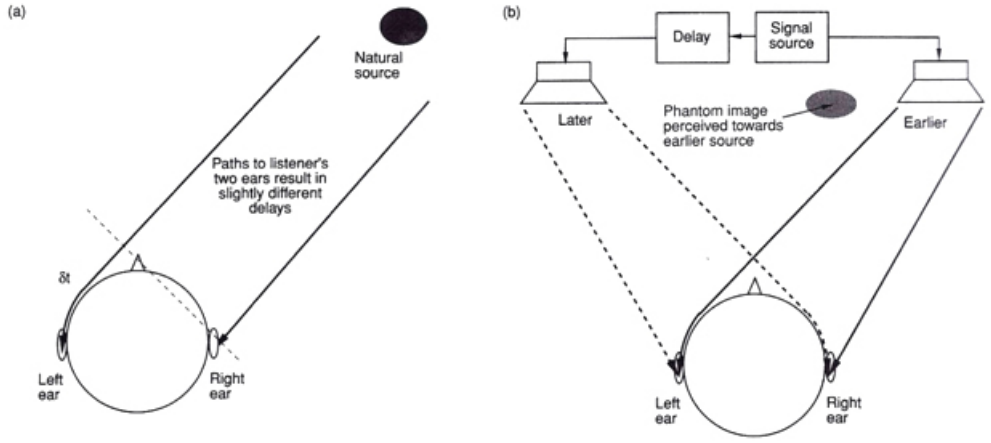


Figure 2.43: (a) Arrival of a single wavefront (from a single physical sound source) to both ears; (b) Slightly delayed arrivals of an identical sound (two wavefronts) approaching both ears from different directions (adapted from Rumsey, 2001, p. 27).

The distance cue also depends on three factors: 7) The ratio between the direct and the reverberated sound, when the intensity of the direct sound decreases according to the inverse square law (see section 2.1.3.1); 8) The loss of high-frequency components with distance; and 9) The loss of detail with distance.

Moore (2013, p. 280) refers that the human hearing system is best at localizing sounds in the horizontal plane, reasonable in the vertical plane, and worst for distance.

However, sometimes there are conflicting cues causing ambiguous localization. One of the examples is the case in which physical sound sources provide almost identical inter-aural signal differences. This occurs when a physical sound source is located in the median plane (Hollerweger, 2006, pp. 16–17; Castellanos, 2006, pp. 14–15) (see sections 2.3.2 and 2.5.2). As a result, other cues come into play, such as context dependent cues and those from other senses (see section 2.5.4.4), as visual ones (see section 2.5.4.3), if related to a sound source. The environment also provides additional cues in the form of sound scattering, absorption and reflections caused by objects, as well as by the human body (Hollerweger, 2006, pp. 20–21; Rumsey, 2001, pp. 34–35). In addition, Blauert (1997, pp. 96–97) found out that there is no advantage in immobilizing the listener's head in experimental investigations of spatial hearing with one sound source in the median plane, because the head does not normally move more than about 1° to the left or right to allow localization of the sound source in that plane.

In order to determine the limits of localization capacity of the human hearing system, several studies have been made on absolute localization performance, on just noticeable differences in direction or distance (see section 2.1.3.2), and on direct source identification (Hammershøi, 2009, p. 4).

In this sense, localization blur is defined by Blauert (1997, pp. 21, 37) as the smallest change in one or more particular attributes of a physical sound or of another event that is somehow correlated with a perceptual sound which leads to a change in the location of the perceptual sound. The auditory system has therefore less spatial resolution than that of physical space, so that a physical point sound source is always perceived as a perceptual sound spread out to a certain degree in space.

A measure of accuracy in azimuth or elevation location is determined by the Minimum Audible Angle (MAA),

which is defined as the smallest perceivable angular sound displacement from a given starting position (Stern et al., 2005, p. 8), as already mentioned in section 2.2.4. Another measure is the Minimum Audible Movement Angle (MAMA), which is defined as the angular distance a moving sound has to traverse before its movement is perceived by a listener (Marentakis et al., 2008, p. 2; Moore, 2013, p. 276). The MAA is best directly in front of a listener, in the horizontal plane, being equal to about 1º (Saberi et al., 1991, p. 58; Blauert, 1997, pp. 38, 96; Holman, 2000, p. 207; Pulkki & Karjalainen, 2001, p. 743, quoting Blauert; Stern et al., 2005, p. 8) (see section 2.5.1). In front of a listener, in the vertical plane, it is approximately equal to 3º (Holman, 2000, p. 207) (see section 2.5.2). The MAA gets progressively worse to the sides, rear, and above and below, which leads Holman (2000, pp. 203, 207) to argue that this is the reason why psycho-acoustically designed multichannel sound reproduction systems (cf. section 2.3.4) use more front channels than rear ones. Thus, whereas the MAA is greater with lateral or elevated sounds, MAMA grows linearly with the speed of a moving sound.

In addition, the spatial cues provided to the brain by the pinnae and by the reflections off the shoulders and body at different source positions and angles of incidence contribute to the filter functions called Head-Related Transfer Functions (HRTF) (see section 2.4.1), which are unique to each person. Localization of perceptual sound sources is therefore frequency-dependent and is influenced by the different temporal structures of sound signals (Pulkki, 2001b, p. 26; Pulkki & Karjalainen, 2001, p. 739). In this respect, the vertical localization of the perceptual sound source depends on its frequency contents rather than on the actual physical sound source position, that is, if signals with frequencies above 7 kHz are produced by a physical sound source in front of a listener, then the perceptual sound source will be naturally localized at a vertically displaced position (Power et al., 2013, p. 1), described in terms of directional bands (Blauert, 1997, pp. 102, 112) (see sections 2.3.4.7 and 2.5.2).

2.5.1 Localization in the Horizontal Plane

It is in the horizontal plane that the localization of sound sources is carried out as effectively as possible by the human hearing system (Odowichuk, 2012, p. 30; Moore, 2013, p. 280). In this plane, the localization blur or the Minimum Audible Angle (MAA) (see section 2.5) are minimum in the forward direction (0º), having a value of about 1º (Saberi et al., 1991, p. 58; Blauert, 1997, pp. 38, 96; Holman, 2000, p. 207; Pulkki & Karjalainen, 2001, p. 743, quoting Blauert; Stern et al., 2005, p. 8; Moore, 2013, p. 250) (see section 2.5), increasing to the left or to the right, until the value becomes 3 to 10 times greater than the value in the forward direction at angles of -90º or +90º, which makes localization at those angles 3 to 10 times more difficult to be performed (Blauert, 1997, pp. 40–41).

Nevertheless, as there is a strong dependence on the frequency of sound, additional relative minima of localization blur can appear in various directions to the sides, so that localization of sinusoidal sounds and of other narrow-band sounds is different from that of broadband sounds, where various simultaneous or successive perceptual sound sources can arise in different directions (Blauert, 1997, p. 42). With narrow-band sounds, perceptual sound sources can also appear in different directions other than that of the incidence of sound, that is, in directions more or less axially symmetric relatively to the axis of intersection between the horizontal and the frontal planes, which passes through both ears (Blauert, 1997, p. 43). These two more or less

symmetrical directions are usually differentiated by the auditory system based on the spectrum of the incident sounds at both ears, but with narrow-band sounds this information does not exist. This kind of phenomenon can be eliminated in most cases if the duration of sound (see section 2.1.3.1) is long enough and a listener can move his or her head freely, in order to make sufficient exploratory movements, so that the direction of the perceptual sound source coincides with that of the physical sound source (Blauert, 1997, pp. 43–44). Behind a listener the value of localization blur decreases normally to about two times the value in the forward direction from -90° to -180° or from +90° to +180° (Blauert, 1997, p. 41).

2.5.2 Localization in the Median Plane

The paths and the arrival times of sound, produced by a physical sound source that is located in the median plane (see section 2.3.2), are identical at both ears, so that there are almost no inter-aural signal differences (Blauert, 1997, p. 44; Pulkki, 1999, p. W99-1). In this plane, the localization blur or the Minimum Audible Angle (MAA) (cf. section 2.5) in the forward direction are equal to about $\pm 4°$ for white noise (Blauert, 1997, pp. 44, 310) (see section 2.1.4), approximately 9° for continuous speech by a familiar person, and about 17° for continuous speech by an unfamiliar person (Blauert, 1997, p. 44). The localization blur increases up to $\pm 10°$ for white noise at 90° overhead, a value which is doubled if low-pass filtered noise, cut-off at 4 kHz with a 30 dB per octave slope, is used (Blauert, 1997, p. 310).

If very short unfamiliar sounds with impulse contents are produced in the median plane, the corresponding perceptual sound sources are displaced to the rear area of the median plane. However, this phenomenon does not occur if such sounds are previously presented to the listener (Blauert, 1997, pp. 44–45). Anyway, Blauert (1997, pp. 103, 105) states that the coincidence of the direction of the perceptual sound source with that of the physical sound source does not depend on a previous familiarity with any kind of sound, although the performance is improved by it.

Furthermore, localization and localization blur relatively to the position of a physical sound source in the median plane cannot be determined for narrow-band sounds with a smaller bandwidth than about two thirds of an octave (Blauert, 1997, p. 45). In this case, the direction of the perceptual sound source only depends on the frequency contents of the sound. Thus, spectral cues and exploratory movements of the head are usually used in this plane to detect elevation of sounds and to resolve any front-back ambiguities (Pulkki, 1999, p. W99-1).

In addition, it has been found out that the direction of a perceptual sound source usually matches that of the physical sound source in the median plane, when broadband sounds of long duration or repeated several times are used (Blauert, 1997, pp. 97, 101). This happens in front-back directions, and, with respect to the elevation angle, this also happens with broadband sounds including components above 7 kHz, where the localization blur decreases to 4° (Blauert, 1997, p. 102). Since the direction of a perceptual sound source is determined according to the centre frequencies of sounds at both ears, the localization can be described in terms of directional bands (Blauert, 1997, p. 112) (cf. section 2.5). It is essentially the direction-dependent filtering produced by the pinna that provides information about the location of sounds in the median plane (Pedroso de Lima, 2012, p. 20; Moore, 2013, p. 280). As a result, a broadband sound will be perceived from the forward direction if it contains spectral peaks around the directional bands associated with forward direction perception

of about 250 Hz to 500 Hz and 2 kHz to 6 kHz, a notch around 500 Hz to 2 kHz, and increased energy above 13 kHz (Blauert, 1997, pp. 109–114, 311) (see figure 2.44). It will be perceived in an upward direction if it contains a spectral peak around the directional band of upward direction perception of approximately 7 kHz to 9 kHz. If it contains spectral peaks around the directional bands associated with backward direction perception of about 500 Hz to 2 kHz and 10 kHz to 12 kHz, and a notch around 2 kHz to 10 kHz, then it will be perceived from a backward direction.

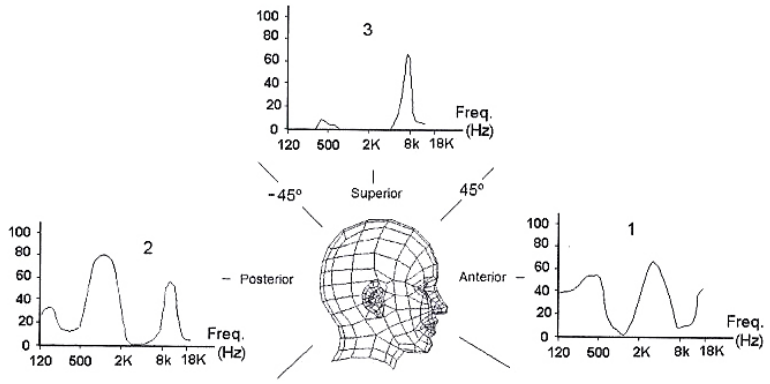


Figure 2.44: Directional bands: 1) Forward direction; 2) Backward direction; 3) Upward direction (Pedroso de Lima, 2012, p. 321).

2.5.3 Localization in the Distance

Distance of a perceptual sound source is defined by Blauert (1997, p. 117) as its distance from the point of intersection between the horizontal, frontal, and median planes of a head-related reference system (see section 2.3.2). The same definition can be applied relatively to the distance of a physical sound source. Localization between the distance of a physical sound source and that of the perceptual sound source can therefore be assessed.

Localization in the distance is greatly influenced by familiarity with the produced sound, that is, localization in the distance is very difficult to judge with unfamiliar sounds (Blauert, 1997, p. 45; Marentakis et al., 2008, p. ICAD08-2). Furthermore, in an enclosed space, the moments in which sound reflections are carried out and their respective levels relatively to the direct sound are indicative of the type of space and of the distance to a physical sound source (Blauert, 1997, pp. 276, 280; Malham, 1998, p. 170; Rumsey, 2001, p. 35). The loudness of reverberation remains therefore more or less constant in a fairly reverberant space and the loudness of the sound source decreases with distance. In this respect, Marentakis et al. (2008, p. ICAD08-2) mention that the relationship between direct and reverberant energy of sound defines the perceived distance of a perceptual sound. Thus, localization of a sound is more accurate if the level of the direct sound is higher than that of the reflections, so that the direct sound forms a perceptual sound, and the generation of other possible following perceptual sounds below the listener's echo threshold of about 30 milliseconds (Howard & Angus, 2001, p. 105) is eliminated by the hearing system, according to the law of the first wavefront (Blauert, 1997, p. 279) (see sections 2.2.2 and 2.5.5). If, however, the level of reflections is higher than that of the direct sound, then a diffusely located reverberant perceptual sound can be perceived, where the perceptual sound

due to the direct sound is masked by the reverberant sound (see section 2.1.3.2).

For broadband direct sounds and physical sound sources in the median plane, certain attributes of sound at both ears depend on the distance of the physical sound source from a person, as follows (Blauert, 1997, pp. 118–119):

A) For point sound sources (see section 2.2.1) at distances of less than approximately 3 metres: 1) the curvature of the wavefronts (see section 2.2.2), which reach the head, is noted; 2) sound distortions caused by the head and by the outer ear vary with distance; and 3) the sound pressure level and the shape of the spectrum of sounds at both ears change with distance, so that the sound pressure level decreases according to the inverse square law as in a free sound field, that is, 6 dB for each doubling of the distance (see section 2.1.3.1). In this distance range, the transfer function of the outer ears (see sections 2.4 and 2.4.1) depends largely on distance (Blauert, 1997, p. 129), and twice the loudness (see section 2.1.3.2) is perceived if the sound pressure level at both ears is increased by approximately 10 dB.

B) For point sound sources at distances of about 3 to 15 metres, if the original sounds do not vary, only the frequency-independent sound pressure level of sounds at both ears changes with distance, so that the sound field becomes more and more similar to a plane wave with distance (see section 2.2.1). However, sound pressure level is also related to loudness and timbre (see section 2.1.3.2) of a perceptual sound (Blauert, 1997, p. 120). Thus, loudness increases with sound pressure level and low-frequency components become more perceptible relatively to high-frequency components (see figure 2.12 in section 2.1.3.2), which means that the timbre becomes darker. In turn, loudness and timbre depend on the shape of the spectrum of sound at both ears. As a result, in this situation, distance of a perceptual sound source decreases with increasing level of sounds at both ears, so that the distance of the perceptual sound source is independent of the distance of the physical sound source (Blauert, 1997, pp. 121–122). In addition, the distance of the perceptual sound source tends to increase more slowly than that of the physical sound source, if sound pressure level is the only attribute accessible to the hearing system. The doubling of the distance of a perceptual sound is therefore only achieved when the level decreases about 20 dB.

C) For sound sources at distances greater than about 15 metres: 1) the frequency-independent sound pressure level of sounds at both ears still changes with distance, as in item B), but higher frequencies are attenuated additionally, more than lower ones, due to air attenuation caused by the moisture contents of the air and by wind speed (see sections 2.1.1 and 2.1.3.1); and 2) the shape of the spectrum of sounds at both ears also changes with distance. In this case, the distance to a perceptual sound is limited (Blauert, 1997, pp. 122, 127).

Sound distance can be simulated in music by using weaker sounds or lower sound pressure levels for longer distances, and louder sounds or higher sound pressure levels for shorter distances, so that a sound can be perceived in the foreground or in the background (Zelli, 2009). The use of more or less instruments, the dynamic variation of music, and the distance of the instruments themselves cause the listener to perceive

variable sound spectra. As already mentioned in section 2.2.4, when a physical sound source approaches or moves away from a listener, the perception of auditory movement is primarily dominated by sound pressure level cues. *Dreamsong* (1978) is an example of a composition, in which the composer, performer, installation artist, and computer music veteran Michael McNabb[56], uses distance as an important compositional element.

2.5.4 Localization Cues

The main cues used to determine the localization of a sound source are: 1) the Inter-aural Time Differences (ITD) (see section 2.5.4.1.1); 2) the Inter-aural Level Differences (ILD), also known as Inter-aural Intensity Differences (IID) or Inter-aural Amplitude Differences (IAD) (see section 2.5.4.1.2); 3) the Head-Related Transfer Functions (HRTF) (see section 2.4.1); and 4) the possibility of head movement (Malham, 1998, pp. 169–170). The first two cues, also called binaural cues in this context because they are dependent on both ears (see section 2.5.4.1), are part of the duplex theory of Lord Rayleigh, as already referred to in section 2.5. The HRTFs are in turn monaural cues, because they depend on one ear only (see section 2.5.4.2). Apart from the monaural and binaural cues, other cues, such as visual ones (see section 2.5.4.3), also contribute to the localization of a sound source. Nevertheless, localization cues are degraded for perceptual sound sources (Marentakis & McAdams, 2013, p. 22:2).

2.5.4.1 Binaural Cues

The duplex theory of Lord Rayleigh (see section 2.5) states that the two main cues of sound source localization in lateral positions are the frequency-dependent Inter-aural Time Differences (ITD) and the Inter-aural Level Differences (ILD), to which humans are very sensitive (Blauert, 1997, p. 177; Pulkki, 1999, p. W99-1; Pulkki & Karjalainen, 2001, p. 740; Pulkki, 2001a, p. 754; Stern et al., 2005, p. 2; Zelli, 2009; Moore, 2013, p. 280). Given that the meaning of the word *inter-aural* is *difference or relationship between sounds at both ears* (Blauert, 1997, p. 93), the ITD is essentially due to time differences of sounds at both ears below about 1.5 kHz, where the size of the head is smaller than the wavelengths of sounds, so that the waves diffract around the head as if it were not there. The ILD, where spectral differences occur, is in turn due to the shadowing effect of the head for frequencies above approximately 1.5 kHz, where the wavelengths are smaller than the size of the head (Pulkki, 2001b, p. 7; Pedroso de Lima, 2012, p. 19). In addition, since ITD and ILD can be represented by direction angles, it is possible to compare sound source direction perception between listeners (Pulkki, 1999, p. W99-2). According to Moore (2013, p. 281), binaural processing can diminish the perception of unwanted echoes and improve localization of sound sources in reverberant enclosed spaces, and allow the hearing system to detect and analyse sounds in noisy environments.

Musical works, such as ”Bye Butterfly” (1965), ”The Bath” (1966), and ”I of IV” (1966), composed by the American composer and accordionist Pauline Oliveros (born in May 30, 1932)[57], or ”Vanishing Point” (1989), composed by the American musician, scientist, and director of the Stanford University Center for Computer Research in Music and Acoustics (CCRMA) (founded by John M. Chowning - see section 2.2.4), Christopher

[56] Retrieved 18/06/2016, from `http://www.mcnabb.com/music/`
[57] Retrieved 19/06/2016, from `https://en.wikipedia.org/wiki/Pauline_Oliveros`

David Chafe (born in Switzerland in December 5, 1952)[58], best known as Chris Chafe, are examples where differences of time and differences of level are explored (Zelli, 2009).

2.5.4.1.1 Inter-aural Time Differences (ITD)

In sound source localization, the meaning of the word *inter-aural* is *difference or relationship between sounds at both ears* (Blauert, 1997, p. 93), as already mentioned in section 2.5.4.1. Thus, in sound source localization in lateral positions, different arrival times of a sound to both ears, due to the difference between the paths that sound has to travel in order to reach both ears, lead to an arrival time difference between both ears, which is usually called Inter-aural Time Difference (ITD). In this situation, a delay or a time shift of the entire sound, or of parts of it, is perceived at one ear relatively to the other (Blauert, 1997, p. 141), so that lateral displacements of a perceptual sound can be detected closer to the ear at which the wavefront arrives first (see section 2.5.5).

In sinusoidal sounds, the periodicity can lead to an ambiguity with respect to the ITD, that is, there can be two different ITDs, Δt_1 and Δt_2, as shown in figure 2.45.

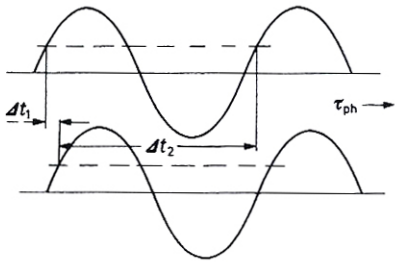


Figure 2.45: Ambiguous ITD (Blauert, 1997, p. 147).

In this case, the perceptual sound closer to the median plane, which corresponds to the shortest inter-aural time difference Δt_1, is dominant (Blauert, 1997, p. 147). Since the ITD Δt_1 is shorter than a period, the ITD represents a phase difference which corresponds to a single direction (Cheng & Wakefield, 2001, p. 232; Stern et al., 2005, p. 3). The main cue for low frequencies appears therefore to be the phase difference (Castellanos, 2006, p. 14). If the sound reaches both ears at the same time, then the perceptual sound appears in the median plane (see section 2.3.2).

In figure 2.46, the head is deemed to have spherical form for simplification reasons, assuming that the ears are positioned across a diameter, although it is known that the head is rather ellipsoidal and the ears are positioned behind and below the centre of the head (Duda et al., 1999, p. 965) (cf. sections 2.4 and 2.4.2).

The maximum difference between the paths that sound has to travel in order to reach both ears is actually equal to about 21.5 centimetres, although it varies from person to person essentially due to the different head sizes. Thus, considering a radius of the head $r = 0.0836\ \mathrm{m}$ and an incidence angle of sound of $\frac{\pi}{2}$ radians or 90º in figure 2.46, this maximum difference is the result of (Howard & Angus, 2001, pp. 98–99; C. H. Brown & May, 2010, pp. 128–129):

$$\Delta d + y = r \times \theta + r \times \sin\theta = r \times (\theta + \sin\theta) = 0.0836 \times (\frac{\pi}{2} + \sin\frac{\pi}{2}) \approx 0.215\ \mathrm{m} = 21.5\ \mathrm{cm}\ .$$

[58] Retrieved 19/06/2016, from `https://ccrma.stanford.edu/~cc/pub/pdf/Rescc15.pdf`

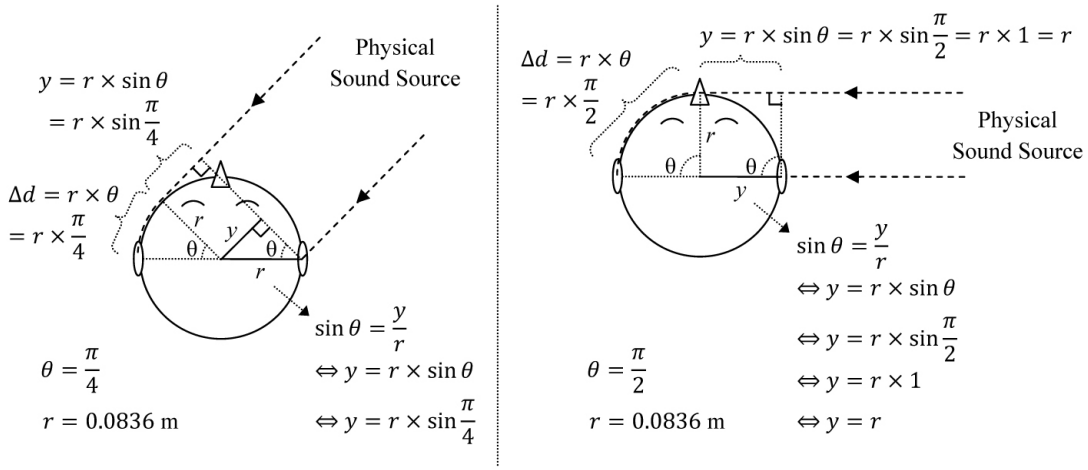


Figure 2.46: A simple spherical model of the head seen from above, assuming that the ears are positioned across a diameter: to the left, the incidence angle of sound is equal to $\frac{\pi}{4}$ radians or 45º; to the right, the incidence angle of sound is equal to $\frac{\pi}{2}$ radians or 90º.

This means that a sound reaching both ears from a side can arrive to the farthest ear maximally about 630 microseconds after it arrives at the nearest ear (Blauert, 1997, p. 143; Howard & Angus, 2001, p. 99; Stern et al., 2005, p. 3; Thigpen, 2009), at a temperature of about 15 ºC (see section 2.1.3.1):

$$c_{\text{air}} = \frac{x}{t} \Leftrightarrow t = \frac{x}{c_{\text{air}}} \Leftrightarrow t = \frac{\Delta d + y}{c_{\text{air}}} \Leftrightarrow t \approx \frac{0.215 \text{ m}}{331.3 + 0.606 \times 15 \text{ ºC}} \Leftrightarrow t \approx 632 \times 10^{-6} \text{ s} = 632 \ \mu\text{s} \approx 630 \ \mu\text{s} \ ,$$

where c_{air} is the sound propagation speed in the air, and x is the distance travelled by sound in time t.

Since a single perceptual sound is most accurately localized when a sound reaches both ears in phase (Blauert, 1997, p. 145), when the sound reaches both ears in an out-of-phase condition of approximately 180º, which corresponds to an inter-aural phase shift of about half a period or $\frac{T}{2}$, listeners clearly perceive two perceptual sounds, one to the left and one to the right (Blauert, 1997, p. 148).

When the difference between the paths that a sound has to travel in order to reach both ears is shorter than half a wavelength or the respective ITD Δt_1 is less than half a period or $\frac{T}{2}$, then both ears are first simultaneously stimulated with a compression or a rarefaction (see figure 2.47), which is a condition needed for a listener to localize a single perceptual sound in a lateral position in a more precisely way without ambiguities (Stern et al., 2005, p. 3). In addition, the lateral displacement of a perceptual sound depends almost linearly on the ITD before its maximum is reached (Blauert, 1997, p. 144).

If the ITD Δt_1 is longer than half a period or the phase angle ϕ is greater than π radians, then two perceptual sounds, one to the left and one to the right, are possible (Howard & Angus, 2001, p. 100). Furthermore, if frequencies above more or less 1.5 kHz are involved, that is, if the ITD Δt_1 is longer than a period of about 630 microseconds, where the difference between the paths that a sound has to travel in order to reach both ears is longer than one wavelength, then the ITD cues for lateral localization lead to phase differences that correspond to more than a single spatial location, because aliasing problems occur (Cheng & Wakefield, 2001, pp. 231–232).

As a result, the inter-aural delay of about 630 microseconds corresponds to the full lateral displacement of a perceptual sound, which, however, occurs only for sounds whose half-period is longer than about 630 μs

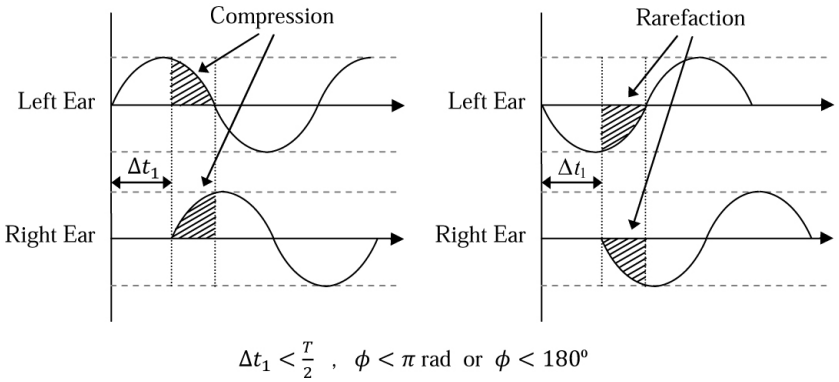


Figure 2.47: ITD $\Delta t_1 < \frac{T}{2}$, where the phase angle $\phi < \pi \text{ rad}$ or $\phi < 180^{o}$.

(Blauert, 1997, p. 148; Stern et al., 2005, p. 3):

$$\frac{T}{2} > 630 \times 10^{-6} \text{ s} \Leftrightarrow T > 1260 \times 10^{-6} \text{ s} \Leftrightarrow T > 1.26 \text{ ms} ,$$

so that their frequency f is lower than about:

$$f < \frac{1}{T} \Leftrightarrow f < \frac{1}{1260 \times 10^{-6} \text{ s}} \Leftrightarrow f < 793.65 \text{ Hz} \approx 800 \text{ Hz} .$$

Above approximately 800 Hz, the maximum lateral displacement of a perceptual sound becomes increasingly smaller, because neurons between the inner ear and the central nervous system are not ready for a second trigger for about 1 to 2 milliseconds after having been triggered once (Blauert, 1997, pp. 148–149), which corresponds to a recovery time of neurons, known as refractory period, that is longer than periods of sounds above about 800 Hz. Above around 1.5 kHz to 1.6 kHz there is usually no perceptible lateral displacement any more, so that above that frequency region it is not possible to perceive binaural beats, that is, a listener cannot perceive the oscillation of a perceptual sound inside the head between both ears at a frequency equal to the difference between two very close frequencies presented separately to both ears (see section 2.1.3.2).

With noise and broadband signals, lateral displacements of perceptual sounds can be easily perceived below and above around 1.6 kHz. Below about 1.6 kHz, the hearing system takes usually the ITDs of the fine structures (see figures 2.5 and 2.6 in section 2.1.3.1) into consideration (Blauert, 1997, p. 173). Above approximately 1.6 kHz, ITDs of low-frequency envelopes of high-frequency stimuli (Stern et al., 2005, p. 5) (see figures 2.5 and 2.6 in section 2.1.3.1) are normally evaluated (Blauert, 1997, pp. 150–151, 173; Pulkki, 2001b, p. 7; Stern et al., 2005, p. 5).

Blauert (1997, p. 152) defines lateralization blur of ITDs as the smallest change in the inter-aural phase delay that leads to a lateral displacement of a perceptual sound. In this sense, lateralization blur of ITDs decreases with the increase of the sound level or duration (Blauert, 1997, p. 155). For low-frequency sinusoidal sounds, the Just Noticeable Difference (JND) (see section 2.1.3.2) for ITDs is of the order of 10 microseconds (Stern et al., 2005, p. 5). It depends on the ITD, on the Inter-aural Level Difference (ILD) (see section 2.5.4.1.2), and on the frequency of sound.

Summing up, the ITD is commonly useful and effective for the localization of sounds comprised of frequencies below about 1.5 kHz (Thigpen, 2009), where the size of the head is smaller than the wavelengths of sounds, so that the waves diffract around the head as if it were not there.

2.5.4.1.2 Inter-aural Level Differences (ILD)

In sound source localization, the meaning of the word *inter-aural* is *difference or relationship between sounds at both ears* (Blauert, 1997, p. 93), as already mentioned in sections 2.5.4.1 and 2.5.4.1.1. Thus, if a single physical sound reaches both ears at the same time but with different sound pressure levels, then the resulting perceptual sound moves towards the ear which is presented with the highest sound pressure level (Blauert, 1997, p. 155), that is, lateral displacements of the perceptual sound can be perceived in this way (cf. section 2.3.4.3). If the sound pressure level of the physical sound is equal at both ears, then the perceptual sound appears in the median plane (see section 2.3.2). The difference between sound pressure levels at both ears is therefore usually called Inter-aural Level Difference (ILD), although this phenomenon is also known as Inter-aural Intensity Difference (IID) or Inter-aural Amplitude Difference (IAD) when intensity or amplitude differences are respectively considered, as referred to in section 2.5 as well.

In a free sound field, the ILD depends on frequency (Blauert, 1997, p. 157) and horizontal angle of a sound source, being most effective for physical sounds containing frequencies above approximately 1.5 kHz, which are perceived with different sound pressure levels by both ears as a result of the shadowing effect of the head for these sounds at one ear, where the wavelengths are smaller than the size of the head (Pulkki, 2001b, p. 7; Stern et al., 2005, pp. 2–3) (see also sections 2.4 and 2.5).

Blauert (1997, p. 160) defines lateralization blur of ILDs as the smallest change in the inter-aural sound pressure level difference that leads to a lateral displacement of a perceptual sound. In this sense, lateralization blur of ILDs increases as the perceptual sound becomes broader with increasing lateral displacement. The maximum lateral displacement of a perceptual sound due to ILD is difficult to measure, because its width and the corresponding lateralization blur increase for level differences of more than 8 to 10 dB, although it is known that the displacement depends almost linearly on the ILD before its maximum is reached (Blauert, 1997, p. 158).

According to Stern et al. (2005, p. 3), ILDs resulting from distant physical sound sources can reach values of the order of 25 dB at high frequencies. This difference can be even greater if the physical sound source is closer to one of the ears. The Just Noticeable Difference (JND) (see section 2.1.3.2) for ILDs is of the order of 1 dB for low-frequency sinusoidal sounds (Stern et al., 2005, p. 5).

Sensibility of the human hearing system decreases by a given quantity if it is stimulated for long periods of time, which causes adaptation and fatigue (Blauert, 1997, pp. 162–163) (cf. section 2.1.3.2). However, this depends on the type, level, and duration of the sound that is used. Adaptation means that sensibility is rapidly lost after a few seconds, reaching its maximum after about three to five minutes. The return to the original sensibility is normally achieved after one to two minutes. Fatigue is due to high level and long duration sounds. The return to the original sensibility takes usually a longer rest time. As a result, both adaptation and fatigue influence lateralization to some extent, so that the lateral displacement of a perceptual sound decreases towards the centre during sound presentation, because the sensibility of the ear that receives a greater sound

pressure level decreases relatively more than that of the other ear (Blauert, 1997, p. 163). Furthermore, lateralization of perceptual sounds associated with ILDs is a phenomenon that varies over time. Whereas short-term variations related with adaptation and fatigue can be observed, long-term variations associated with learning processes can be observed.

2.5.4.1.3 Interactions between Inter-aural Time and Level Differences

Both Inter-aural Time Difference (ITD) cues and Inter-aural Level Difference (ILD) cues work together in sound source localization (Blauert, 1997, p. 174), so that low-frequency ITD cues nearly suggest the same directions as high-frequency ILD cues (Pulkki & Karjalainen, 2001, p. 739). For sounds with frequencies below approximately 800 Hz, the size of the head is smaller than half the wavelength of sounds, so that the hearing system can determine the location of a perceptual sound based on the phase differences between the two ears without any confusion, as already explained in section 2.5.4.1.1. In this case, ITDs of the fine structures of sounds (see figures 2.5 and 2.6 in section 2.1.3.1) with frequencies below about 1.5 kHz at both ears are evaluated in a dominant way, since sound pressure levels at both ears are more or less identical (Odowichuk, 2012, p. 29). For sounds with frequencies above approximately 1.5 kHz, the size of the head is greater than the wavelength of sounds, so that the hearing system determines the location of a perceptual sound rather based on ILDs (see section 2.5.4.1.2) and, to a lesser extent, on ITDs of low-frequency envelopes of high-frequency stimuli. However, whereas ILDs generally influence the lateral position of any sound over the entire audible spectrum, ITDs affect the lateral position of a sound consisting of frequencies below about 1.5 kHz, or of a sound consisting of frequencies above about 1.5 kHz only if low-frequency envelopes are produced (Stern et al., 2005, p. 6) (see figure 2.48).

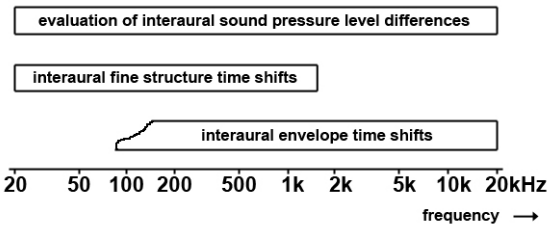


Figure 2.48: Evaluation of inter-aural sound differences (adapted from Blauert (1997, p. 164)).

The equivalence of a time difference to a certain level difference or the equivalence of a level difference to a specific time difference, in order to obtain the same lateral displacement of a perceptual sound, can be observed (Howard & Angus, 2001, p. 104) and it depends on loudness, essentially for sounds with frequencies above approximately 1.6 kHz (Blauert, 1997, p. 165). The trading between time difference and level difference is most effective for inter-aural time differences shorter than about 630 $\mu\mathrm{s}$ (Howard & Angus, 2001, p. 104). This relationship, expressed in $\mu\mathrm{s/dB}$, is therefore called *compensation factor* or *trading ratio* and is not linear (Blauert, 1997, pp. 165–166). This means that a greater sound pressure level difference is necessary to compensate for a certain time difference when sounds have greater sound pressure levels. As a result, ILDs are most significant when sounds have a lower level and frequencies above about 1.6 kHz are present. In this

case, the displacement of a perceptual sound is determined by inter-aural envelope shifts and ILDs, and the compensation factor varies between around 200 and 70 $\mu\mathrm{s}/\mathrm{dB}$, decreasing with increasing loudness (Blauert, 1997, p. 172). On the other hand, ITDs dominate for sounds with frequencies below about 1.6 kHz (Blauert, 1997, p. 170). For inter-aural time differences between approximately 630 $\mu\mathrm{s}$ and 30 milliseconds, where the perceived direction of a perceptual sound is usually not changed by the small inter-aural level differences, the sound seems to come from the source which arrives first (see section 2.5.5), except if the delayed sound has a level of more than about 12 dB greater than the other one which reaches one of the ears first (Howard & Angus, 2001, p. 104). For inter-aural time differences greater than 30 milliseconds, the delayed sound is perceived as an echo.

Nevertheless, since no time and no level differences exist at both ears when a physical sound source is positioned in the median plane, cues other than ITD and ILD cues have to be used, so that a listener can differentiate between a sound source in front or behind him or her.

2.5.4.2 Monaural Cues

Monaural cues are comprised of time and level differences between the various spectral components of a sound at one ear only, described by its Head-Related Transfer Function (HRTF) (see sections 2.4.1 and 2.5.4), where inter-aural interactions are not taken into account (Blauert, 1997, p. 177; Pulkki, 2001b, p. 6; Audio Products Division of National Semiconductor Analog Products Group, n.d.; Calmes, 2013). The main functions of monaural cues are therefore to allow a listener to distinguish between a sound source in front or behind him or her, and to define its elevation and its distance (Blauert, 1997, pp. 177, 304; Pulkki, 2001b, p. 6; Zelli, 2009).

2.5.4.3 Visual Cues

Sound source localization can be affected by visual cues, which means that the perception of the position of a physical sound source can be improved by visual cues together with the auditory cues (Moore, 2013, p. 281), or simply that the perception of the position of a physical sound source can be changed by what the listener sees and where he or she sees it while sound is presented (Blauert, 1997, pp. 193–194, 196).

Whereas visual cues were found to influence auditory movement perception during gesture control of spatialization (Marentakis et al., 2008; Marentakis & McAdams, 2013, p. 1) (cf. section 2.6.3), auditory cues were found to improve or complement the visual ones (de Götzen, 2004, p. 9).

Odowichuk (2012, p. 1) considers that one of the most important factors that leads an audience to regard an audio-visual performance as exciting is the connection that exists between the senses of hearing and vision. In this sense, Tsay (2013, p. 1) found out that judgements made about music performance are fundamentally affected by visual information, although subjects usually believe that sound is the most important factor in that task.

2.5.4.4 Other Cues

The position of a sound source can be judged in a better way by a listener if the head is free to move (Blauert, 1997, pp. 178, 180; Odowichuk, 2012, p. 30) and if the sound source can be seen (see section 2.5.4.3).

According to Blauert (1997, pp. 178–179), head movements associated with sound source localization can be of two types: 1) A more or less unconscious and spontaneous or reflexive movement towards the direction of the perceptual sound source that corresponds to the most probable position of the physical sound source, so that the localization blur decreases to its minimum; and 2) More or less conscious searching and orienting movements in order to determine the definite position of a perceptual sound source while it is still being located. As a result, these movements cause changes in the monaural and binaural attributes.

Furthermore, the position, the direction, and the amplitude of rotation of the head, which can be defined by the vestibule and the three semicircular canals (see section 2.4), by the sense of vision, and by the receptors of position, tension, and posture in the neck muscles, are useful to determine the position of a sound source (Blauert, 1997, pp. 181, 197–198).

In 1967, it was shown by W. R. Thurlow, J. W. Mangels, and P. S. Runge that the most used movements in sound source localization are those of rotation to the left and to the right in the horizontal plane around the vertical intersection axis between the median and frontal planes, and of upwards and downwards rotation in the median plane around the horizontal intersection axis between the horizontal and frontal planes (Blauert, 1997, pp. 181–182, 185) (see section 2.3.2). The former type of rotation is commonly used in order to determine if sound sources are in the frontal or rear hemisphere. When a listener needs to determine if sound sources are in the upper or lower hemisphere, upwards and downwards rotation in the frontal plane around the horizontal intersection axis between the horizontal and median planes can be further used. In addition, long or repeated head movements are normally used at the beginning if the duration of sound is long enough.

Sound can also be perceived by the sense of touch if high sound pressure levels and low frequencies are produced. In this case, the location of the perceived sound is inside the body and not at the position of the perceptual sound (Blauert, 1997, p. 200).

Summing up, the movement of the head improves sound source localization, so that direction differences between physical and respective perceptual sound sources are minimized (Blauert, 1997, pp. 190–191).

2.5.5 Precedence Effect

The term *precedence effect* was first proposed by the German-American experimental psychologist Hans Wallach (November 28, 1904 - February 5, 1998)[59], by Edwin B. Newman[60], and by the American research psychologist Mark Richard Rosenzweig (September 12, 1922 - July 20, 2009)[61], in 1949, to describe the phenomenon in which listeners localize physical sound sources based on sound that arrives first at the ears directly from the sound source (direct sound - see section 2.1.3.1), rather than from reflected sound (A. D. Brown et al., 2015, p. 1). They showed that two identical brief click sounds radiated by two loudspeakers, at symmetric positions and equidistant from a listener, are perceived as a single fused sound if one is produced approximately 1 to 5 milliseconds after the other.

However, the term *precedence effect* has been also widely used to describe other auditory phenomena that are related with the perception and localization of sound sources as well, which occur between a direct

[59] Retrieved 06/12/2013, from http://en.wikipedia.org/wiki/Hans_Wallach
[60] Retrieved 10/07/2016, from https://www.jstor.org/stable/1418275?seq=1#page_scan_tab_contents
[61] Retrieved 10/07/2016, from https://en.wikipedia.org/wiki/Mark_Rosenzweig

sound and a reflection, where the spatial information carried by the sound that reaches the ears first is usually dominant (A. D. Brown et al., 2015, p. 1). Terms such as *fusion*, *localization dominance*, and *lag discrimination suppression* were therefore suggested by Litovsky et al. (1999, p. 1634), as an attempt to categorize all these phenomena. Whereas *fusion* is a temporal aspect of the precedence effect, *localization dominance* or *law of the first wavefront* and *discrimination suppression* are spatial aspects of the precedence effect (A. D. Brown et al., 2015, p. 3).

If the delay between two sounds radiated by two loudspeakers at symmetric positions and equidistant from a listener, one simulating a direct sound and the other a reflection of it, is relatively short, depending mainly on the type of sound and acoustics of the surrounding space, a listener tends to: 1) perceive one single fused sound (*fusion*), rather than the two original ones; 2) localize the fused sound at, or close to, the location stimulated by the first of the two sounds, that is, by the direct sound (Stern et al., 2005, p. 7; Thigpen, 2009) (*localization dominance* or *law of the first wavefront*); and 3) essentially lose the perception of the localization of the second sound, that is, the reflected sound (A. D. Brown et al., 2015, p. 2) (*discrimination suppression*). The suppression of the natural reflections of the direct sound by the surrounding environment is therefore what allows a listener to also localize sound sources in reverberant spaces (Pulkki, 2001b, p. 8).

Thus, when the delay between two identical sounds, one considered as the direct sound and the other as its simulated reflection, radiated by two loudspeakers at symmetric positions and equidistant from a listener, has a value between 0 and about 1 millisecond, a single perceptually fused sound, rather than the two original ones, is perceived at an average location between both directions, exactly in the median plane if the delay is null, or closest to the lead loudspeaker if the delay is greater than zero to approximately 1 millisecond, which corresponds to the position of a perceptual sound source (see section 2.2.2), rather than that of a physical sound source (cf. section 2.2.1). This phenomenon was identified by Hans Warncke in 1941 and defined as *summing localization* (Blauert, 1997, pp. 203–204; Litovsky et al., 1999, p. 1634; A. D. Brown et al., 2015, p. 2) (see also sections 2.2.2 and 2.3.4).

When the delay between both sounds has a value between approximately 1 millisecond and the value at which the fused sound perceptually splits into two, that is, at the listener's echo threshold of about 30 milliseconds (Howard & Angus, 2001, p. 105), depending on the type and level of the direct sound, and on the acoustic behaviour of the surrounding environment, known as echo threshold (Blauert, 1997, p. 229), a fused sound is perceived closest to the loudspeaker which produces sound first (Litovsky et al., 1999, p. 1634; A. D. Brown et al., 2015, p. 2). In fact, if the direct sound level is increased, the echo threshold has shorter times, and it seems to have longer times for ongoing sounds than for single impulses (Blauert, 1997, p. 230). In addition, if the level of the reflected sound is increased relatively to that of the direct sound, the reflected sound appears at a shorter delay time, and vice-versa. This phenomenon was called by the German electrical engineer and acoustician Lothar Cremer in 1948 the *law of the first wavefront* (see also section 2.2.2).

Furthermore, in 1951, the Dutch scientist Helmut Haas (Rumsey, 2001, p. 28) found out that the perceived level and the width or spaciousness of a sound source can be increased if a reflection of a sound arrives at the ears more than 1 millisecond after the direct sound and before the echo threshold of about 30 milliseconds (Howard & Angus, 2001, p. 105), even if the level of the delayed sound is substantially higher than the first arriving sound (Blauert, 1997, p. 226), a phenomenon known as *Haas Effect* (see also section 2.2.2). As

the delay is increased, the perceived timbre of the perceptual sound source changes and its width increases correspondingly (Blauert, 1997, p. 224). In this respect, Gibson (1997, p. 15) adds that this effect can be used in an elementary stereophonic reproduction system (see section 2.3.4.3) in order to increase the perceived width or spaciousness of a sound source in a mix if its delayed version is produced within the above-mentioned delay range, calling it *fattening* of a sound.

In 1961, H. P. Seraphim defined the threshold of perception for reflections as the level difference between the direct sound and the reflection at which a listener perceives the reflection as perfectly audible (Blauert, 1997, p. 223). It is most commonly known as *masked threshold*, because the threshold of audibility of one perceptual sound, usually the reflected sound, is increased by the other one, usually the direct sound, called *masker* (Howard & Angus, 2001, p. 231) (see section 2.1.3.2).

If the delay is equal to, or longer than, the echo threshold, a second sound (echo) is perceived, initially closest to the loudspeaker that produces sound first, and at longer delays nearest to the other loudspeaker (Litovsky et al., 1999, p. 1634).

However, if the time difference between two ping-ponging sounds produced by two loudspeakers, at symmetric positions and equidistant from a listener, is in the range of about 5 to 50 milliseconds, a single sound is perceived at a location of the sound that arrives at the ears first (Thigpen, 2009) (see section 2.2.4).

The precedence effect is normally more robust if sounds with longer duration, such as speech and music, are used (A. D. Brown et al., 2015, p. 7). Furthermore, the precedence effect is more robust if sounds are displaced laterally by Inter-aural Time Differences (ITDs) (cf. section 2.5.4.1.1), rather than by Inter-aural Level Differences (ILDs) (see section 2.5.4.1.2) (A. D. Brown et al., 2015, p. 9).

Summing up, the precedence effect makes use of the direct sound, which remains the same independently of the surrounding space, in order to allow a listener to also localize sound sources in reverberant spaces. The precedence effect seems to be ruled by low-frequency ITD cues and is most robust if sounds of longer duration and larger bandwidth are used (A. D. Brown et al., 2015, p. 13).

2.5.6 Cocktail Party Effect

The human ear system has the ability to filter or extract specific information from a complex sound, such that for example a particular conversation among multiple conversations taking place simultaneously with the same level in a reverberant environment (Henrique, 2007, p. 874) (see section 2.3.2) can be attended to, as already mentioned in section 2.1.3.2. The interfering sound a listener is not intended to focus on can be perceived with a lower loudness than that of the sound which is his or her target (Pedroso de Lima, 2012, p. 17). This process is known as the *cocktail party effect*.

2.6 Gesture

As pointed out by Cadoz & Wanderley (2000, p. 71), Yoshioka (2005, p. 21), Nehaniv (2005), and Godøy & Leman (2010, p. x), there are numerous definitions of the term *gesture* within the human-human interaction (HHI), human-computer interaction (HCI), and human-robot interaction (HRI) domains, as well as within the

musical and other artistic domains, although none of them is more accurate or complete than the others. Many of these definitions are related to movement, motion, or action, as an expression of feeling, emotion, sentiment, passion, idea, attitude, intention, opinion, information, or meaning. For instance, the Collins Cobuild English Language Dictionary (1993, p. 606) defines the term *gesture* as "... a movement that you make with a part of your body, especially your hands or your head, to express emotion or information, either instead of speaking or while you are speaking." Thus, gesture is commonly related to the movements of hands or arms, or to a part of the body, which represent a non-verbal form of communication (Rodrigues, 2007, p. 113; Schneider, 2010, pp. 76, 82) that can be used in combination with, or instead of, verbal communication (Bhuiyan & Picking, 2009), to express an idea or meaning (Leman & Godøy, 2010, p. 5; Schneider, 2010, p. 71). In this regard, Jensenius et al. (2010, p. 30) refer that most authors appear to accept that gestures have to do with both body movement and meaning.

According to the views of the English experimental psychologist Adam Kendon (born in 1934)[62] and the American psychologist and writer David McNeill (born in 1933)[63], it seems that gesture and speech are linked together as belonging to the same communication system (Rodrigues, 2007, p. 125), an idea that appears to be nowadays strongly supported with the advent of brain imaging techniques (Godøy, 2010, p. 108). McNeill considers that the connection between them is made through meaning, time, function, development, and dissolution (Rodrigues, 2007, p. 126), but that gesture and speech do not have to co-occur (Jensenius et al., 2010, p. 15). This relationship is substantiated by means of Kendon's sequence of conventionality values, which McNeill designates *Kendon's continuum*, in which the obligatory presence of speech relatively to the type of gesture decreases from the first to the last item in the sequence, while the presence of linguistic properties of gestures increases accordingly (Mulder, 1996, p. 6; Yoshioka, 2005, p. 22; Rodrigues, 2007, p. 134; Jensenius et al., 2010, p. 15; McNeill, 2011, p. 344): 1) *Gesticulation* is a type of gesture that is only made during speech, normally by the hands and arms, so that the presence of speech is obligatory; 2) *Language-like* or *speech-linked* gestures are similar to gesticulation, but now they grammatically form part of speech in a sequential rather than concurrent way, replacing it; 3) *Emblems* are culturally conventionalized well-formed gestures, in which the presence of speech is optional; 4) *Pantomimes* refer to the use of gestures that depict objects and actions, which communicate meanings, in which speech is by definition obligatorily absent (see also McNeill, 2000, p. 2; Schneider, 2010, p. 71); and 5) *Signed languages* are sets of gestural signs that form complete linguistic communication systems, so that speech is obligatorily absent.

Nevertheless, in all of these definitions of gesture a more or less direct or indirect reference to human physical or motor behaviour is made, where movement is an important part of gesture. Therefore, muscular activity can be taken into consideration and the term *motor unit* can be introduced as its basic functional element (Cadoz & Wanderley, 2000, p. 75). Consequently, the type of movement that humans do can be analysed, that is, it can be studied if they do a slow or sustained movement, or a fast, short, or ballistic one. The origin of the movement (passive or active) being performed and the nature of it (reflex, automatic, or intentional) are further characteristics that can also be investigated. Movements can therefore be objectively measured (Jensenius et al., 2010, p. 19).

[62] Retrieved 05/08/2016, from `https://web.archive.org/web/20070629060355/http://www.semioticon.com/semiotix/semiotix9/sem-9-03.html`

[63] Retrieved 05/08/2016, from `https://en.wikipedia.org/wiki/David_McNeill`

On the other hand, whereas movement has to do essentially with physical displacement of a body part or an object, gesture relates to other concepts beyond movement, so that it can be considered as both a bodily and a mental process, "... in which humans interact with their environment from the perspective of embodied cognition" (Leman & Godøy, 2010, p. 8). As Rodrigues (2007, p. 125) states, the relationship between thought and gesture was first systematically studied by McNeill. Based on his cognitive point of view (Rodrigues, 2007, pp. 115–116), gesture can be approached from a communicational, controlling, and metaphoric perspective (Jensenius et al., 2010, p. 14), although many other perspectives, which can be found in Yoshioka (2005), Rodrigues (2007), and Godøy & Leman (2010), are possible. It should be stressed here that our goal is not to approach every typology of gesture, since there are many of them because "... a gestural typology, like all other typologies, is not intended to create an absolute classification system..." (Jensenius et al., 2010, p. 25), but rather focus on the most recent and relevant ones for our purpose.

Relatively to the first standpoint (gesture as communication), the typological classification of gestures proposed by McNeill in 1992, involving the representational and functional aspects of some kind of physical body movement closely connected with speech, can be considered (Yoshioka, 2005, pp. 25–28; Rodrigues, 2007, pp. 129–133; Jensenius et al., 2010, p. 14): 1) *Iconic* or *concrete* gestures are those whose shapes describe real objects or real-world actions, such as a knocking-at-the-door imitating movement; 2) *Metaphoric* gestures resemble the iconic gestures, but their shapes represent an abstraction or an abstract idea of the mental world; 3) *Beats* are usually fast, short, and two-phase horizontal or vertical gestures, or flicks, which, along with speech, highlight specific moments of it, or, in a narrative, point out important information; 4) *Deictic* gestures are those which are used to indicate a point, place, object, person, direction, or event in the real or in an imaginary or virtual world, or where a referent in discourse is missing, usually with the stretched index finger, although other parts of the body or objects can be used; 5) *Cohesive* gestures or catchments are repetitive gestures that are used to link thematically-related parts of speech that are temporarily separated; and 6) *Emblems* are culturally conventionalized gestures, such as the thumbs-up 'OK' gesture, but are not completely organized as a language. Since gestures can perform more than one function (poly-functional) (Rodrigues, 2007, pp. 124, 140, 142) or concentrate several meanings (polysemic) (Rodrigues, 2007, p. 126, 142) at a time (Jensenius et al., 2010, p. 30), all these types of gestures can be used individually or co-occur in different ways with each other in social interaction situations. Furthermore, these gestures are mainly empty-handed movements, also known as free, naked, semaphoric, or even semiotic gestures (Jensenius et al., 2010, p. 17; Odowichuk, 2012, p. 8), because they only convey meaningful information without making use of any objects or physical tools (Coutaz & Crowley, 1995, p. 1) (see also section 2.6.1).

The space where a gesture preferably occurs is divided by McNeill into four areas (McNeill, 1992, pp. 88–91; Yoshioka, 2005, p. 29; Rodrigues, 2007, p. 133): 1) Centre-centre; 2) Centre; 3) Periphery; and 4) Extreme periphery (see figure 2.49). Iconic gestures are usually performed in the centre-centre area, the metaphoric ones in the lower centre, the beats in the lower centre and periphery, and the deictic gestures in the periphery.

From the point of view of control, gesture can be used as an input to computer and interactive systems in the HCI domain. In this context, gesture can be mainly manipulative, depending therefore on physical contact, also called haptic[64], instrumental, or ergotic contact, because it "... is associated with the notion of work" and

[64] A survey about haptics can be found in `http://www.ijcsi.org/papers/IJCSI-9-5-3-234-244.pdf`, visited on 28/08/2016.

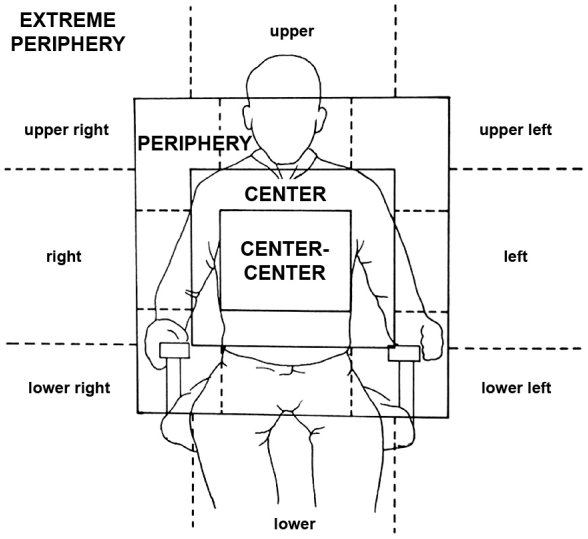


Figure 2.49: McNeill's gesture space (adapted from McNeill, 1992, p. 89).

"... the capacity of humans to manipulate the real world, to create artefacts, or to change the state of the environment by 'direct manipulation'" (Coutaz & Crowley, 1995, p. 1). However, nowadays input to computer and interactive systems can also be accomplished by empty-handed gestures. In any case, computers can be used in order to understand the meaning of human gesture by extracting expressive characteristics from human body movement (Choi, 2000, p. 140; Jensenius et al., 2010, p. 16) (cf. section 2.7).

With respect to the third perspective (gesture as metaphor), a gesture can be seen as a mental entity, image, or shape that is suggested or conveyed by something, but that is not directly associated with any physical movement, such as a gesture evoked from musical sounds (Jensenius et al., 2010, p. 18). In this case, this kind of gesture seems to be similar to a physical gesture, but at a mental level.

In the HRI domain, an attempt of a basic classification of types of gestures is proposed by Nehaniv (2005) and Nehaniv et al. (2005), in order to identify human intent in this context and to understand the movement of the human body: 1) *'Irrelevant'/Manipulative* gestures are those that affect the environment or that relate to it, such as the movement of hands and arms while walking; 2) *Side effect of expressive behaviour* is observed during communication with others, in which hands, arms, and face move without any particular function; 3) *Symbolic* gestures are conventionalized gestures in communicative interaction that are equivalent to the above *emblems* proposed by McNeill; 4) *Interactional* gestures are used to regulate communicative interaction, but without conveying any of its contents (equivalent to *regulators* in Paul Ekman's and Wallace Friesen's typology of gestures, which will not be presented here, because it is beyond the scope of this thesis); and 5) *Referential/Pointing* gestures are those which are used to indicate a point, place, object, person, direction, or event in the real or in an imaginary or virtual world, or where a referent in discourse is missing, usually with the stretched index finger, although other parts of the body or objects can be used (equivalent to deictic gestures proposed by McNeill, who uses exactly the same terminology).

In the musical domain, some gestures are learned body movements that musicians make in order to play musical instruments, called sound-producing movements, musical gestures (Leman & Godøy, 2010, pp. 3, 6), or excitatory movements (Jensenius et al., 2010, p. 22), or that a conductor uses to "... coordinate actions

among musicians...” (Leman & Godøy, 2010, p. 5), which ”... relate directly to the musical structure as well as to its reproduction by musicians in a performance...” (Schneider, 2010, p. 72), called conducting gestures. In a broader context, the term *musical gesture* is defined by de Götzen (2004, p. 5) and Jensenius et al. (2010, pp. 13, 19) as body movement that is made by performers (musicians, dancers, etc.) in order to produce sound or music, as already mentioned above, and as a response to the perceived sounds in a continuous feedback loop, or by listeners and dancers as a response to the perceived sounds only. In this regard, Choi (2000, p. 149) states that ”... musical gestures have two components, an auditory sequence and a performer’s motion sequence.”

Similarly to what has been said at the beginning of this section about gesture relatively to verbal communication, gesture can be used in combination with musical communication as well (Schneider, 2010, p. 71), so that expressiveness is also connected with musical gestures. According to Leman & Godøy (2010, p. 3), musical gestures or music-related gestures are a means that links music with movement, thus ”... gestures are intrinsic to music” (Schneider, 2010, p. 73). Since the vestibule and the three semicircular canals (see section 2.4) are responsible for the sense of movement, and movement over time and space can be described by trajectories, music can in turn cause sensations of movement in the ear through melodic lines which describe melodic movement or through musical rhythm (Guedes, 2005, p. 12), that is, sensations of movement can be evoked by music (Schneider, 2010, p. 94; Godøy, 2010, pp. 103, 104). For instance, a musical phrase can be regarded as ”... a gesture which can express an emotion using only musical parameters, where the music is the abstract space” (de Götzen, 2004, p. 6).

Nevertheless, according to Godøy (2010, pp. 104–106), ”... listeners [can also] use images of sound-related movement in making sense of what they hear. Thus, (...) sound [can induce sensations or] images of movement, and conversely, (...) previously learned images of sound-related movement [can be] projected onto sound, (...). [This] means that music perception is (...) closely linked with bodily experience (...), and that music [is perceived in a supplementary way] with the help of both visual/kinematic images and effort/dynamics sensations.” For instance, the movement of a soloist or conductor is therefore ”... guided by auditory perception and gestural principles” (Choi, 2000, p. 146).

2.6.1 Basic Types of Gesture and Movement

Motion, as an action, gesture, or movement, ”is the process of continually changing position or moving from one place to another” (Collins Cobuild English Language Dictionary, 1993, p. 940). Thus, it is dynamic by nature (Choi, 2000, p. 142). In addition, the internal state of a subject is frequently changed under the influence of some sort of incoming signal. As a result, this change can either trigger a detectable movement as a response or emotion that affects the environment or an imperceptible movement that does not modify the environment at all.

In any case, movement can be regarded as slow or sustained, or as fast, short, or ballistic, as already referred to in section 2.6. The speed of movement becomes usually slower as the curvature of its trajectory gets bigger (Cadoz & Wanderley, 2000, p. 77). Furthermore, the relation of speed of a gesture to the length of its trajectory is commonly more or less maintained, independently of the scale in which it is performed.

Although a distinction between posture and gesture can be made and information can be communicated by both individually, they actually co-occur (Cadoz & Wanderley, 2000, p. 76). The former is usually associated with the way the whole human body is positioned (Rodrigues, 2007, p. 91), such as sitting or standing, for example. In the context of a single hand, a posture can be considered as an individual static position of that hand and a gesture as a dynamic sequence of postures which describes the hand movement (Cadoz & Wanderley, 2000, p. 72; Odowichuk, 2012, pp. 7–8). Nevertheless, when human body movement or gesture is subjected to analysis, it is common to divide it into smaller parts, in order to make it easier to be understood (Rodrigues, 2007, p. 126; Bhuiyan & Picking, 2009) (see section 2.6.4).

An *instrumental gesture* is defined by Cadoz & Wanderley (2000, p. 79) as a gesture that "... is applied to a material object...", where "... physical interaction [exists] with it", so that "... specific (physical) phenomena are produced, whose forms and dynamic evolution can be mastered by the subject." Furthermore, instrumental gestures perform three different but complementary functions (Cadoz & Wanderley, 2000, p. 78): 1) the *semiotic* function has to do with communication of information without using any objects or physical tools (cf. section 2.6); 2) the *ergotic* function involves the material manipulation of the environment, as has also been referred to in section 2.6; and 3) the *epistemic* function which "... allows humans to learn from the environment through tactile experience" (Coutaz & Crowley, 1995, p. 1). This kind of gesture is complementary to empty-handed gestures, which are not instrumental, but rather semiotic, since they only convey information without using any objects or physical tools.

In the context of Human-Computer Interaction (HCI), Choi (2000, p. 147) defines the term *Gestural Primitives* as "... fundamental human movements that relate the human subject to dynamic responses in an environment" (see also Choi, 2003, p. NIME03-202), and proposes therefore three types of device- and signal-independent primitives (Choi, 2000, p. 148), bearing in mind that they are performed by "... an observer having a chosen physical disposition to a movement sensor, with an intent to modify a dynamical process" (Choi, 2000, p. 147): 1) *Trajectory-based* primitives have to do with changes of orientation, where the trajectory of a movement can be performed based on a target, such as the movements that involve direct manipulation of objects (*pointing*, *grabbing*, *throwing towards*, or *putting there*), or where the trajectory of a movement is not dependent on a target, such as the *sweeping*, *twisting clockwise*, *waving*, and *bending further* movements; 2) *Force-based* primitives have to do with gradient movements, where linear changes can be detected by the fine sense of weight distribution (*leaning*, *pushing*, *pulling*, *squeezing*, *bending*, or *twisting*); and 3) *Pattern-based* primitives have to do with quasi-periodic movements, such as *walking* or *scratching*.

From the computational point of view, it is easier to deal with both trajectory-based and force-based primitives than with pattern-based primitives (Mustard, 2006, p. 9). In the former case, thresholds can be used to detect any changes, demanding little use of memory and resources. The latter involves recognition of patterns, which requires large amounts of data of learned patterns that have to be constantly analysed and compared with incoming data in order to verify if there is any match.

A survey study of different types of gestures used from 1980 until 2009 in research works on gesture-controlled technology for user interactions is presented in Bhuiyan & Picking (2009). The authors conclude that the hand gesture is the most used gesture in these systems, although other gestures, such as the head gesture and the gesture with voice are also found.

2.6.2 Gesture and Sound

Gesture and sound are frequently influenced by each other (Odowichuk, 2012, pp. 1, 6). Sometimes, gestures and motions affect sound, such as those that a musician uses to produce sounds in musical instruments. Conversely, in certain occasions, it is the sound that influences motion, as in the case of a dance performance, in which information is conveyed in only one direction.

Following what has been said about the mutual influence of sound on movement and of movement on sound in section 2.6, a musician, while producing sounds on a musical instrument, also responds to them through movements in the form of *sound-producing* gestures, necessary for producing sound again, or *sound-accompanying* gestures, not necessary to produce sound, but that are made along with sound, such as dancing or nodding the head (Godøy, 2010, p. 110), which can affect perception (Jensenius et al., 2010, p. 27). A dancer or performer, while reacting to sound with *sound-accompanying* gestures (Jensenius et al., 2010, p. 30), can also use gestures and motions in several human-computer interactive systems in order to affect sound, such as in 'Jeux de Modes' (version I)[65], 'Jeux de Modes' (version II)[66], and 'Seine hohle Form'[67], performed by the Palindrome Inter.media Performance Group, founded by the American dancer and choreographer Robert Wechsler[68]. As stated by de Götzen (2004, p. 6), "... the connection between music and body movement is explicit in dance: (...) the emotional states carried by the music are the same ones expressed by body movement". In either case, when the motion of a performer affects sound and vice-versa, "... a feedback loop with useful and expressive possibilities [results]" (Odowichuk, 2012, p. 6). de Götzen (2004, p. 5) and Jensenius et al. (2010, pp. 13, 19) name these movements *musical gestures* (cf. section 2.6). Furthermore, the auditory and visual senses are very closely linked together as well (see also section 2.5.4.3).

The motor synchronization with a given sound stimulus plays an important role in music, essentially when several musicians play together, and in dance. In this sense, Fraisse (1982, p. 154) states that "As a general rule, our reactions succeed the stimuli. In synchronization the response is produced at the same time as the appearance of the stimulus. A similar behavior is possible only if the motor command is anticipated in regard to the moment when the stimulus is produced. More precisely, the signal for the response is not the sound stimulus but the temporal interval between successive signals." Thus, synchronization between sounds and taps is best at time intervals between successive sounds in the range of 400 to 800 milliseconds, although intervals of 200 to 1800 milliseconds between consecutive sounds are possible for synchronization (Fraisse, 1982, p. 155). Nevertheless, a coincidence error between a tap and sound can be observed in synchronization. Its value depends on the body part that is used to perform the tap. Whereas a tap performed by the index finger usually anticipates sound by about 30 milliseconds, a tap performed by a foot anticipates sound by a greater value. "The difference between hand and foot permits us to think that the subject's criterion for synchronization is the coincidence of the auditory and of the tactile-kinaesthetic information [(see section 2.7)] at the cortical level. For this coincidence to be as precise as possible, the movement of tapping should slightly precede the sounding order to make allowance for the length of the transmission of peripheral information" (Fraisse, 1982,

[65] Retrieved 19/08/2016, from https://player.vimeo.com/video/113498257
[66] Retrieved 19/08/2016, from https://player.vimeo.com/video/126678981
[67] Retrieved 19/08/2016, from https://player.vimeo.com/video/150689124
[68] Retrieved 19/08/2016, from http://www.palindrome.de/; see also the topic 'Transforming Dance to Song' from Deutsche Welle's Euromaxx edition of May 15, 2012, available in http://tv-download.dw.de/Events/mp4/eme/eme20120514_taenzer_sd_dwdownload.mp4 and https://player.vimeo.com/video/176465488

p. 155).

Although a cause-effect relationship can be observed when human gestures are used to produce sounds, in computer music it is possible to separate them completely from each other, since "... computers can carry out all aspects of sound production, from composition up to interpretation and performance" (de Götzen, 2004, p. 7).

2.6.3 Gesture and Sound Spatialization

The control of sound spatialization with gesture has been experienced for the first time by Pierre Schaeffer and Pierre Henry in 1951 with the invention of Schaeffer's *potentiomètre d'espace*, as stated before in section 2.3.1. Other studies have been made in this respect, such as those presented below, just to mention a few.

Marshall et al. (2009) present the development of three gesture-controlled sound spatialization systems, based on three groups of controllable parameters: 1) sound source position and orientation; 2) sound source characteristics; and 3) environmental and room model parameters. In the first proposed system (*Spatial Performer*), sound sources can be moved in a virtual space with gesture in real-time by using, on the one hand, a Polhemus Liberty[69] magnetic position and orientation tracker of both hands, to directly position multiple sound sources in this space, and a pair of custom-built data gloves, to measure the posture of both hands (see section 2.6.1), which turned out to be both suitable for fine control in three dimensions. On the other hand, a Roland V-Drum MIDI drum kit is used to launch sound sources in the horizontal plane at different angles and velocities as a ballistic control system, and a weight-sensitive floor is used to steer a single sound source, both latter systems enabling a coarse control in two dimensions only. The second system (*Instrumental Performers*) makes use of existing musical performance gestures in order to indirectly manipulate parameters of sound sources. This is accomplished by a wireless accelerometer-based system which can be either worn by a performer or directly attached to an instrument. In the third proposed system (*Spatial Conductor*), room and environment model parameters can be directly controlled by several hand gestures, which are recognized by a data-glove system as a combination of hand postures and movements, so that the whole system sound and every sound within it can be affected.

An approach to the development of a gesture-controlled sound spatialization system for a performance set-up of a small ensemble is reported by Marshall et al. (2006), in which a Gesture Description Interchange Format (GDIF) for storing, retrieving, and sharing gesture-related information in a standardized way, already proposed by Jensenius et al. (2006) and still under development, is used. Based on Sound Description Interchange Format (SDIF), which was developed by the *Institut de Recherche et Coordination Acoustique/Musique* (IRCAM)[70] and the *Center for New Music and Audio Technologies* (CNMAT)[71] in the late 1990's (Jensenius et al., 2006, p. 176), in GDIF the movement-related data is divided into raw data (unprocessed data coming from sensing devices), body data (information about orientation and motion of the body, and limb motion in relation to the body, based on biomechanical properties and Rudolf von Laban's concepts - see section 2.6.4), and meta data (information about general motion qualities) (Marshall et al., 2006, p. 364). "The idea is to cre-

[69] `http://polhemus.com/motion-tracking/all-trackers/liberty`, visited on 30/11/2017.
[70] Retrieved 23/08/2016, from `http://www.ircam.fr/`
[71] Retrieved 23/08/2016, from `http://cnmat.berkeley.edu/`

ate a flexible setup where various sensing devices can communicate with different sound processing modules running in a large networked computer setup" (Marshall et al., 2006, p. 365).

Marentakis et al. (2008) investigate in turn the effects of the visibility of a performer's gestures on the identification of spatial sound trajectories in a concert hall. Four sound trajectories with identical start and end points are used, each trajectory starting at the back of the audience and ending at its front: a straight line across the middle of the hall, an arc and a wobbly arc to the left side of the hall (facing the stage), as well as a wobbly line swinging from the middle of the hall to the right and to the left. In this experimental study, the authors found out that the identification of spatial sound trajectories is improved when the shape of the performer's gestures, aligned and synchronized with that of the spatial sound trajectory, is visible to the audience. When there is no visual feedback of the performer's gestures to the audience, the identification of the spatial sound trajectories is made worse and degraded depending on the listening seat. In this situation, the wobbly line trajectory is the easiest to be identified and the wobbly arc is the worst. These latter results are consistent with what has already been mentioned in section 2.3, that is, that the human auditory system is relatively inefficient in processing spatial information and that the spatial audio systems, which are designed for the centre of the listening area, lead to the degradation of identification of the trajectories when the listener is not seated in the best listening position or sweet-spot (see section 2.3.4).

Marentakis & McAdams (2013) study the perceptual impact of gesture control of spatialization in the case of direct manipulation of auditory movement within the listening area in two different experiments, resorting to the use of four sound trajectories with identical start and end points, each trajectory starting at the back of the audience and ending at its front: a straight line across the middle of the audience, an arc and a wobbly arc to the right side of the audience (facing the stage), as well as a wobbly line swinging from the middle of the listening area to the left and to the right. The first experiment, performed in a concert hall, deals with "... the identification of spatial sound trajectories in the absence and presence of congruent visual cuing from the performer's gestures (...) [by listeners] seated in different listening locations, (...) within and outside (...) the optimal listening area" (Marentakis & McAdams, 2013, p. 22:5). In the second experiment, performed in a controlled laboratory space in which only one subject is sitting at a time in the best listening position, "... the congruency of audiovisual stimulation, the sensory focus of attention, and the attentional process involved (selective or divided) are manipulated" (Marentakis & McAdams, 2013, p. 22:5). The authors found out that the identification of spatial sound trajectories is substantially improved when the performer's gestures are visible to the listeners, but that the listeners' attention is consequently directed to vision. As a result, the auditory motion information is not properly retained, which makes the identification in incongruous audiovisual motion stimulation situations more difficult or even impossible. However, when the attention is primarily focused on audition, the auditory motion information is preserved in the case of unambiguous auditory motion trajectories and performance is therefore improved, being only altered by the visible performer's gestures in the situation in which the auditory motion trajectories are ambiguous. When attention is given to both auditory and visual movement feedback, the auditory motion information is also badly retained.

2.6.4 Analysis of Gesture and Movement

Although we did not directly use any of the methods of analysis of gesture and movement, presented in the following paragraphs, in our research work, but we rather did an analysis of triggered sounds by gesture, time spent to localize sound by means of gesture, and respective hand angle relative to the direction of sound (cf. appendix E.8), we nevertheless consider that it is important to expose the existence of these tools.

Thus, the Hungarian dancer, choreographer, teacher, and theoretician Rudolf von Laban (December 15, 1879 - July 1, 1958)[72], one of the pioneers of modern dance in Europe and one of the most important dance theoreticians of the twentieth century (Gambetta, 2005, p. 25; Campbell, 2005, p. 7), was responsible for the development of a method and language, called Laban Movement Analysis (LMA), in order to study, perceive, describe, interpret, and visualize the phenomenon of human movement as a medium complete in itself in a qualitative and quantitative manner (Sutil, 2013, p. 173). He has therefore laid the foundations of a system of movement notation, called Labanotation or Kinetography, which he published in 1928 under the name of *Kinetographie Laban*, in order to record and analyse movement and choreography as a psycho-physical process (Gambetta, 2005, pp. 29–30). Although originally developed for dance, LMA can be used in the analysis of any movement of the human body in any circumstance (Campbell, 2005, p. 6). LMA is currently one of the most important tools used in areas beyond dance, such as music, actor training, education, athletics, medicine, physical and occupational therapy, psychology, work and industrial efficiency, management and business consultancy, and conflict resolution (Gambetta, 2005, p. 30; Sutil, 2013, p. 174).

The language of LMA consists of terms which allow the description of whole body movements or of parts of it (Gambetta, 2005, p. 30). A brief overview will be presented next. Thus, movement can be divided into four categories (Gambetta, 2005, p. 30; Campbell, 2005, p. 13; Jensenius et al., 2006, p. 178), which are part of the Laban/Bartenieff Movement Fundamentals, so that a particular characteristic of movement can be observed individually, although they are interconnected: 1) *Body*; 2) *Effort*; 3) *Space*; and 4) *Shape*. The combination of these four categories give a comprehensive view of the movement as a whole.

The first category (*Body*) has to do with the way the body is used, that is, if postures, gestures, or whole body movements are used (Gambetta, 2005, pp. 32–33). Furthermore, it is possible to identify the exact location in the body where the movements are initiated, such as the torso, shoulders or hips, elbows or knees, hands and fingers or feet and toes. Simultaneous, successive, sequential, or whole body movements can also be identified.

The second category (*Effort*) describes the energy that a person applies to movement (Campbell, 2005, p. 15) through which his or her feelings, emotions, inner intentions, etc., are expressed (Gambetta, 2005, p. 41). *Weight* (strong or light), *Time* (quick or sudden, or sustained), *Space* (direct, or indirect or flexible), and *Flow* (free or bound) are terms which Laban found to classify human movement based on effort (see figure 2.50).

With respect to the third category (*Space*), Laban defines *kine-sphere* as an imaginary spherical space commonly centred at a person's body centre, whose boundaries can be reached physically with the limbs, regardless of the performed movements (Gambetta, 2005, p. 34; Campbell, 2005, p. 15), and where "... a person's sense of influence and ownership" can be considered psychologically (Campbell, 2005, p. 15). Nevertheless, it can be centred at any other location in the body, usually the nearest articulation joint depending

[72] Retrieved 27/07/2016, from `https://en.wikipedia.org/wiki/Rudolf_von_Laban`

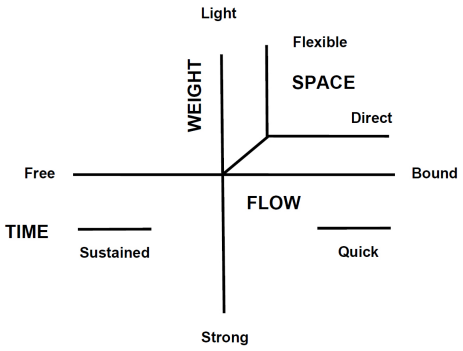


Figure 2.50: Laban Effort Graph (Gambetta, 2005, p. 37).

on which body parts are active (Longstaff, 2005, pp. 10–11), or even outside of the body in the external space (see figure 2.51).

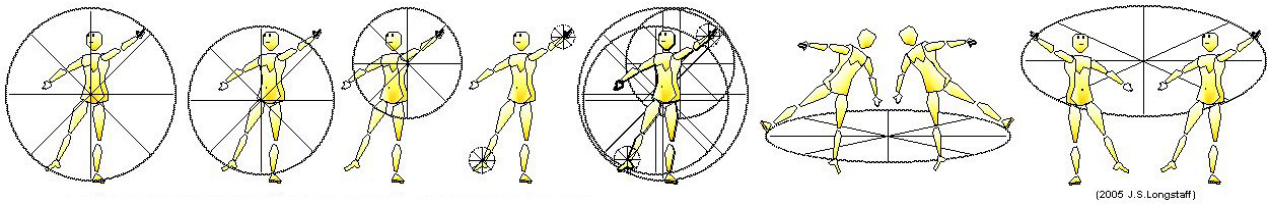


Figure 2.51: The centre of the Kine-sphere (Retrieved 03/08/2016, adapted from `http://www.laban-analyses.org/laban_analysis_reviews/laban_analysis_notation/space_harmony_choreutics/kinesphere_scaffolding/center_of_kinesphere.htm`).

Although the kine-sphere has a spherical form, Laban considered that a trained body should make harmonic or aesthetically agreeable movements within it following the forms of the five regular Platonic solids (tetrahedron, hexahedron or cube, octahedron, dodecahedron, and icosahedron) (Sutil, 2013, p. 177–178) (see section 2.3.4.7), which are the basis of Choreutics or Space Harmony (Gambetta, 2005, p. 27).

Similarly to what has been defined for the head-related reference system in section 2.3.2, in this case the vertical (door), horizontal (table) and sagittal (wheel) planes of the body's dimensional cross of axes (height: up–down; width: left-right; depth: forward-backward) are defined relatively to the orientation of a movement (Gambetta, 2005, p. 35) (see figure 2.52).

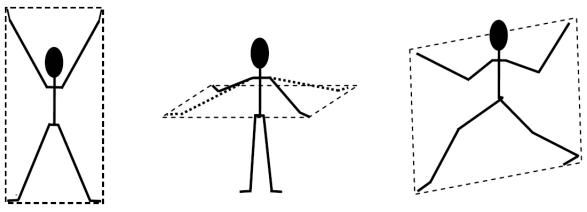


Figure 2.52: The vertical (door), horizontal (table), and sagittal (wheel) planes in Laban Movement Analysis (Gambetta, 2005, p. 35).

The forth category of movement (*Shape*) describes the continuously changing visual aspect of the body or of parts of it during movement (Gambetta, 2005, pp. 38–39). Three modes can be distinguished: a) *Shape flow*, in which qualities of movement, such as growing and shrinking or opening and closing, can be identified;

b) *Directional movement*, in which spoke-like or arc-like movements can be described; and c) *Shaping*, in which the shape of real or imaginary objects can become perceptible through movement.

Movement patterns can therefore be decomposed according to *what* is moved, *where*, *when*, and *how* (Campbell, 2005, p. 7).

In 1991, Christophe Ramstein proposes the analysis of instrumental gestures (see section 2.6.1), based on *descriptive* or *phenomenological* (speed of the movement, space taken up by the gesture, and frequency of movement), *functional* (functions that a gesture performs in a given situation), and *intrinsic* approaches (performer's perception of the parts of the body that are most suitable for a certain situation) (Cadoz & Wanderley, 2000, p. 74; Miranda & Wanderley, 2006, p. 7). Analyses of the orientation of a movement and behaviour of gestures within a system can be made based on frontal (moving away and towards a device), vertical (where gravity accentuates or opposes a gesture), and lateral action (side to side movement), similarly to what has been proposed in LMA.

According to de Götzen (2004, p. 6), the process of analysis of expressive content in human movement and gesture and in musical gesture performance "... starts from gesture-derived information (physical movements or audio signals), captured by sensors into a computing system." This information can then be statistically processed and analysed.

In turn, a method for analysing musical gestures is proposed by Jensenius et al. (2010, pp. 28–30): 1) Observation and/or introspection; 2) Documentation of the perceived musical gestures and sound by applying qualitative and/or quantitative methods; 3) Motion capture by using technology, such as video-based computer vision techniques, infra-red, electromagnetic, ultrasound, mechanical and inertial motion capture systems; 4) Processing and representation of motion capture data; 5) Simulations and/or animations; and 6) Annotation and interpretation. "The use of high-resolution motion capture systems has enabled the quantitative study of these movements" (Winters & Wanderley, 2012, p. 227). Although they are most often used in the analysis and design of interactive systems, Jensenius et al. (2010, pp. 24–25) also use dimension spaces in order to visualize the relationships between different functions of the gestures of a musician and a dancer.

2.7 Gestural Controllers

The gesture or body movement of a human being can be used as a command or interface to control many different tools, instruments, musical instruments, or devices. In the case of traditional musical conducting, although the movements of the conductor do not control directly the sounds produced by an orchestra, they can indicate the tempo, dynamics, and progression of music, as well as the entries of the musicians, allowing them to play as synchronized as possible (see section 2.6.2).

Thus, several haptic/tactile or non-haptic/non-tactile/empty-handed (see section 2.6) gestural controllers or input devices have been developed in domains such as the human-machine interaction (HMI) and human-computer interaction (HCI) domains, as well as in the musical and other artistic domains. Considered as a single device which consists of one or several sensors (Wanderley & Depalle, 2004, p. 635), many of these gestural controllers, whose description is beyond the scope of this thesis, are reviewed in Rovan & Hayward (2000), Miranda & Wanderley (2006), Bhuiyan & Picking (2009), and Tanaka et al. (2012), just to mention a

few sources. In Wanderley & Battier (2000) an extensive list of resources with hyper-links, containing gestural controllers, is presented.

For instance, the real-time control of the position and movement of sound sources in a three-dimensional space by using gestures with a data glove is presented in Harada et al. (1992). In turn, whereas Rovan & Hayward (2000) propose a tactile feedback system for the hands and feet to be used simultaneously with non-haptic gestural controllers in order to improve their performance, three-axis accelerometers in consumer game-controllers are used in Bencina et al. (2008) to control sound with whole body gesture.

In a traditional musical performance, the use of "... gestures to control sound and music is intuitive..." (Odowichuk, 2012, p. 7), so that musical gestures of a performer result in the production of sound, and visual, auditory, proprioceptive or kinaesthetic, ego-location, and tactile feedback are usually obtained and used in order to evaluate and immediately correct the result if necessary, forming a closed loop (Rovan & Hayward, 2000, p. 356) (see also section 2.6.2). Whereas proprioceptive or kinaesthetic feedback, as part of haptic sensations, has to do with the "... awareness of one's body state, including position, velocity and forces supplied by the muscles through a variety of receptors located in the skin, joints, muscles, and tendons" (Rovan & Hayward, 2000, p. 357), ego-location feedback is related with "... the awareness of one's overall position within a defined space, or with respect to objects in that space" (Rovan & Hayward, 2000, p. 357). In turn, tactile feedback, also forming part of haptic sensations, "... is associated with discriminative touch as in the perception of surfaces" (Rovan & Hayward, 2000, p. 356). Furthermore, feedback can also be used for short-term and long-term learning (Rovan & Hayward, 2000, p. 356).

Haptic/tactile gestural controllers are used in order to obtain similar results in a virtual world as in the real world, but a distinction between haptic devices and tactile simulators is made (Rovan & Hayward, 2000, p. 361). In the former case, the device uses a mechanical system which transmits large-scale mechanical feedback, also known as force feedback, to a performer, so that he or she can manipulate a virtual mechanical system. In the latter case, a small-scale mechanical feedback is transmitted to the performer to simulate "... the effect of skin touching a surface..." (Rovan & Hayward, 2000, p. 361) through "... some mechanism of controlled skin deformation (matrix of pins, typically) or vibrotactile stimulators (devices vibrating at a given frequency, in contact with the skin at one or several locations)" (Rovan & Hayward, 2000, p. 361).

However, when a non-haptic gestural controller is used, for instance for tracking a performer's gesture, providing freedom of movements to the performer, "... the tactile feedback loop is broken, forcing performers to rely on proprioceptive, visual and aural cues" (Rovan & Hayward, 2000, p. 356), so that "... egolocation becomes the primary feedback skill necessary to develop (...)" (Rovan & Hayward, 2000, p. 357). According to Rovan & Hayward (2000, pp. 356, 357), as a consequence, gestural accuracy is therefore affected, but Miranda & Wanderley (2006, p. 11) consider that "... there is no imminent need to provide the user with tactile or force feedback" when an empty-handed device is used, although it improves performance.

2.8 Mapping

The process in which data elements of two different data models are linked together somehow is known as data mapping (Odowichuk, 2012, p. 8). Thus, it is possible to extract data from gestures "... directly from

individual sensors or as a result of signal-processing techniques...” (Miranda & Wanderley, 2006, p. 14) and to map them somehow to sound-related data. In this sense, mapping ”... determines the degree of control accessible to a user interacting gesturally with sound processes” (Françoise, 2013, p. 1051).

Four types of gesture mapping strategies can therefore be applied in this relationship (Miranda & Wanderley, 2006, p. 16; Halmrast et al., 2010, pp. 208–209): 1) *One-to-one* or *direct mapping* (Wanderley & Depalle, 2004, p. 638) is the simplest type of mapping, in which a single gestural data element affects only one sound-related data element, such as moving a hand closer and farther away from a metal antenna of a Theremin, the first empty-handed gestural-based expressive electronic musical instrument ever build, patented in 1928 and named after its Russian inventor Lev Sergeyevich Termen (August 27, 1896 - November 3, 1993)[73], known in the West as Léon Theremin, to control either the sound level or its frequency (de Götzen, 2004, p. 7; Odowichuk, 2012, p. 11); 2) *One-to-many* or *divergent mapping* (Odowichuk, 2012, p. 9) is the relationship, in which a single gestural data element affects numerous sound-related data elements, such as using the speed or velocity of a key-stroke on a single key of a sampled piano instrument in order to select different audio samples, so that louder piano samples are reproduced according to higher velocities, causing spectral changes to be perceived as in a real acoustic piano; 3) *Many-to-one* or *convergent mapping* (Odowichuk, 2012, p. 9) is the mapping, in which many gestural data elements affect only one sound-related data element, such as the overall quantity of motion of many different gestures, captured by a video camera, to adjust the overall level of sound; and 4) *Many-to-many* is the most complex type of mapping, in which many gestural data elements affect numerous sound-related data elements, such as using a video camera to capture left hand horizontal gestures, which adjust both the harmonic contents of sound through a filter and the overall sound level, and right hand vertical wobbly gestures, which adjust both frequency and vibrato of sound. The choice of any of these strategies depends on many variables, such as the context, intention, available technology and equipment, and perception.

For instance, Schacher (2007) uses gesture to move perceptual sound sources with physical object properties in a three-dimensional space. Thus, these sound sources can be grabbed, thrown, pushed, or spun by using a data glove. In turn, Bencina et al. (2008) map body motion with sound in order to ”... give expressive sonic capabilities to the whole body in motion” (Bencina et al., 2008, p. 197), so that sound production is the direct consequence of body movement. This is achieved by using three-axis accelerometer data from Nintendo Wii remotes, obtained from body movement, in order to trigger and modify some sound attributes.

2.9 Interactive Installation

2.9.1 Interactivity

In the 1960’s, many artists became interested in sound and video recording systems, leading them to believe that an artist’s ”... work could respond to and even be controlled by the viewer...” (Gibbs, 2007, p. 33), that is, that there could be interaction between the viewer and an artist’s work (Gibbs, 2007, pp. 32–33). This aspect raised naturally many questions about the relationship between an artist, an artist’s work, and the audience, a

[73] Retrieved 04/09/2016, from `https://en.wikipedia.org/wiki/L%C3%A9on_Theremin`

matter on which the American composer John Cage (see section 2.3.1) had a great influence.

Based on the definition of the American computer game designer Chris Crawford (born in 1950)[74], interactivity can be regarded as an iterative[75] process of listening (input), thinking (processing), speaking (output), and feedback between two or more human or non-human performers (Gibbs, 2007, p. 102) (see section 2.9.2). Thus, the interface, a "... mediating structure between a system and the person using it" (Gibbs, 2007, p. 170), turns out to be one of the most important aspects to be considered along with the way data is mapped (cf. section 2.8). Since interactivity and sound are both related with time (see section 2.1.3.1), interaction with sound makes immediate communication and entertainment possible.

2.9.2 Performance

The terms *performance* and *performative* and the expression *performative utterance* were introduced by the British philosopher of language and linguist John Langshaw Austin (March 26, 1911 - February 8, 1960)[76] in the 1950's. They have been used since then to refer to utterances of natural languages, with which a certain act or activity is performed. Therefore, the performative utterance has to be carried out in a completely appropriate situation regarding the act and under certain conditions, as opposed to those utterances with which something is solely described or noticed (constative utterances). One of the examples that Austin (1962, p. 5) presents as a performance utterance and that clearly reflects the doing of an act is that of a ceremonial ship launching: *I name this ship the Queen Elizabeth* (see also Pinto de Lima, 1983, pp. 43–44).

Traditionally, performance is understood as one in which the artist is on stage and simply acts to an audience. However, a new concept of performance appears in the 1950's and 1960's in the form of events or happenings (Gibbs, 2007, p. 126). At these events, which the American composer John Cage (see section 2.3.1) considers as spontaneous theatrical events, artists do not perform in the usual places as previously, such as auditoriums or theatres, but rather in galleries, trying to explore the idea of performance as art by means of unpredictability, involving the direct or indirect participation of the audience. In this respect, Choi (2000, p. 144) states that "performing art is a formalized presentation of artwork in social and cultural venue where the work of art is intended to reach a "public" audience." Furthermore, according to Choi (2000, p. 141), performance is a formal presentation of high-order emotional synthesis.

Currently, any actor (human or not) can be a performer and interaction can exist between actors (see section 2.9.1). As a curiosity, it should be noted here that the word *theatre*, *theatron* in Greek, originally means *place to see*, and that the word *auditorium*, taken from Latin, means *place to listen*. Since the audience both radiates sound in a theatre or an auditorium - through applause, laughter, coughing, comments with a close spectator - and also captures it, the audience can also be considered as a performer, to some extent.

The increase of the processing capacity of computers in the recent years has made their use as actors possible, too, apart from being capable of playing the role of musical instruments in the least conventional sense, of sound generators, allowing or not the imitation of other traditional or electronic instruments, or allowing playback of music or pre-recorded sounds, among other alternatives. A performer, such as a disc jockey,

[74] Retrieved 04/09/2016, from `https://en.wikipedia.org/wiki/Chris_Crawford_(game_designer)`

[75] "A repeated process whereby the output of a system is (normally) fed back to the input to be processed a second (and subsequent) time/s" (Gibbs, 2007, p. 170).

[76] Retrieved 05/09/2016, from `https://en.wikipedia.org/wiki/J._L._Austin`

can use disc mixing techniques, as well as remixing and scratching techniques, together with a computer to develop further techniques.

2.9.3 Gesture, Sound, and Interactive Installation

An interactive installation is one that responds somehow interactively to the presence, gesture, or action of one or many subjects on an artist's work (Gibbs, 2007, p. 102). The result is therefore affected by the subjects, who become participants in the process. Since many forms of interaction are possible, interactive systems can be quite complex, as in interactive audiovisual installations. For instance, Grigoriou & Floros (2010) present an interactive audiovisual installation, in which one subject at a time interacts with the stars in the space by using hand gestures to control both audio and visual reproduction in real time. High-quality video projection on a curved dome is used to create a representation of the sky and Ambisonics surround sound technology (see section 2.3.4.7) is employed with an authentic imaginary sound environment soundscape, based on sounds of nature recorded during the night. In turn, the collaborative project called *Gedankenscherz*[77] is an interactive audiovisual installation, in which subjects use gestures, captured by a Microsoft Xbox 360 Kinect depth camera, in order to navigate in virtual worlds, based on a baroque historical collection described by the German polymath and philosopher Gottfried Wilhelm Leibniz (July 1, 1646 - November 14, 1716)[78] in 1675. The gestures are accompanied by atmospheric sound effects.

2.10 Chapter 2 in Retrospect

This chapter has inevitably turned out to be rather long, not only because of internal requirements and reasons to the study itself – theoretical foundations are presented within it, which provide the real presuppositions and the supporting pillars of this work, with all the rich, varied, and indispensable information of diverse origins and the responsibility of renowned authors –, but also due to terminological requirements and above all due to didactic concerns.

In fact, we have been engaged in teaching Acoustics and Sound Engineering for many years now and we consider that such a precise and meticulous research work can enclose several sources, which can be explored in a specific deliberate way, so that the tools or information thus selected and collected can be made available to anyone who comes to consult or use it by means of the exposition contained in our work.

This is the reason why our option has been not to concentrate only on one single aspect or idea without following its evolution, and particularly its updating, based on the initiative taken by their own author(s), filtering out of these texts what appeared to be to us the most coherent and clearly formulated line of thought. At the same time, this is in our view a guarantee that the information obtained in this way will be not only the starting point but the veritable basis of what we have been doing and whose results are the subject of chapters 3 and 4.

On the other hand, it is also for pedagogical reasons that we often did not avoid some redundancy in our

[77] Retrieved 09/09/2016, from `http://www.humboldt-forum.de/humboldt-lab-dahlem/projektarchiv/probebuehne-1/gedankenscherz/teaser/`

[78] Retrieved 09/09/2016, from `https://en.wikipedia.org/wiki/Gottfried_Wilhelm_Leibniz`

text, so that there are no doubts regarding the discussed concepts and points of view. In this sense, we also used frequently the expression 'that is', that functions as a marker to show that what follows is an explanation of what has previously been said in other words or in rather technical terms.

Chapter 3

System Development

Figure 3.1 shows the interactive installation (see section 2.9), which we set up in the Motion Capture Laboratory at the School of Arts, at the Portuguese Catholic University (EA-UCP), in Porto, from the 24th to the 29th September 2012, in order to carry out the experimental part of the proposed study. A laptop computer with proper software, a television set serving as a computer monitor, a sound card, a depth camera and eight active loudspeakers were therefore used.

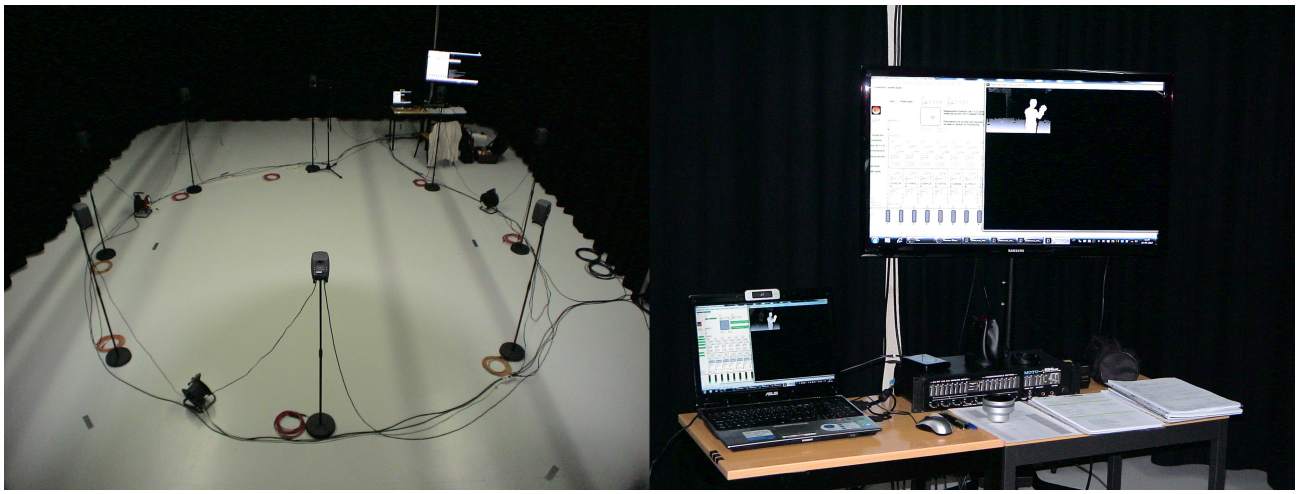


Figure 3.1: System set-up. Left: eight active loudspeakers in a circular arrangement. Right: a sound card and a television set, both attached to a laptop computer.

3.1 Selected Room for the Research

The room that we selected for the research was the Motion Capture Laboratory at the School of Arts, at the Portuguese Catholic University, because it was a common reverberant room (broadband $RT_{60} < 1$ s when black flannel crimped curtains were covering the walls) (see section 2.3.2)[1], it had the suitable dimensions and a part of the necessary equipment for the experiments, and it was available for a whole week.

[1] In what follows, all these references to the several sections of the previous chapter are intended to clearly show that in chapter 3 we are using and exploring relevant information already dealt with.

The Motion Capture Laboratory has the working dimensions of about 10.67 m (length) $\times$ 7.66 m (width) $\times$ 4.94 m (height) (see figure 3.2 and visit `http://artes.porto.ucp.pt/VisitaVirtual/EdArtes/Mocap/tour.html` for a virtual tour). The floor is made of concrete, coated with grey vinyl. The walls and the ceiling are plastered and painted in white. Only the walls are covered with black flannel crimped curtains with the purpose of lowering the reverberation of the room in order to achieve an ordinary reverberant room and to allow more accurate localizations of sounds (see section 2.5.3).

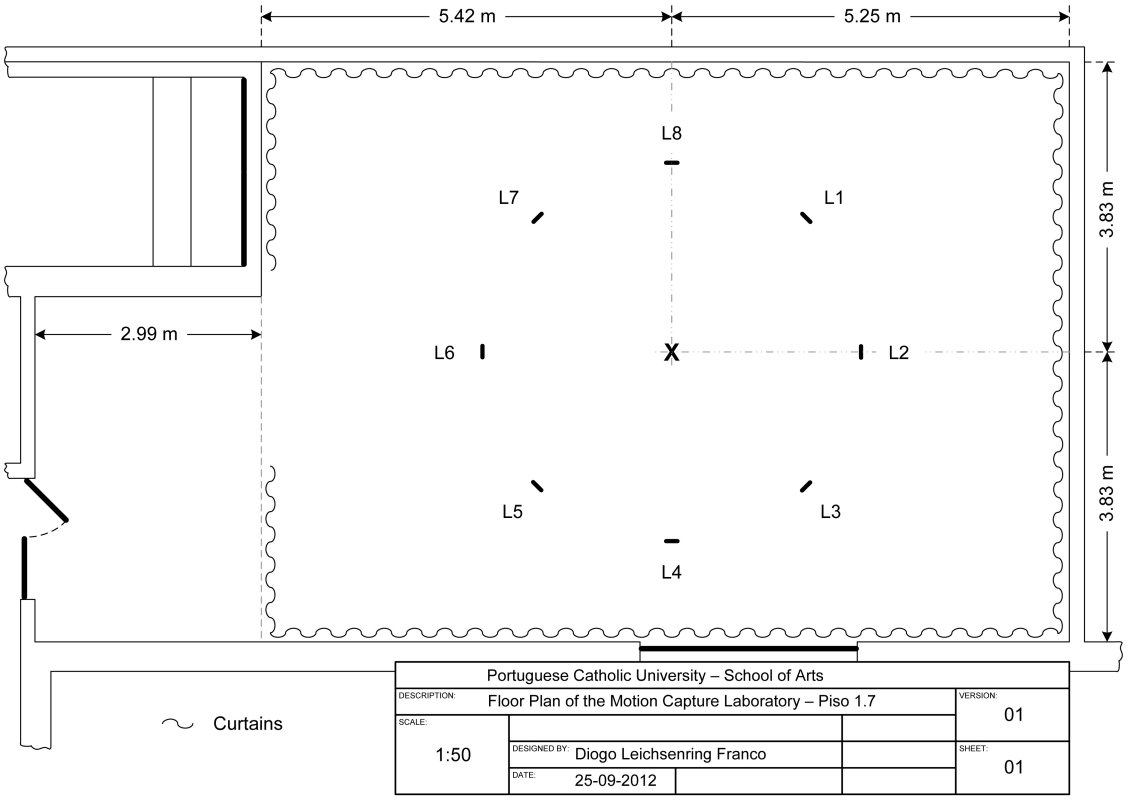


Figure 3.2: Floor plan of the Motion Capture Laboratory and the location of the eight loudspeakers (L1 to L8).

3.2 Equipment Used in the Research

A Microsoft Xbox 360 Kinect for Windows, version 1.5, depth camera[2] is placed close to a loudspeaker (L2 in figure 3.2), which serves as a spatial reference to the system (see section 3.6), so that its infra-red projector, RGB camera, and infra-red camera centres are at a height of 1.525 m from the ground, which corresponds to the height of the main centre axis or acoustic axis (see section 2.2.1 and figure 3.4) of each loudspeaker to the ground (mean hearing height) (see figure 3.3). This height was chosen to make it possible for us to capture and track the desired gestures of the participants (see section 3.5) from a frontal position around the 'X'-marked area shown in figure 3.2, taking into account the user calibration process described in section 3.6.

We selected this camera firstly because it is a depth camera suitable for whole-body tracking, able to capture a large range of gestures from about 50 centimetres to circa 7.6 metres distance (Borenstein, 2012, p.

[2]Retrieved 20/05/2014, from `http://download.microsoft.com/download/B/A/4/BA4D9FA4-7E68-447E-9C63-17C1C62850FB/kinect_en_nl-NL_pt-PT_es-ES.pdf`

Figure 3.3: Left: a GENELEC 6010A loudspeaker (L2 in figure 3.2) with its main centre axis at the horizontal 0º position and a Microsoft Xbox 360 Kinect depth camera. Right: a close-up of the same loudspeaker and the same camera with its infra-red projector (left), RGB camera (centre), and infra-red camera (right).

67), with a field of view of 57º in the horizontal direction and 43º in the vertical direction (Odowichuk, 2012, p. 22), and without being sensitive to the light conditions in a room. Secondly, because it allows an empty-handed tracking, that is, there is no need for holding controllers with the hands, as explained in section 3.5. In addition, it is a relatively inexpensive device, which is usable on various platforms with open source drivers and software. These reasons led us to buy it on purpose for us and to use it in this research.

The camera is attached to our own ASUS PRO57VR-AS069C laptop computer[3] via an USB 2.0 port. The computer is equipped with an Intel®Core™2 Duo CPU P8600@2.40 GHz processor, a 4 GB RAM memory (although only 3 GB are usable by the following indicated operating system), and an ATI Radeon HD 3470 graphics card, running a 32 bits Microsoft Windows 7 Professional Portuguese operating system.

In addition, a Samsung UE46D5500RWXXC 46" D5500 Series 5 SMART Full HD LED TV set[4], property of the EA-UCP, is used as a computer monitor and connected to the computer via the VGA port, so as to allow the researcher to have larger image visualization and better observation during the experiments.

Our own MOTU 896HD (firmware 1.01; hardware 1.0; driver 4.0.5.3503) sound card[5] is linked via IEEE 1394 FireWire 400 to the computer as well, in order to provide eight audio analogue outputs to an octophonic sound reproduction system composed of eight visible active GENELEC 6010A loudspeakers[6] belonging to the EA-UCP, arranged in a horizontal circular form. The horizontal angle between each loudspeaker's main centre axis or acoustic axis (see figure 3.4 and section 2.2.1), that is, an imaginary perpendicular line to the loudspeaker's front panel, which results from the intersection between the planes formed by the vertical and the horizontal coverage angles of the loudspeaker, and where the best response of the loudspeaker is achieved, is equal to 45 degrees. The horizontal distance from each loudspeaker to the 'X' central hearing point is equal to 2.5 m. The height of the acoustic axis of each loudspeaker to the ground is equal to 1.525 m (mean hearing height), as already mentioned above. These measurements were made by using our own calibrated BOSCH

[3]Retrieved 20/05/2014, from `http://support.asus.com/Download.aspx?SLanguage=en&p=3&s=123&m=Pro57Vr&os=29&ft=12&f_name=E3840_M51_HW.zip#E3840_M51_HW.zip`

[4]Retrieved 20/05/2014, from `http://downloadcenter.samsung.com/content/UM/201112/20111216084511364/Web_W_X6DVBEUA_Eng.pdf`

[5]Retrieved 20/05/2014, from `http://cdn-data.motu.com/manuals/firewire-usb-audio/896HD_Manual_Win.pdf`

[6]Retrieved 20/05/2014, from `http://www.genelec.com/documents/opmans/OM6010A.pdf`

PLR50 measuring laser device[7] and a Parkside Laser-type Spirit Level 670 (see appendix D). The horizontal distance from each loudspeaker to the 'X' central hearing point was also confirmed by impulse responses (cf. section 2.4.1), measured from each loudspeaker to this point with SATlive software, version 1.40.18, developed by Thomas Neumann (Tomy Soft)[8], a Behringer ECM8000 Omnidirectional Measurement Microphone[9], and a Rion sound calibrator NC-74[10], so that propagation delays were equal.

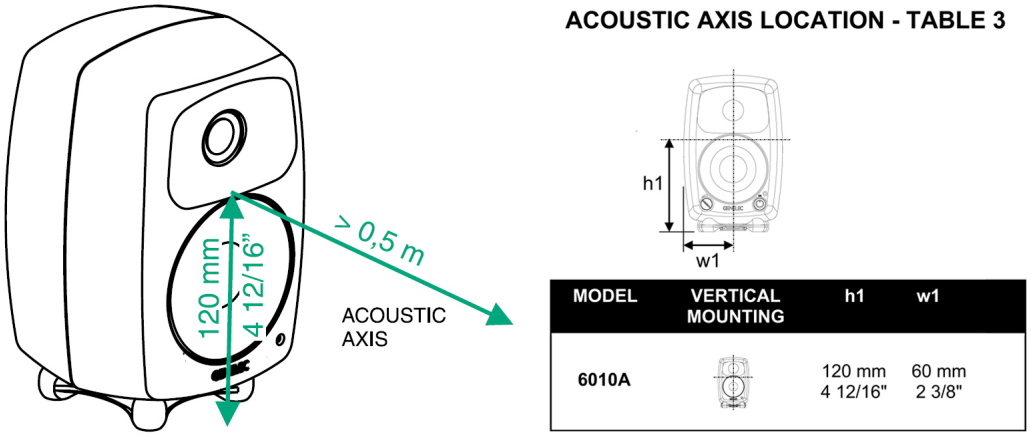


Figure 3.4: Genelec 6010A main centre axis or acoustic axis location (adapted from `http://www.genelec.com/documents/opmans/OM6010A.pdf`, p. 2, and `http://www.genelec.com/documents/other/acousticaxis.pdf`, p. 3, both retrieved 20/05/2014).

Each loudspeaker was turned on and calibrated individually with a pink noise signal, generated by the Cycling'74's visual programming MAX MSP Jitter 6.0.1 (50928) software[11] at a sampling rate of 48 kHz (see section 2.3.4.6), in order to produce an A-weighted sound pressure level of about 66.6 $\mathrm{dBA} \approx 67.0\ \mathrm{dBA}$ (see section 2.1.3.2) at 1 m distance on the acoustic axis and the flattest frequency response as possible. The patch that was used to produce the pink noise can be seen on the left side of figure 3.5. The right side of that figure shows that the Crest Factor (CF) of the normalized pink noise signal generated by MAX MSP Jitter (see section 2.1.4) is equal to 4.53, which corresponds to a CF level of about $20 \times \log_{10} 4.53 \approx 13.11\ \mathrm{dB}$. The amplitude of the signal that was sent to all loudspeakers was equal to 0.1, except for loudspeakers in position L4 and L5 in figure 3.2, which had to be increased from 0.1 to 0.22, so that they produced the same level at 1 m distance as the others. This difference of $20 \times \log_{10} \frac{0.22}{0.1} \approx 6.85\ \mathrm{dB} \approx 7\ \mathrm{dB}$ was due to an electric potential level problem in the output channels 4 and 5 of the MOTU 896HD, which was solved by increasing the output amplitude of the signal in those channels in MAX MSP Jitter. Their characteristics remained identical to those of the other channels.

The A-weighted sound pressure level was measured with a calibrated IVIE IE-35 Audio Analysis System[12], applied to a Dell Axim X51v PDA[13], both property of the Music Academy of Espinho (AME). As already men-

[7] Retrieved 20/05/2014, from `http://www.bosch-do-it.de/media/media/diy/diymedia/199928/199929/24116/24116_bedienungsanleitung_westeuropa/447928_2609140621_201107pdf.pdf`

[8] Retrieved 15/06/2017, from `https://www.satlive.audio/en/`

[9] Retrieved 15/06/2017, from `http://www.music-group.com/Categories/Behringer/Microphones/Condenser-Microphones/ECM8000/p/P0118`

[10] Retrieved 15/06/2017, from `http://rion-sv.com/products/10000%EF%BC%9A10013/NC740009`

[11] `https://cycling74.com/`

[12] Retrieved 14/01/2014, from `http://www.ivie.com/download/IE-33%2635%20Man_060110.pdf`

[13] Retrieved 14/01/2014, from `http://downloads.dell.com/manuals/all-products/esuprt_electronics/esuprt_axim/dell-axim-x51_owner%27s%20manual_en-us.pdf`

tioned in section 2.1.3.2, the A-weighting is recommended in measurements of loudness of any sound at any sound pressure level, in order to maintain the consistency between measurements, although actual hearing response changes with sound pressure level (Stark, 2002, p. 62; Howard & Angus, 2001, p. 85). When all loudspeakers played the same pink noise signal simultaneously, the A-weighted sound pressure level was equal to approximately 72.8 $\mathrm{dBA} \approx 73.0\ \mathrm{dBA}$ at the central hearing point in the room, at 2.5 m distance from each loudspeaker.

We also obtained identical results with the use of the SATlive software, the Behringer ECM8000 Omnidirectional Measurement Microphone, and the Rion sound calibrator NC-74. The actual frequency response of each individual loudspeaker was measured to be within $\pm 3\ \mathrm{dB}$ in the frequency range of about 80 Hz to 18 kHz.

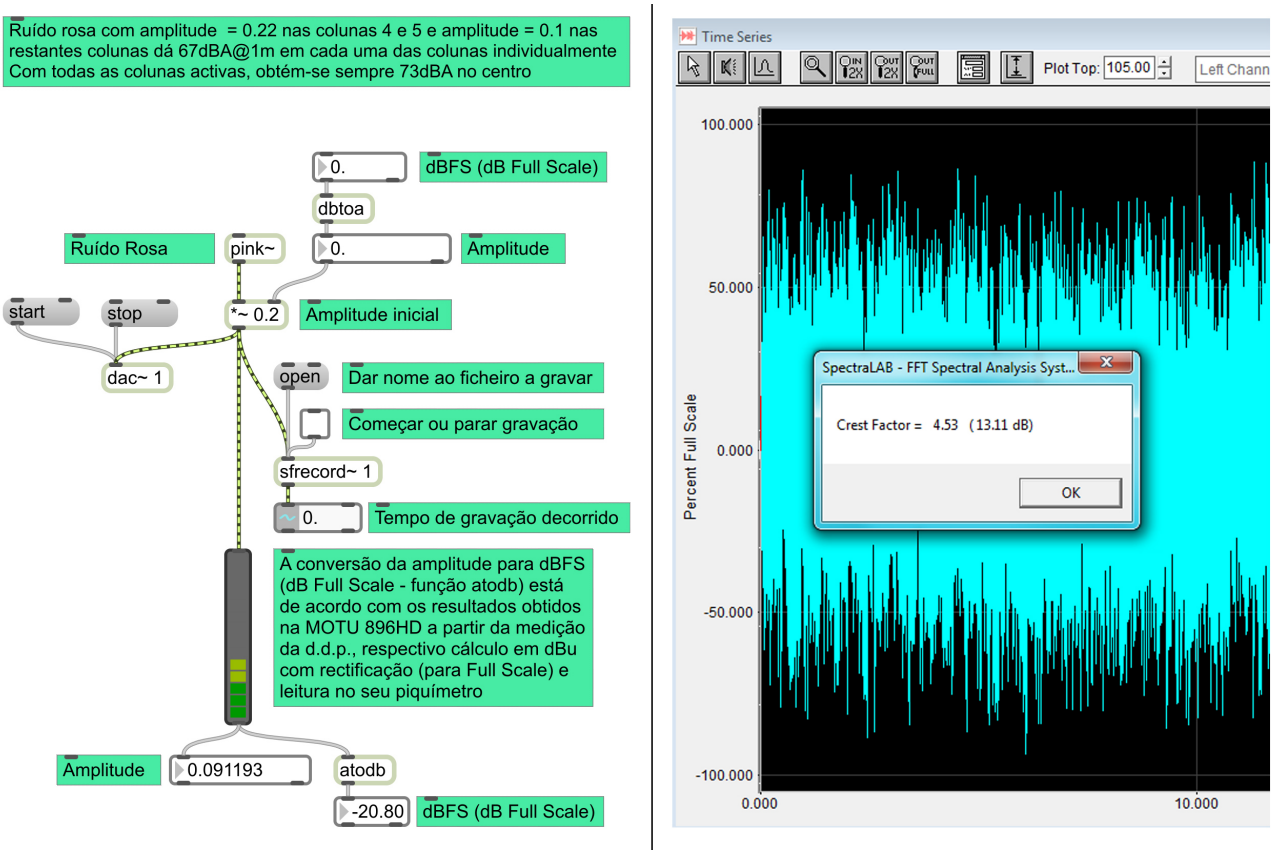


Figure 3.5: Left: MAX MSP Jitter pink noise generator patch. Right: Crest factor of the normalized pink noise signal generated by MAX MSP Jitter, measured by SpectraLAB 4.32.17 audio analysis software.

The signal flow diagram of the whole system can be seen in figure 3.6.

3.3 Sound Reproduction System Used

In this research, we decided to use eight equally spaced active GENELEC 6010A loudspeakers (see section 3.2), whose free field frequency response goes from 74 Hz to 18 kHz ($\pm$ 2.5 dB)[14], positioned in a horizontal circular form, because the quality of perceptual sound sources is relatively good for a large listening area in

[14] See pp. 5–6 of the loudspeaker's operating manual, retrieved 20/05/2014, from `http://www.genelec.com/documents/opmans/OM6010A.pdf`

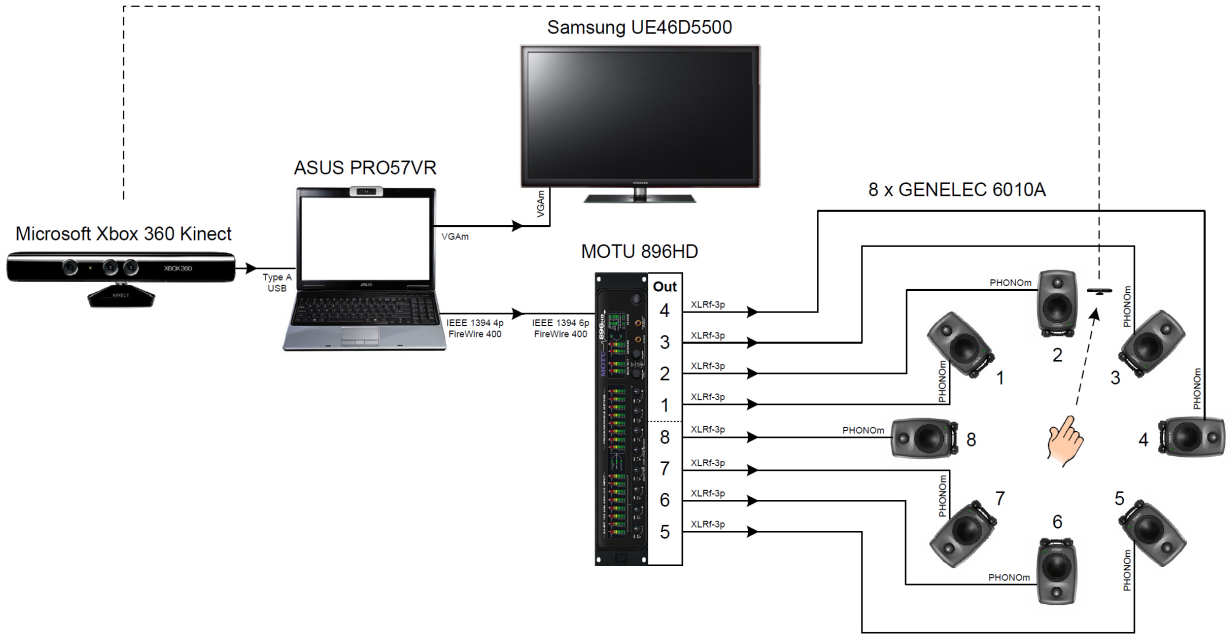


Figure 3.6: Signal flow diagram of the whole system.

an eight loudspeaker sound reproduction system where the angle between loudspeakers is of 45º (Pulkki, 2001b, p. 29) (see section 2.3.4). In addition, in our research Ambisonics Equivalent Panning (AEP), proposed by Neukom & Schacher (2008), is used to synthesize the horizontal sound field only, known as pantophonic reproduction, so that a signal is sent to the eight loudspeakers around a listener, but with different gain factors (see section 2.3.4.7).

Since the same signal is present in all loudspeakers, although with different amplitudes, the spatial width of the perceptual sound source is almost kept constant (see section 2.3.4), which provides stable perceptual sound sources (Odowichuk, 2012, p. 28), and a sound field that remains the same if the listener rotates his head (Grigoriou & Floros, 2010, p. 432). Furthermore, easy implementation (cf. section 2.2.4), simplification, and fast calculations are achieved in real-time for moving sound sources in AEP. Since we were expecting gesture associated with moving sound speeds below 50 meters per second, because the Kinect depth camera we use has a significant amount of latency (average latency equal to about 218 milliseconds - Tanaka et al., 2012, p. 77) and a slow data capture frame rate of 30 frames per second (Borenstein, 2012, p. 60), which makes it difficult to track sudden events (Odowichuk, 2012, p. 20), the Doppler effect was not taken into account (cf. section 2.2.4). Another reason that has led us not to simulate the Doppler effect is that this is a common practice when musical items are used (Marentakis & McAdams, 2013, p. 22:8), as described in section 3.4.

Although the AEP function narrows increasingly with the order and higher sound source directionality is therefore achieved (Neukom & Schacher, 2008), so that fewer loudspeakers are needed, the size of the listening area is increased with higher orders, including off-centre listening positions, yielding superior immersiveness. Since in AEP the exponent in the equivalent panning function, or order of the ambisonic resolution, can be any positive number, we experimentally tried out different orders, first of all in an Excel file, named `Localização de Colunas e Fonte.xlsx` (on the DVD-ROM in the Appendix E.1), and later in the actual sys-

tem, finally ending up by choosing the order equal to 4.64, the same as shown in Neukom & Schacher (2008), because it gave us satisfying results during our own experiments with the system.

3.4 Selected Sounds for the Research

Test signals such as pink noise, speech, pulsed tones, tone bursts, and clicks have been broadly used in localization tests (Blauert, 1997; Frank et al., 2008; Power et al., 2013) (see also section 2.1.4). Nevertheless, in our research we use essentially recorded music as stimulus, since it is more stimulating and familiar to a human being. The contents is part of the Western Music hearing universe and includes some western musical instruments of different timbres.

Recommendations ITU-R BS.1284-1 (International Telecommunication Union, 2003, p. 4) and ITU-R BS.1116-3 (International Telecommunication Union, 2015, p. 9) mention that the audio excerpts used in auditory tests should be typically 10 to 25 s long, because of the short-term human memory limitations, although they may be shorter for some tests. Furthermore, the phrases of musical items should not appear to be interrupted. In our case, from the following eighteen stereo musical items we chose, only one is shorter than 10 s (item 18) and two are longer than 25 s (items 5 and 12), and the phrases are not interrupted, except for items 3 and 7, which do not contain the conclusion of the musical phrase (the items can be found in appendix E.2). We consider that the shorter and longer durations and the endings of the aforementioned items do not significantly affect the results of our research, because the outcomes are not directly dependent on them (see chapter 4):

1. Clarinet and Orchestra (stereo .wav format file, named `Clarinete.wav`, sampling rate of 44.1 kHz (see section 2.3.4.6), 16 bit)
   Excerpt from the 2nd movement (Adagio) of the Wolfgang Amadeus Mozart's Concert for Clarinet and Orchestra in A Major, KV 622; CD "Wolfgang Amadeus Mozart – Klarinettenkonzert, Flötenkonzert No.1, Fagottkonzert", track 5; Clarinet: Harold Wright; Boston Symphony Orchestra; Conductor: Seiji Ozawa; © 1980 Polydor International GmbH (Deutsche Grammophon); from 05'48" until 06'05" (total time: 17.113")[15]

2. Double Bass (stereo .wav format file, named `Contrabaixo.wav`, sampling rate of 44.1 kHz, 16 bit)
   From the CD-ROM "Microsoft Musical Instruments – An Interactive Journey into the World of Musical Instruments"; © 1992 Microsoft Corporation, © 1992 Dorling Kindersley Limited; Path: STRINGS\DBAS\DBASSOLO.WAV (total time: 13.444")

3. Harpsichord (stereo .wav format file, named `Cravo.wav`, sampling rate of 44.1 kHz, 16 bit)
   Excerpt from the Johann Sebastian Bach's Sinfonia IX in F minor, BWV 795; Musicassette 2 "Musique et Tempérament", track 9; Harpsichord: Yannick Legaillard; © Éditions Costallat 1985 (total time: 20.296")[16]

4. Glockenspiel (stereo .wav format file, named `Glockenspiel.wav`, sampling rate of 44.1 kHz, 16 bit)
   From the CD-ROM "Microsoft Musical Instruments – An Interactive Journey into the World of Musical

[15] P. 35, cc. 83–87, retrieved 06/12/2017, from `https://imslp.nl/imglnks/usimg/f/f7/IMSLP29515-PMLP03144-Mozart_Clarinet_Concerto_K622.pdf`

[16] P. 1, cc. 1–4, retrieved 06/12/2017, from `https://imslp.nl/imglnks/usimg/f/f2/IMSLP00770-BWV0795.pdf`

Instruments"; © 1992 Microsoft Corporation, © 1992 Dorling Kindersley Limited; Path: PERCUSSN\GLOC\GLOCSOLO.WAV (total time: 18.621")

5. Trumpet and Piano (stereo .wav format file, named `Korsakov.wav`, sampling rate of 44.1 kHz, 16 bit)
   "Flight of the bumble-bee", Nikolai Rimsky-Korsakov; CD "Ravel, Gershwin, Bernstein, Glazunov – Trumpet Works", track 10; Trumpet: Sergei Nakarjakov; Piano: Alexander Markovich; © 1992 Teldec Classics International GmbH (total time: 1'01.906")[17]

6. Marimba (stereo .wav format file, named `Marimba.wav`, sampling rate of 44.1 kHz, 16 bit)
   From the CD-ROM "Microsoft Musical Instruments – An Interactive Journey into the World of Musical Instruments"; © 1992 Microsoft Corporation, © 1992 Dorling Kindersley Limited; Path: PERCUSSN\MARI\MARISOLO.WAV (total time: 15.000")

7. Oboe and Orchestra (stereo .wav format file, named `Oboé.wav`, sampling rate of 44.1 kHz, 16 bit)
   Excerpt from the 2nd movement of the Franz Schubert's Symphony No. 9 in C Major, D. 944; CD "Yehudi Menuhin erklärt die Instrumente des Orchesters", track 45; EMI Studio DRM (CDM 7 69816 2); Royal Philharmonic Orchestra; Conductor: Rafael Kubelik; © 1988 EMI Electrola GmbH (total time: 14.475")[18]

8. Piano (stereo .wav format file, named `Piano.wav`, sampling rate of 44.1 kHz, 16 bit)
   Excerpt from the Frederic Chopin's Waltz for Piano No. 6 in D Flat Major – 'Minute' / 'Dog Waltz' – Op. 64/1, B. 164/1; CD "Classic Experience II – 30 das mais populares peças da música clássica", Compact Disc 2, track 11; © 1991 EMI-Valentim de Carvalho Música, Lda. (total time: 22.248")[19]

9. Tubular Bells (stereo .wav format file, named `Sinos Tubulares.wav`, sampling rate of 44.1 kHz, 16 bit)
   From the CD-ROM "Microsoft Musical Instruments – An Interactive Journey into the World of Musical Instruments"; © 1992 Microsoft Corporation, © 1992 Dorling Kindersley Limited; Path: PERCUSSN\OCHM\OCHMSOLO.WAV (total time: 17.552")

10. French Horn (stereo .wav format file, named `Trompa.wav`, sampling rate of 44.1 kHz, 16 bit)
    Excerpt from "Grand Ur", Erich Avinger; CD "Tom Bacon – The Flipside – Jazz Horn Solos", track 4; French Horn: Tom Bacon; © 1989 Summit Records (total time: 17.020")

11. Vibraphone (stereo .wav format file, named `Vibrafone.wav`, sampling rate of 44.1 kHz, 16 bit)
    From the CD-ROM "Microsoft Musical Instruments – An Interactive Journey into the World of Musical Instruments"; © 1992 Microsoft Corporation, © 1992 Dorling Kindersley Limited; Path: PERCUSSN\VIBR\VIBRSOLO.WAV (total time: 18.970")

12. Violin (stereo .wav format file, named `Violino.wav`, sampling rate of 44.1 kHz, 16 bit)
    Chapter 3 – Example 7: "Brahms, Symphony No. 3, third movement, mm. 13-24"; CD "SO–1 The Study

[17] Retrieved 06/12/2017, from `http://www.wenatcheemusic.com/index.php/learning-zone/sheet-music-and-articles/category/128-trumpet-sheet-music?download=115:flight-of-the-bumble-bee-for-trumpet-and-piano`

[18] P. 37, cc. 8–14, retrieved 06/12/2017, from `https://imslp.nl/imglnks/usimg/d/d5/IMSLP245918-PMLP25384-FSchubert_Symphony_No.9_CFPeters_1871_fs.pdf`

[19] Pp. 1–2, cc. 1–36, retrieved 06/12/2017, from `https://imslp.nl/imglnks/usimg/d/d7/IMSLP114892-PMLP02373-FChopin_Waltzes,_Op.64_BH9.pdf`

of Orchestration, Second Edition", Samuel Adler, Compact Disc 1, track 40; Violin: Zvi Zeitlin; © 1989 Thomas Frost Productions, Inc., © W. W. Norton & Company, Inc. (total time: 29.767")[20]

13. Xylophone (stereo .wav format file, named `Xilofone.wav`, sampling rate of 44.1 kHz, 16 bit)
From the CD-ROM "Microsoft Musical Instruments – An Interactive Journey into the World of Musical Instruments"; © 1992 Microsoft Corporation, © 1992 Dorling Kindersley Limited; Path: PERCUSSN\XYLO\XYLOSOLO.WAV (total time: 10.064")

14. Piano (stereo .wav format file, named `Piano2.wav`, sampling rate of 48 kHz, 16 bit)
Cubase 5.0 software, HALionOne VSTi, Preset "Hard Grand Piano", chord consisting of the three notes C_3, E_3, G_3 [21] (total time: 11.000")

15. Long Swells (stereo .wav format file, named `LongSwells.wav`, sampling rate of 48 kHz, 16 bit)
Cubase 5.0 software, HALionOne VSTi, Preset "Long Swells", chord consisting of the three notes C_3, E_3, G_3 [21] (total time: 11.633")

16. Chill Detune (stereo .wav format file, named `ChillDetune.wav`, sampling rate of 48 kHz, 16 bit)
Cubase 5.0 software, Korg MS-20 VSTi, Preset "Chill Detune", chord consisting of the three notes C_3, E_3, G_3 [21] (total time: 14.100")

17. Ultimate Trance (stereo .wav format file, named `UltimateTrance.wav`, sampling rate of 48 kHz, 16 bit)
Cubase 5.0 software, Korg LegacyCell VSTi, Preset "Ultimate Trance Anthem", chord consisting of the three notes C_3, E_3, G_3 [21] (total time: 14.100")

18. Orchestra (stereo .aiff format file, named `sacre.aiff`, sampling rate of 44.1 kHz, 16 bit)
Excerpt from "Rite of Spring", Igor Stravinsky; Max 6.0 Cycling'74 Software; Path: Cycling'74\Max 6.0\patches\docs\tutorial-patchers\msp-tut\sacre.aiff (total time: 1.428")

Each of these items is identified in the Cycling'74's visual programming MAX MSP Jitter 6.0.1 (50928) software (see also section 3.7.1), which is used for their reproduction, with the integer number of the item presented above plus one. For instance, item 6 (Marimba) is identified in the reproduction software with number 7. In addition, only the left channel of each item is intentionally used by the software MAX MSP Jitter and reproduced with a sampling rate of 48 kHz (real-time conversion performed by the sound card MOTU 896HD), so that the Ambisonics Equivalent Panning (AEP) algorithm can process it accordingly.

Long-time average spectra (LTAS) (see section 2.1.3.1) and spectrograms of these items, respectively calculated by Sound Forge Pro software[22], version 11.0 (build 272), and by Adobe Audition[23], version 3.0 (build 7283.0), show that:

a) in item 1 (Clarinet and Orchestra), there are basically two frequency regions (from about 10 Hz to 100 Hz and from approximately 300 Hz to 2 kHz) with higher energy or magnitude, the former basically relative

[20] Pp. 56–57, cc. 12–24, retrieved 06/12/2017, from `https://imslp.nl/imglnks/usimg/a/a0/IMSLP23120-PMLP01698-BraWV,_S._372.pdf`

[21] According to the International Pitch Notation (IPN).

[22] `http://www.magix-audio.com/us/sound-forge/`

[23] `http://www.adobe.com/support/downloads/product.jsp?platform=Windows&product=92`

to background noise and low frequencies of string instruments, such as double basses, and the latter corresponding to the clarinet melodic line (essentially the fundamental frequencies of the played musical sounds) and to frequencies of the accompanying string instruments, such as violins (see figure 3.7). The Crest Factor (CF) (cf. section 2.1.4), measured by SpectraLAB 4.32.17 audio analysis software, is equal to 5.86 (a CF level of about 15.35 dB).

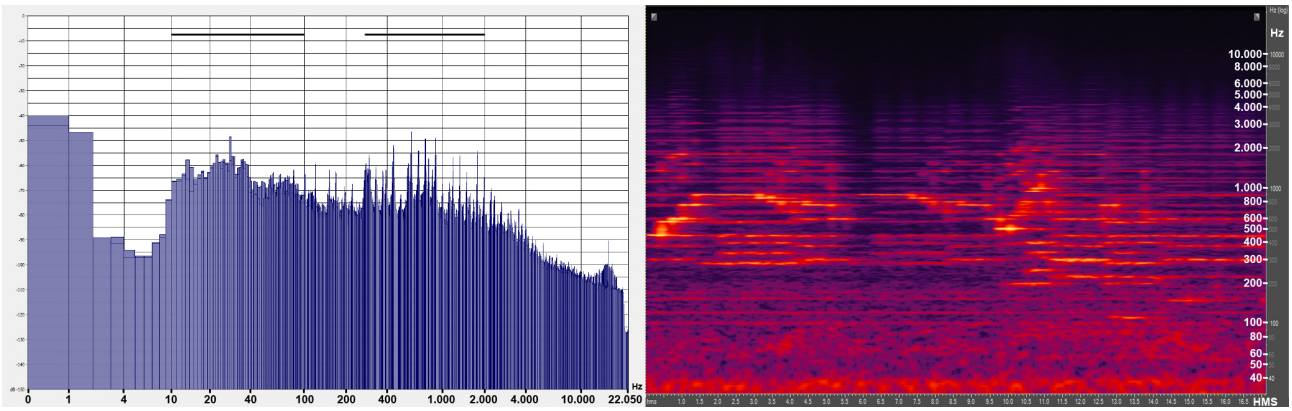


Figure 3.7: Long-time average spectrum (left) and spectrogram (right) of item 1 - Clarinet and Orchestra.

b) in item 2 (Double Bass), the region from about 50 Hz to 400 Hz stands out with higher energy and is associated with the melodic line (essentially the fundamental frequencies of the played musical sounds) (see figure 3.8). The CF is equal to 5.14 (a CF level of about 14.22 dB).

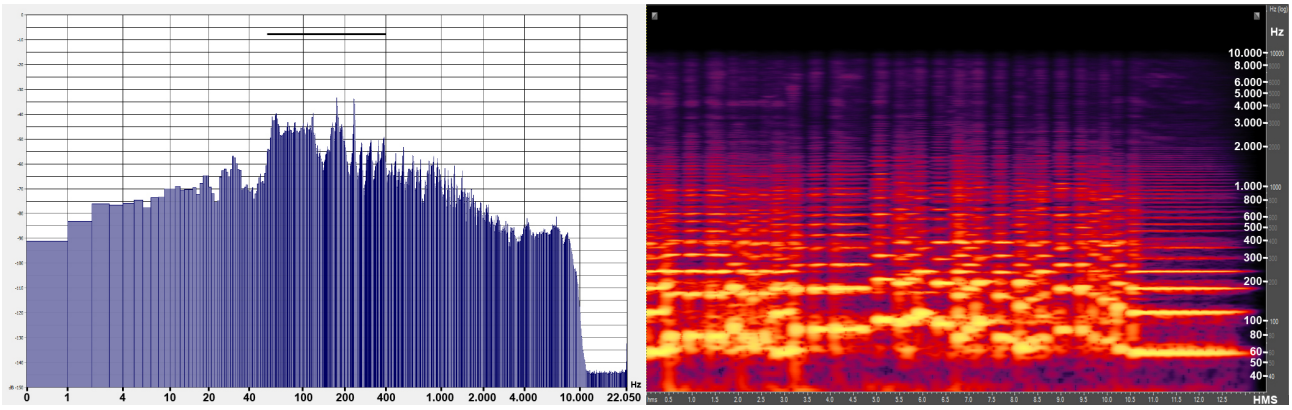


Figure 3.8: Long-time average spectrum (left) and spectrogram (right) of item 2 - Double Bass.

c) in item 3 (Harpsichord), most of the energy is distributed in a practically even way throughout the range from approximately 16 Hz to 2 kHz, as in pink noise (compare figure 3.9 with figure 3.25). The CF is equal to 5.29 (a CF level of about 14.47 dB).

d) in item 4 (Glockenspiel), three regions with higher energy can be distinguished, one around 60 Hz, relative to low frequency resonances of the body of the instrument when a steel bar is hit, another from 130 Hz to 400 Hz due to the same reasons, and the third one from more or less 1.7 kHz to 10 kHz relative to the melodic line (see figure 3.10). The CF is equal to 24.88 (a CF level of about 27.92 dB).

e) in item 5 (Trumpet and Piano), two regions are relevant, one from about 5 Hz to 200 Hz, mostly related to the left hand piano part (essentially the fundamental frequencies of the played musical sounds), and another

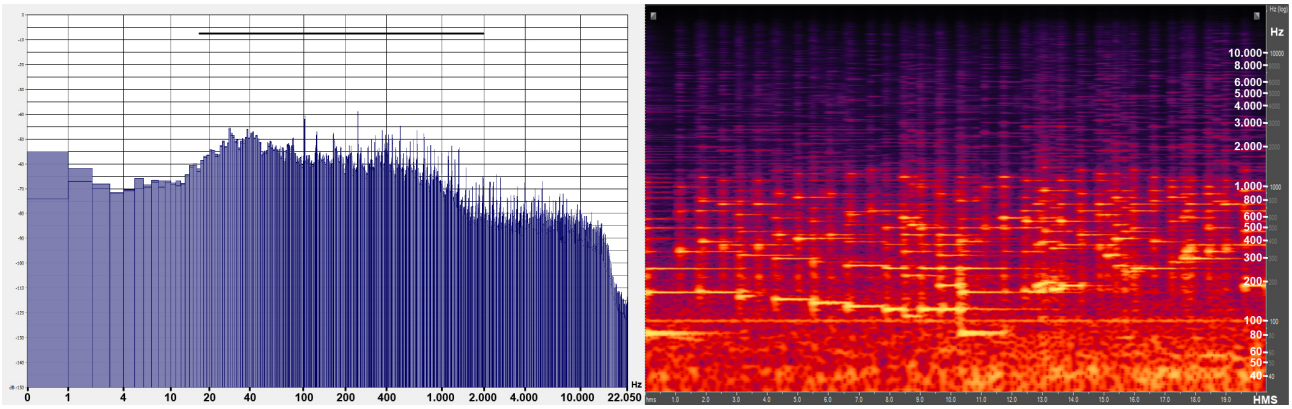


Figure 3.9: Long-time average spectrum (left) and spectrogram (right) of item 3 - Harpsichord.

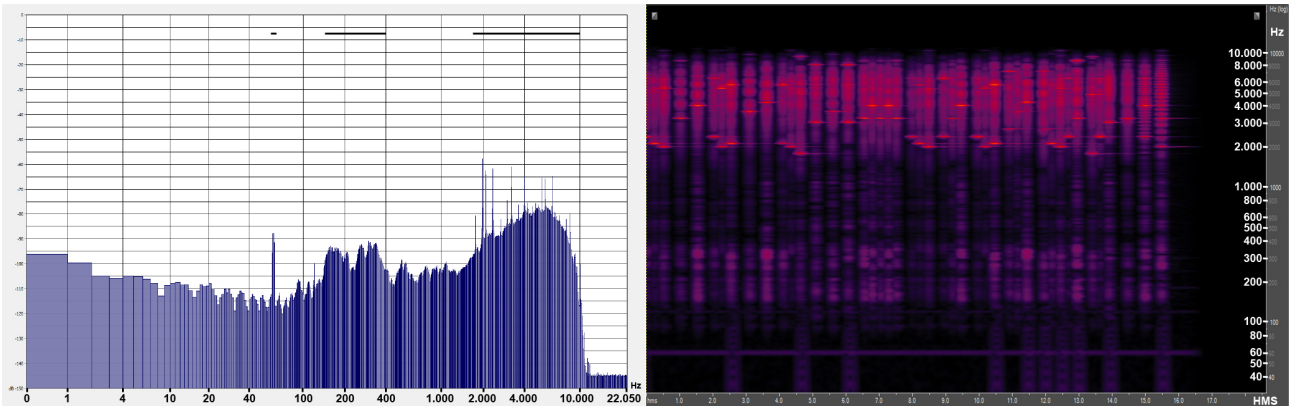


Figure 3.10: Long-time average spectrum (left) and spectrogram (right) of item 4 - Glockenspiel.

from around 200 Hz to 3 kHz, for the most part associated with the trumpet's melodic line and the right hand's piano part (see figure 3.11). The CF is equal to 12.44 (a CF level of about 21.90 dB).

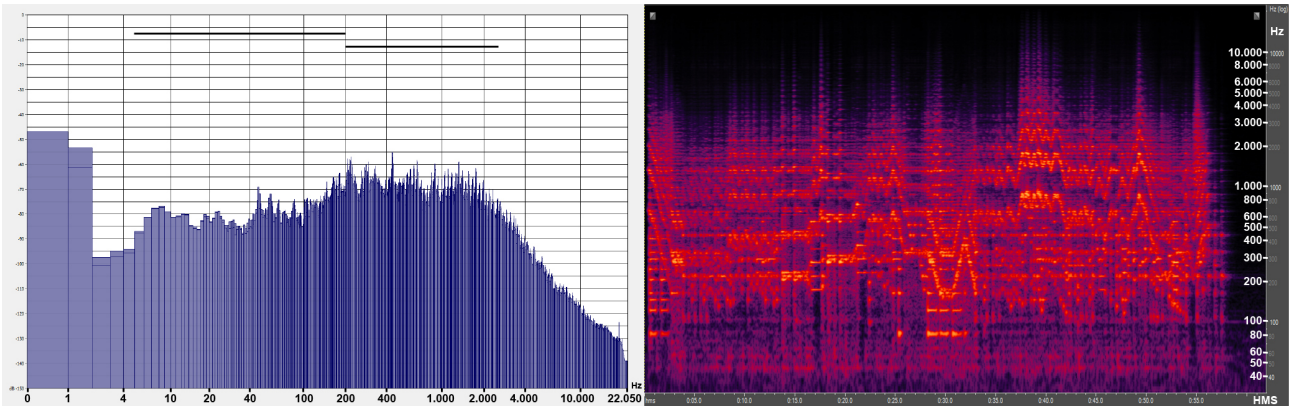


Figure 3.11: Long-time average spectrum (left) and spectrogram (right) of item 5 - Trumpet and Piano.

f) in item 6 (Marimba), the energy is essentially found around 18 Hz, relative to low frequency resonances of the body of the instrument when a bar is hit, and in the region of about 300 Hz to 5 kHz, relative to the melodic line (see figure 3.12). The CF is equal to 9.30 (a CF level of about 19.37 dB).

g) in item 7 (Oboe and Orchestra), the energy is higher in the regions from approximately 20 Hz to 120 Hz (relative to the low frequency range of string instruments, such as double basses) and from about 150 Hz to

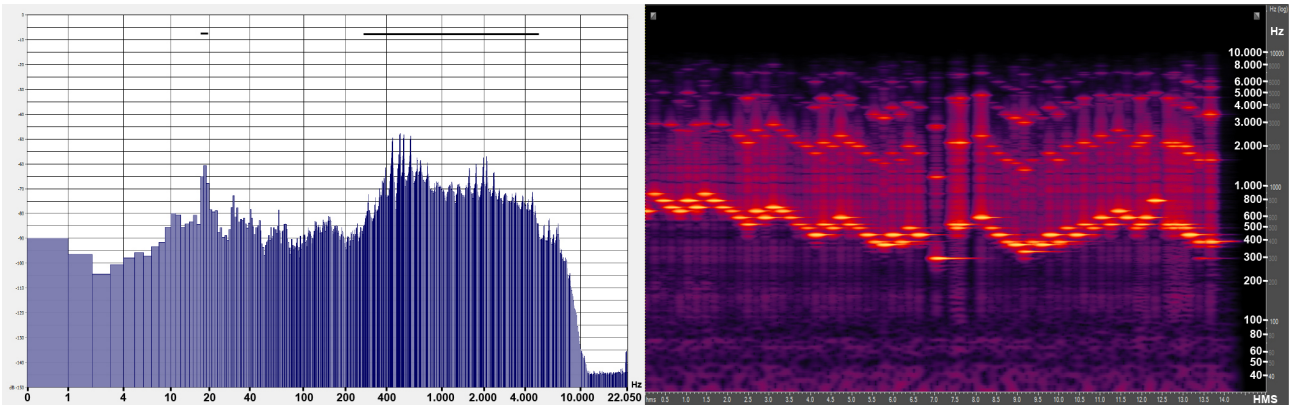


Figure 3.12: Long-time average spectrum (left) and spectrogram (right) of item 6 - Marimba.

3 kHz (essentially due to the accompanying part of the string instruments, such as violins, and to the oboe's melodic part, whose fundamental frequencies start at around 440 Hz) (see figure 3.13). The CF is equal to 6.66 (a CF level of about 16.47 dB).

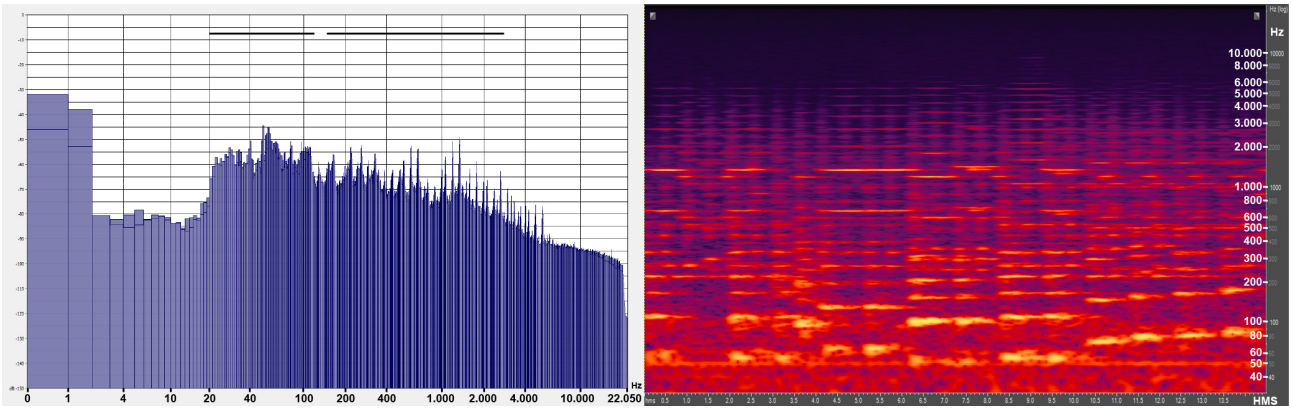


Figure 3.13: Long-time average spectrum (left) and spectrogram (right) of item 7 - Oboe and Orchestra.

h) in item 8 (Piano), whereas the region from approximately 10 Hz to 200 Hz, associated with the low frequency range of the piano and some background noise, has low energy but is almost even throughout it, the region from about 200 Hz to 3 kHz stands out with higher energy and is associated with the melodic lines of both hands (essentially the fundamental frequencies of the played musical sounds) (see figure 3.14). The CF is equal to 6.59 (a CF level of about 16.38 dB).

i) in item 9 (Tubular Bells), the energy is essentially present at more or less 18 Hz, relative to low frequency resonances when the pipes are hit, and in the region of approximately 200 Hz to 8 kHz (see figure 3.15). The CF is equal to 6.10 (a CF level of about 15.71 dB).

j) in item 10 (French Horn), the energy is higher in two regions, one from more or less 35 Hz to 45 Hz, due to background noise, and another from about 300 Hz to 1.2 kHz, relative to the melodic line (see figure 3.16). The CF is equal to 6.67 (a CF level of about 16.48 dB).

k) in item 11 (Vibraphone), there is basically one region with higher energy (essentially the fundamental frequencies of the played sounds from about 250 Hz to 750 Hz), although low frequency resonances with less energy can also be detected when the bars are hit (from around 20 Hz to 40 Hz) (see figure 3.17). The CF is

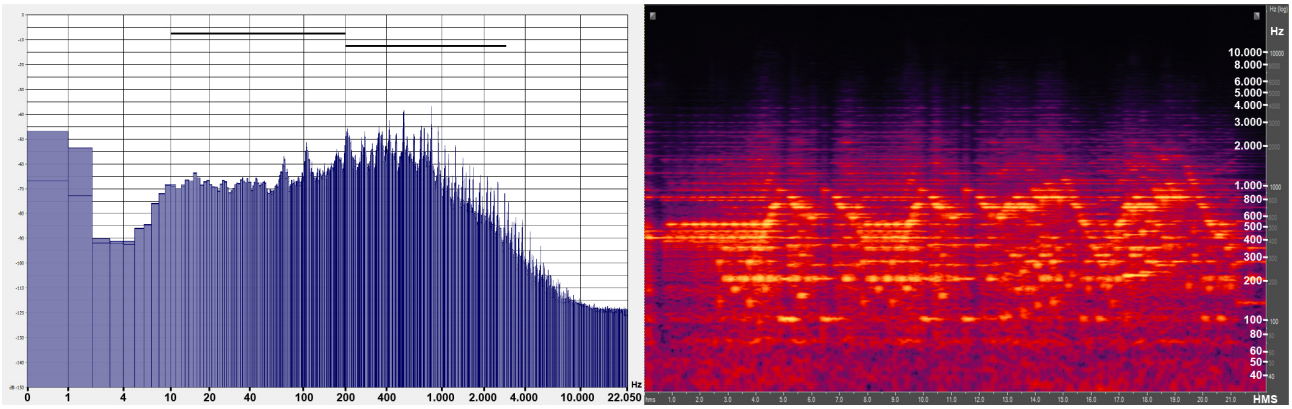


Figure 3.14: Long-time average spectrum (left) and spectrogram (right) of item 8 - Piano.

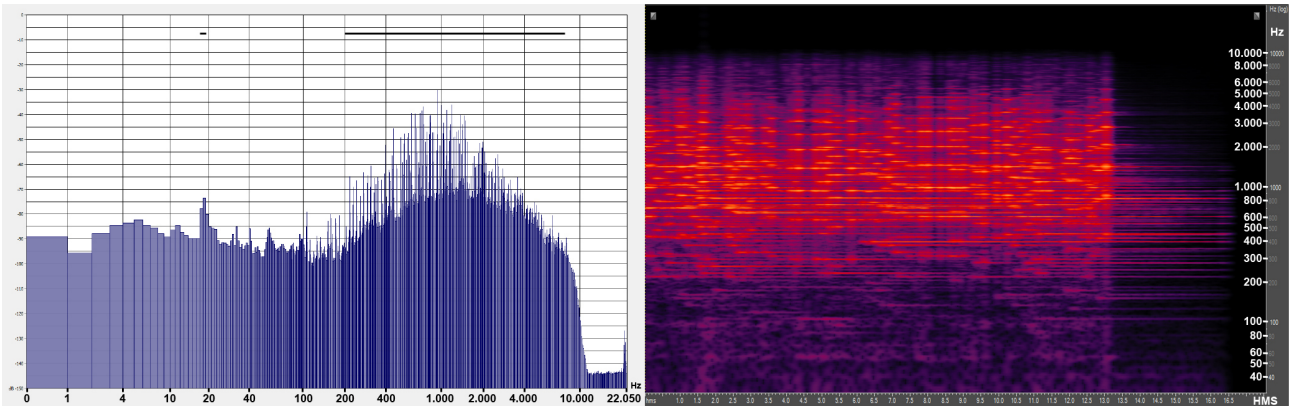


Figure 3.15: Long-time average spectrum (left) and spectrogram (right) of item 9 - Tubular Bells.

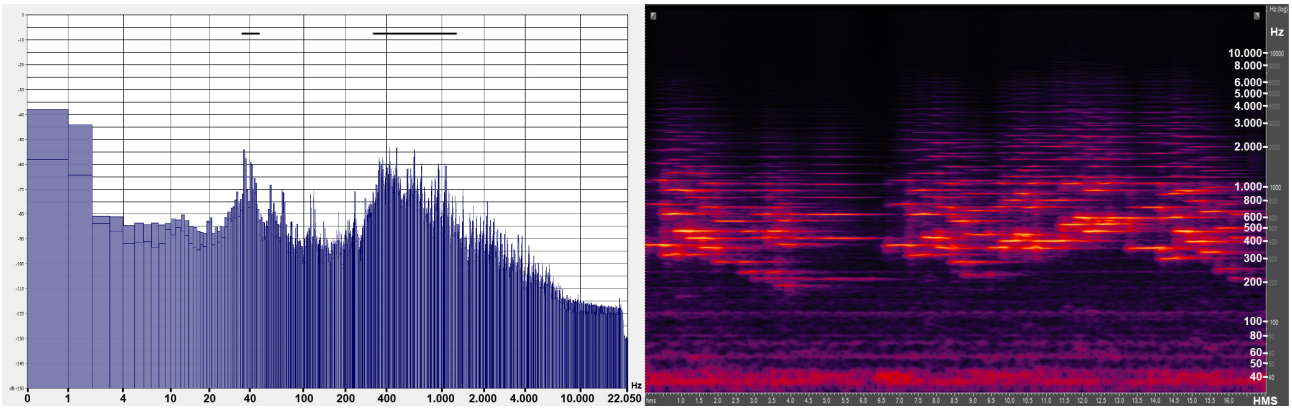


Figure 3.16: Long-time average spectrum (left) and spectrogram (right) of item 10 - French Horn.

equal to 6.81 (a CF level of about 16.66 dB).

l) in item 12 (Violin), while the region from approximately 10 Hz to 50 Hz is associated with a very present background noise, as well as the frequency around 120 Hz, the region from more or less 500 Hz to 5 kHz is related to the violin's melodic line, although its energy tends to be lower in the upper end (see figure 3.18). The CF is equal to 8.29 (a CF level of about 18.37 dB).

m) in item 13 (Xylophone), whereas the energy is maximum in the range from about 780 Hz to 9 kHz, which corresponds to the melodic line, noise due to defective sound recording can be detected around 60 Hz, 120

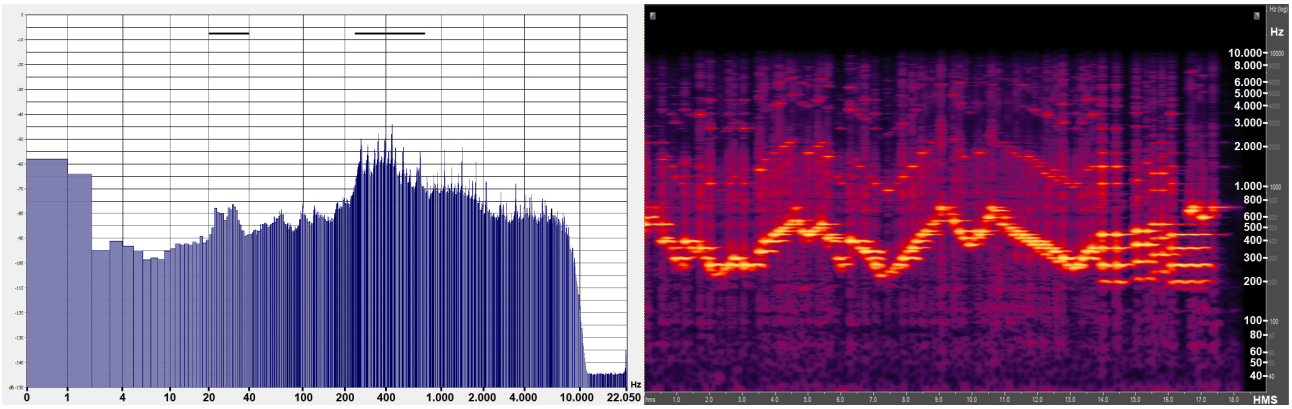


Figure 3.17: Long-time average spectrum (left) and spectrogram (right) of item 11 - Vibraphone.

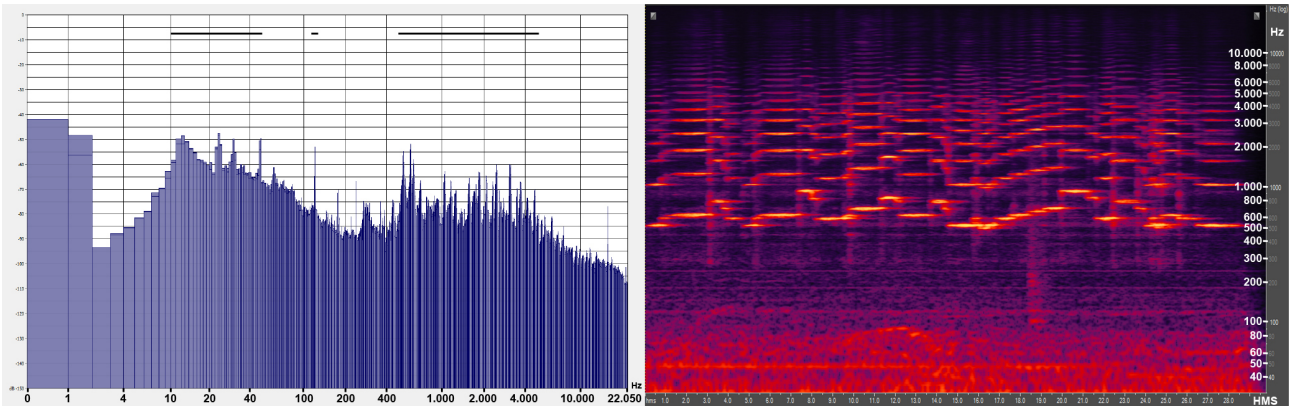


Figure 3.18: Long-time average spectrum (left) and spectrogram (right) of item 12 - Violin.

Hz, and 180 Hz (see figure 3.19). The CF is equal to 23.84 (a CF level of about 27.55 dB).

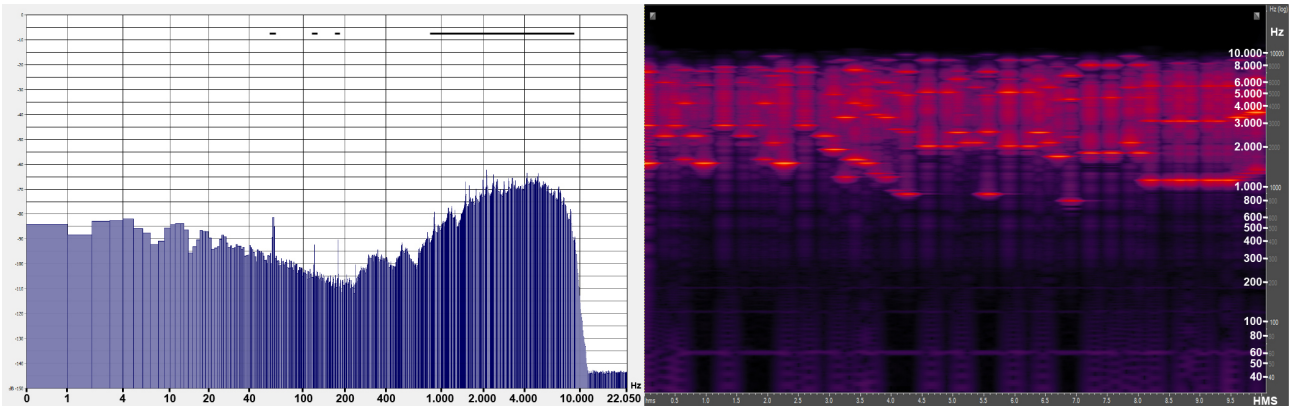


Figure 3.19: Long-time average spectrum (left) and spectrogram (right) of item 13 - Xylophone.

n) in item 14 (Piano), the energy is mostly present from around 250 Hz to 8 kHz (see figure 3.20). The CF is equal to 19.31 (a CF level of about 25.72 dB).

o) in item 15 (Long Swells), the energy is maximum in the region from approximately 125 Hz to 200 Hz, decreasing from then on to the upper limit of the audible spectrum (see figure 3.21). The CF is equal to 5.60 (a CF level of about 14.96 dB).

p) in item 16 (Chill Detune), the result is similar to that of item 15, although less energy than in item 15 can

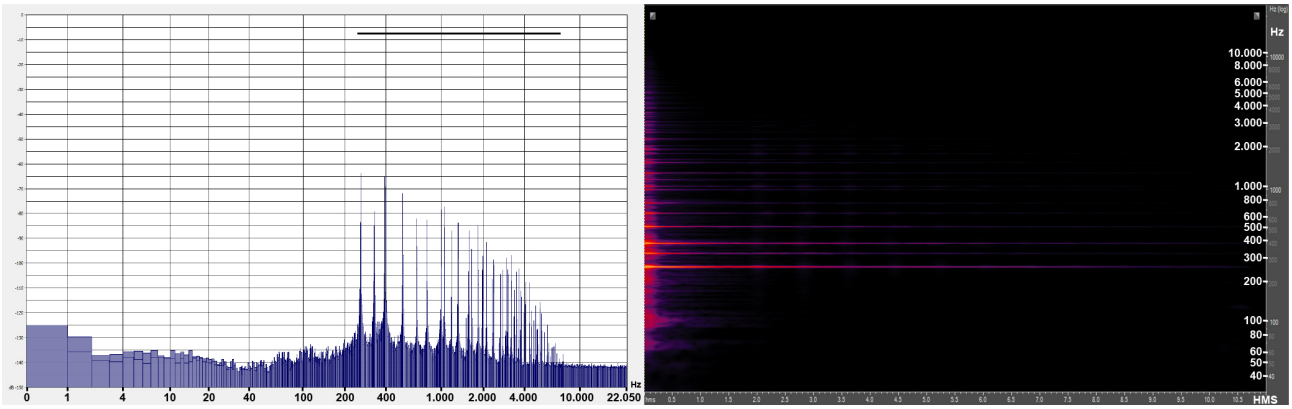


Figure 3.20: Long-time average spectrum (left) and spectrogram (right) of item 14 - Piano.

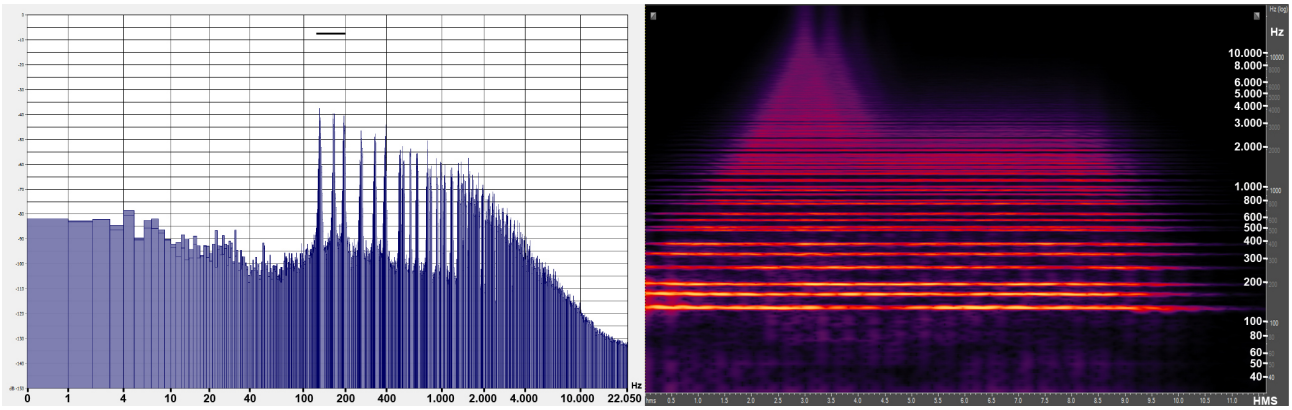


Figure 3.21: Long-time average spectrum (left) and spectrogram (right) of item 15 - Long Swells.

be detected from around 30 Hz to 100 Hz, related to the low frequency range of the sound (see figure 3.22). The CF is equal to 4.78 (a CF level of about 13.59 dB).

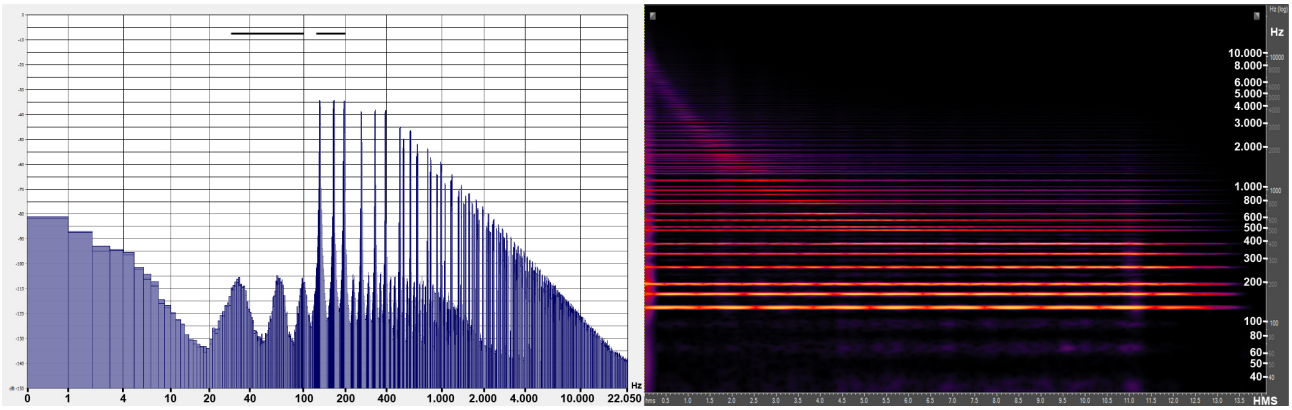


Figure 3.22: Long-time average spectrum (left) and spectrogram (right) of item 16 - Chill Detune.

q) in item 17 (Ultimate Trance), the region with higher energy lies in the same range as in items 15 and 16, but it decreases from then on to more or less 2.5 kHz (see figure 3.23). The CF is equal to 18.19 (a CF level of about 25.20 dB).

r) in item 18 (Orchestra), the energy is higher in the range from approximately 80 Hz to 600 Hz, followed by a range from 10 Hz to 50 Hz (see figure 3.24). The CF is equal to 6.29 (a CF level of about 15.97 dB).

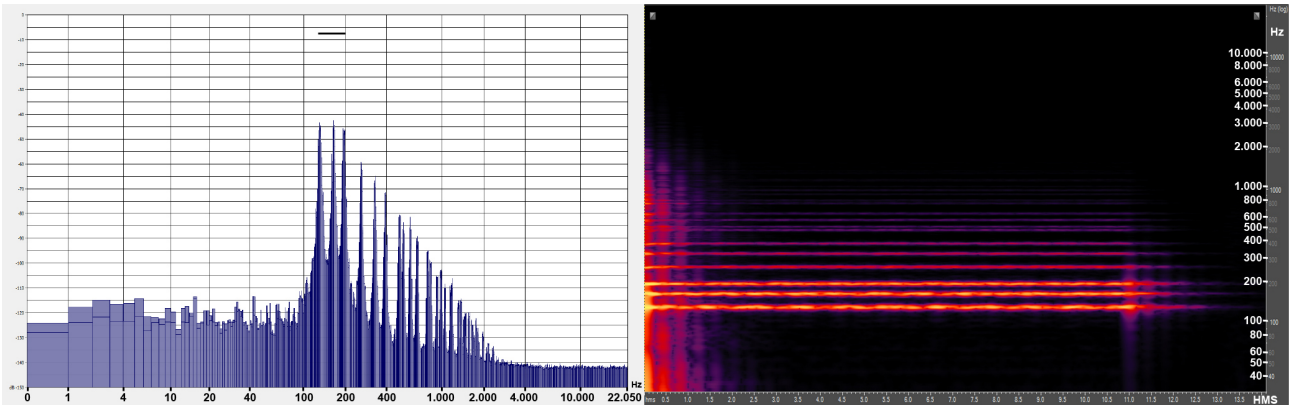


Figure 3.23: Long-time average spectrum (left) and spectrogram (right) of item 17 - Ultimate Trance.

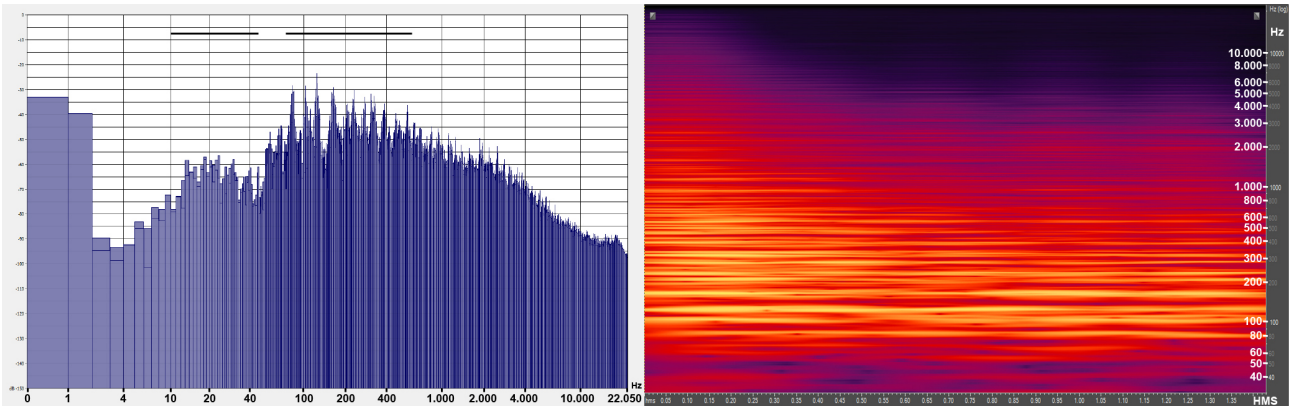


Figure 3.24: Long-time average spectrum (left) and spectrogram (right) of item 18 - Orchestra.

Compared with pink noise (see figure 3.25 and section 2.1.4), items 1, 3, 5, 7, 8, and 18 have a fairly large bandwidth as well, offering many localization cues to a listener, as already explained in section 2.1.4. With regard to items 2, 4, 6, 9, 11, and 13, the bandwidth goes up to approximately 10 kHz. The LTAS of items 14, 15, 16, and 17 show mainly the energy related to the discrete partials of the chords consisting of three notes. In addition, the CF values of all items are higher than that of the normalized pink noise described in section 3.2, which is typical of musical signals in which the dynamic range, that is, "... the number of decibels between the peak level and the noise floor, indicating the 'maximum-to-minimum' range of signal levels which may be handled..." (Rumsey & McCormick, 1997, p. 367), is also relatively large (cf. section 2.1.4).

Furthermore, all complex sounds used in our study contain frequencies below 5 kHz on the melodic lines, which is consistent with the idea that the sense of melody is evoked below that limit, except in item 9 (Tubular Bells), whose spectrum presents much more non-harmonic components, although it is still possible to perceive pitch, but in a much lesser extent (cf. section 2.1.3.2). According to the directional bands described in section 2.5.2, the analysis of the items shows that they all tend to be heard from the forward direction if reproduced frontally in the median plane, although items 4 (Glockenspiel) and 13 (Xylophone) also tend to be heard from a slightly upward direction due to the presence of a frequency region with higher energy around the directional band of upward direction perception of approximately 7 kHz to 9 kHz. In addition, the frequency spectrum varies over time in every item, as well as the transient times (cf. section 2.1.3.1): the initial transient times are of approximately 5 ms in items 9 and 13; around 15 ms in item 11; 20 ms in items 4 and 8; 30 ms in item 6; 30

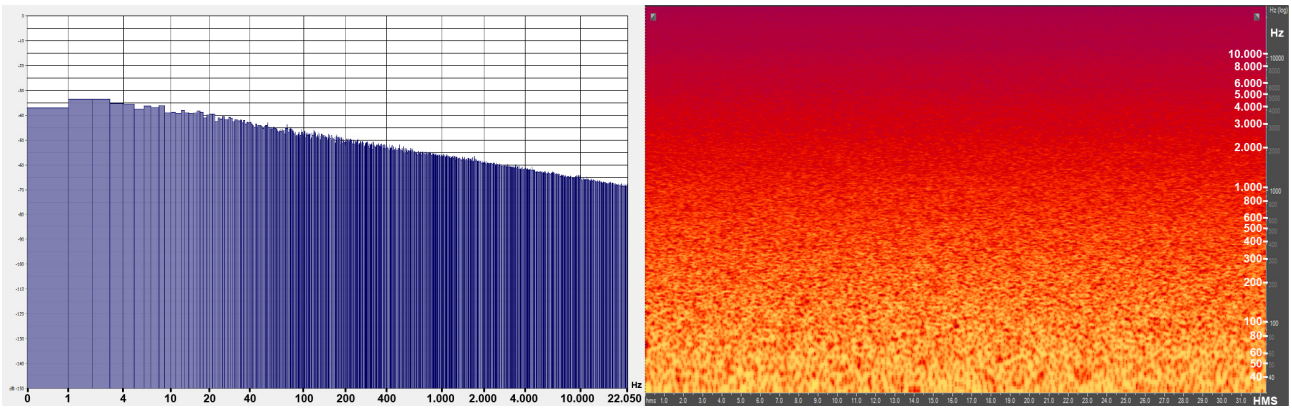


Figure 3.25: Long-time average spectrum (left) and spectrogram (right) of pink noise.

to 500 ms in item 12; 40 to 900 ms in item 16; 50 ms in items 3, 5, and 14; 90 ms in item 18; 100 ms in items 1, 2, 7, and 17; 150 ms in item 10; and 300 ms to 3 s in item 15.

Taking the characteristics of the above mentioned items into account and knowing that the auditory motion perception is more accurate for frontal incidence, for horizontal movements, for broadband sounds with short transient times and varying frequency spectrum over time (Marentakis & McAdams, 2013, p. 2) (see sections 2.2.4 and 2.5), we therefore considered at the time of their selection and still consider currently that they gather the conditions to be used in our sound localization tests.

3.5 Gestures Used in the Research

The type of gesture that has been chosen to be used in our study is essentially a relatively fast downward vertical to horizontal deictic (pointing) gesture, in which the stretched index finger is used (see section 2.6). The space where this gesture preferably occurs is on the periphery of a person's body, as has already been referred to and shown in figure 2.49, in section 2.6. From the computational point of view, it is easy to deal with this trajectory-based primitive, so that thresholds can be used to detect any changes (cf. section 2.6.1), as explained in section 3.7, demanding little use of memory and resources.

This empty-handed gesture, described as having a semiotic function and being therefore a non-instrumental gesture (see sections 2.6 and 2.6.1), is intended for controlling the motion of sound in space. As explained in section 2.7, the consequence of this choice is that it provides fundamentally freedom of movements to the performer, forcing him or her to rely mainly on aural cues, as is desired in our research.

3.6 Experimental Methods

In order to test the main hypothesis of our research, presented in section 1.2, that there is a significantly high relationship between a deictic gesture (see section 2.6) and localization of perceptual sound sources in space (cf. section 2.5), three different practical experiments were designed and developed by us specifically for this purpose for a single user at a time only, the data results and analysis of which will be described and presented in chapter 4. This set of experiments was firstly thought to represent a within-subjects or repeated-

measures design, since the same participants would collaborate in all three experiments (Field, 2009, pp. 15, 317, 458). This option was due to the fact that a repeated-measures design has more power to detect effects of an experimental manipulation by the experimenter than a between-groups, between-subjects, or independent design, in which different groups of participants are used in each experimental condition, because the unsystematic variation, that is, "... small differences in performance created by unknown factors" that exist between the experimental conditions (Field, 2009, p. 16), "... is kept to a minimum and so the effect of the experiment is more likely to show up" (Field, 2009, p. 17). However, this repeated-measures design did not invalidate the possibility of comparing groups of different participants among all who participated in all three experiments, such as the group of participants without musical knowledge and the group of volunteers with musical knowledge, as described and explained in section 4.2.1. In addition, our role would be that of an observer who would gather information about an observed phenomena being studied, but in which we would not take part.

Randomization of participants, that is, the random order by which a volunteer would participate in all three experiments in order to minimize unsystematic variations (Field, 2009, p. 17), was not taken into account for practical reasons: we wanted a participant to adjust him- or herself first to experiment 1, which served as a control experiment, on purpose, since we considered at that time (cf. the beginning of chapter 3) that it would be a new and unusual experience for him or her, because experiences of this kind are not commonly performed, so that we could observe in a clearer way if he or she would detect any differences or react differently in the two following experiments relatively to the first one. Experiment 2 would follow the first one, with the aim of understanding if the same participant would perceive any differences while the direction of sound would be completely opposite to that in experiment 1, as explained later in this section. Experiment 3 would be completed at the end, in order to evaluate the participant's hearing ability together with the use of his or her deictic gesture after being familiarized with the interactive system.

Although experimental methods commonly "... provide a comparison of situations (usually called *treatments* or *conditions*) in which the proposed cause is present or absent" (Field, 2009, p. 14), in our case the comparison has not got to do exactly with a proposed cause being present or absent, but it has rather got to do with a change or manipulation of it. Thus, we manipulated the direction of sound and we then expected participants to use their hearing skills and their deictic gesture to interact with, and to define the origin of, sound. This was done, so that any confounding variables could be excluded. In order for the latter to happen, "... an effect should be present when the cause is present and [...] when the cause is absent the effect should be absent also" (Field, 2009, p. 14). Thus, in our case an effect should be present when the manipulation of the direction of sound was present and when the manipulation of the direction of sound was absent the effect should also be absent. The direction of sound is therefore the proposed cause or the independent variable (cf. section 2.1.4), "... because its value does not depend on any other variables" (Field, 2009, p. 7), whereas the proposed effect or outcome, or dependent variable, "... because the value of this variable depends on the cause..." (Field, 2009, p. 7), that we were interested in measuring, is the ability of perceptual sound source localization by participants together with their deictic gesture. The latter is therefore defined as the percentage of correct perceptual sound source localizations resulting from the participants' responses in the second part of the Inquiry Mode Questionnaire (see appendix A) in comparison with the actual directions of the deictic

gesture and perceptual sound sources given by the software used in our research (cf. section 3.7.1).

All experiments require the participant to stand preferably at the centre of the already shown circular arrangement of the interactive system and to face the depth camera at first (see figure 3.1 in section 3 and figure 3.3 in section 3.2).

Standing at the centre of the system is due to the fact that the listening position should always be at the same distance from the loudspeakers as possible to guarantee the best listening conditions offered by Ambisonics (Frank et al., 2008) (for more details cf. section 2.3.4.7), although the chosen Ambisonics' order of 4.64, explained in sections 2.3.4.7, 3.3, and 3.7.1, increases the size of the listening area (Stitt et al., 2013, p. DAFX-6).

Facing the camera with a ”Psi” or ”submissive” posture at first (Borenstein, 2012, p. 192) (see figure 3.26), that is, a standing posture (see section 2.6.1) in which the feet are close to each other and the hands are raised above the shoulders on the sides of the head, is due to the fact that the system has to detect and calibrate the user, so that his or her joint data are available (cf. section 3.7.1). Once the system starts tracking one of the hands, the user can freely move his or her whole body, legs, arms, and head around the centre by 360 degrees, but hands have to be always visible to the camera.

Figure 3.26: The ”Psi” posture necessary for a participant's detection and calibration.

The same type of deictic (pointing) gesture, described in section 3.5, is used in the three experiments, so that in:

Experiment 1: one can trigger a sound, chosen randomly by MAX MSP Jitter software (see section 3.7.1) from the set of sounds described in section 3.4, making use of this gesture with one of the hands in a certain direction. By doing so, one shall be able to hear the sound from that direction (see figure 3.27a). Then, one shall also be able to move the hand horizontally in a continuous way and control the motion of sound in the horizontal hearing plane (see section 2.3.2) if no other sound has been triggered again. The hand is considered to be in the horizontal plane if the vertical coordinate `coordY0` of the green arrow representing the direction of the deictic gesture is within the range of -0.5 to 0.6 (cf. section 3.7.1), which has been adjusted experimentally by us and confirmed in pilot experiments carried out with the special participation of Inês Franco — the participant's gesture determines the direction of a perceptual sound.

Experiment 2: one can trigger a sound, chosen randomly by MAX MSP Jitter software from the same set of sounds as in experiment 1, making use of this gesture with one of the hands in a certain direction. By doing so, one shall be able to hear the sound coming from the opposite surround direction (see figure

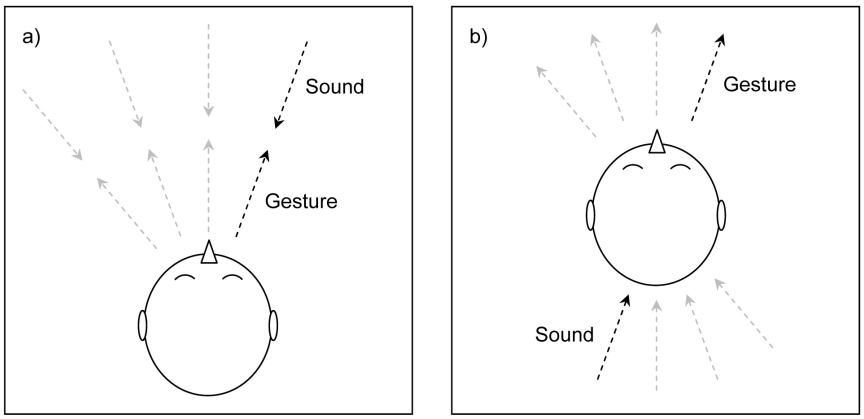


Figure 3.27: Directions of gesture and sound in: a) experiments 1 and 3; b) experiment 2.

3.27b). Afterwards, one shall also be able to move the hand horizontally in a continuous way and control the motion of sound from the opposite surround direction in the horizontal hearing plane if no other sound has been triggered again. As in experiment 1, the hand is considered to be in the horizontal plane if the vertical coordinate `coordY0` of the green arrow representing the direction of the deictic gesture is within the range of -0.5 and 0.6 — the sound follows the participant's triggering gesture from the opposite direction.

Experiment 3: one shall hear a looped sound, chosen randomly by MAX MSP Jitter software from a set of sounds equal to those used in experiments 1 and 2, originating from an also arbitrarily Processing-software-determined surround fixed direction (see section 3.7.1) between -40 and 220 degrees (see figure 3.28). In this particular experiment, the horizontal localization angle has been thought to be smaller and has been determined experimentally by us and confirmed in pilot experiments, once again carried out with the participation of Inês Franco, in order to facilitate the position necessary to have the hands always visible to the camera from 221 to 319 degrees (shadow of 27.5%). Each time one tries and manages to identify the sound's localization by the above suggested gesture within an also experimentally determined margin of $\pm$15 degrees ($\approx \pm$4.2% of 360 degrees), Processing tells the researcher that the participant localized the sound correctly, MAX MSP Jitter stops this sound, and another sound shall originate afterwards again from another random direction after two seconds (see figure 3.27a). This angle of 30º (from -15º to +15º) allows us to later measure the error that the participant may have made in the identification of the software-predetermined angle with this gesture (cf. algorithm 3.24 in section 3.7.1.3) — the participant tries to identify the system-predetermined localization of a perceptual sound source with his or her gesture.

In all three experiments, the type of gesture mapping strategy that has been applied is a one-to-one or direct mapping strategy, as explained in section 2.8. The horizontal coordinates and angle of the direction of a left or right hand deictic gesture, which were calculated by the Processing software (see section 3.7.1) after the gesture was captured by the depth camera (cf. section 3.2), were tested and compared experimentally by Inês Franco and by us with the actual direction of the gesture and with the origin of a perceptual sound source at the position of each of the eight loudspeakers, whose level should be maximum there and that was shown

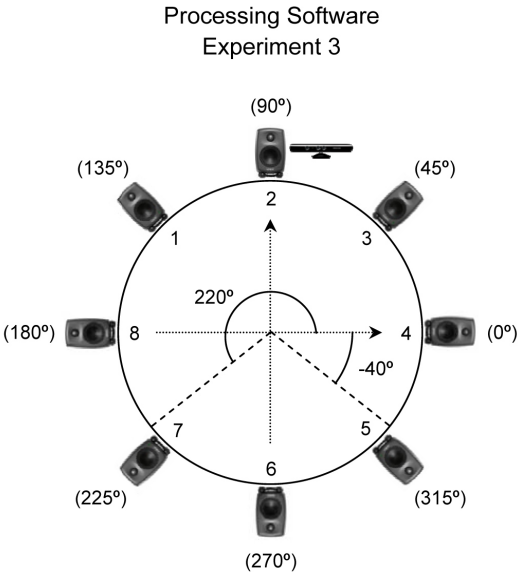


Figure 3.28: Experiment 3: the direction of a perceptual sound source is determined by Processing software between -40 and 220 degrees.

to us as such by the Max MSP software, so that we could determine a spatial reference where the gesture and the origin of sound would coincide, as mentioned in sections 3.2 and 3.7.1.4. This was accomplished using a Parkside Laser-type Spirit Level 670 device, already referred to in section 3.2, attached to a stretched index finger and respective forearm, in order to force that finger to be as parallel as possible to the forearm (see figure 3.29), thus attempting to minimize pointing errors during this process, taking into account that the space where this gesture occurs is on the periphery of a person's body (cf. sections 2.6 and 3.5).

Figure 3.29: Parkside Laser-type Spirit Level 670 device attached to the right hand stretched index finger and respective forearm.

3.6.1 Procedure in Experiment 1

Before starting experiment 1, each participant was informed of the general purpose of the research and of the approximately total duration of the experiments altogether, which was of about 10 to 15 minutes. Additionally,

the participant was also informed that after experiment 3 he or she would be asked to fill in an anonymous Inquiry Mode Questionnaire (InQ) (see appendix A), which is an indirect measuring tool of variables (Field, 2009, p. 10), with the aim of gathering data, that we could not directly measure, for a subsequent statistical analysis.

The participant was in detail informed of the standing and calibrating positions as well, explained in section 3.6, of the gesture to be used to trigger a sound from an expected direction (cf. section 3.5), and of the allowed free body, legs, arms, and head rotating movement around the loudspeaker array's centre to control the motion of sound in the horizontal hearing plane, keeping both hands visible to the camera whenever possible. Then, experiment 1 was loaded into the computer and its full screen information was recorded since then until the end of experiment 3 (see section 3.7.1 and appendix E.7). Afterwards, the subject was asked to choose one of the hands that would be used in all exercises.

As soon as the depth image and the participant's joint data were made available by the system, a green arrow representing the direction of the participant's deictic gesture appeared on the computer screen and the experiment could be finally concluded after our indication. The experiment had the duration of about three minutes.

3.6.2 Procedure in Experiment 2

There has been about one minute interval between experiment 1 and experiment 2 for the participant, so that we could load experiment 2. The only information provided to the participant before he or she began experiment 2 has been that: 1) the same deictic gesture used in experiment 1 should be used again to trigger a sound; 2) he or she was allowed to use a free body, legs, arms, and head rotating movement around the loudspeaker array's centre again, in order to control the motion of sound in the horizontal hearing plane, keeping both hands visible to the camera whenever possible; and 3) the experiment would begin as soon as the depth image and the participant's joint data had been made available by the system after the standing and calibrating positions, and a green arrow, representing the direction of the participant's deictic gesture, had appeared on the observer's computer screen. The experiment had the duration of about three minutes.

3.6.3 Procedure in Experiment 3

There has been about one minute interval between experiment 2 and experiment 3 for the participant, so that we could load experiment 3. Before starting this last experiment, the participant was informed that: 1) the same deictic gesture used in experiments 1 and 2 should be used again, but this time in order to guess or identify the predefined localization of sound in space; 2) if the localization of a sound were identified with the participant's deictic gesture, sound would stop and another one would be produced randomly from another direction after about two seconds; 3) he or she was allowed to use a free body, legs, arms, and head rotating movement around the loudspeaker array's centre again, keeping both hands visible to the camera whenever possible; and 4) the experiment would begin as soon as the depth image and the participant's joint data had been made available by the system after the standing and calibrating positions, after a green arrow, representing the direction of the participant's deictic gesture, had appeared on the computer screen, and after a sound had

been produced from any direction. The experiment had the duration of about three minutes.

After experiment 3, the participant was asked to fill in an anonymous Inquiry Mode Questionnaire (InQ) (see appendix A), as already explained in section 3.6.1.

3.6.4 Research Questions and Hypotheses

In order to complement and support the main hypothesis that there is a significantly high relationship between a deictic gesture (cf. section 2.6) and localization of perceptual sound sources in space (see section 2.5), other research questions were raised and respective hypotheses were formulated, which are presented in the next sections.

3.6.4.1 Experiments 1, 2, and 3 - Questions and Hypotheses

The three experiments, whose descriptions can be found in section 3.6, address the following questions:

Q1. Which of the three experiments do participants understand more quickly?

Q2. In which of the three experiments do participants estimate shorter times needed to understand them?

Q3. In which of the three experiments do participants interact more easily with the installation?

Q4. In which of the three experiments do participants define the origin of sound more easily?

Q5. In which of the three experiments do participants assign a higher level of adequacy of the suggested gesture to the experiment?

Q6. In which of the three experiments do participants most likely reckon that the system's response to gesture is immediate?

Q7. In which of the three experiments do participants estimate shorter times of the system's response to gesture?

Q8. In which of the three experiments do participants more firmly consider that they feel surrounded by sound in the installation?

Q9. In which of the three experiments do participants more firmly consider that their gesture coincides with the origin of sound?

Thus, taking into account that we have one independent variable consisting of three different experimental conditions, in which the same participants are involved, and that the dependent variable consists of an answer on a 5-point scale or of estimated times (Field, 2009, p. 7), we hypothesized that:

H1. there are no significant differences in how quickly participants understand the three experiments, but that they rate their quick understanding highly;

H2. there are no significant differences in the estimation of time it takes for participants to understand the three experimental conditions, but that they take less than 2 seconds to understand them;

H3. there are no significant differences in how easily participants interact with the installation in the three experiments, but that they rate their ease of interaction highly;

H4. participants define the origin of sound more easily in the first and then in the third experimental condition;

H5. there are no significant differences in how highly participants rate the suggested gesture as being adequate to any of the three experiments, but that they rate its adequacy highly;

H6. there are no significant differences in the degree of appreciation whether the system's response to gesture is immediate in any of the three experiments, but that participants more likely consider that the system's response to gesture is immediate;

H7. there are no significant differences in the estimation of the shorter times of the system's response to gesture in the three experiments;

H8. there are no significant differences in how strongly participants feel surrounded by sound in the installation, but that they rate this feeling highly;

H9. participants more firmly consider that their gesture coincides with the origin of sound in the first and then in the third experimental conditions, and that they do not coincide at all in the second.

3.6.4.2 Experiment 1 - Questions and Hypotheses

This experiment, whose description can be found in section 3.6, addresses the following questions:

Q10. Is there any relation between sensing that Experiment 1 is quickly understood and the estimated time needed to understand it?

Q11. Is there any relation between fancying that immediate control over sound is felt and the estimated time needed to control sound in Experiment 1?

Thus, we hypothesized that if participants have the control over the expected perceptual sound source's direction in the horizontal plane, whose sound arises from the front towards the listener, with a congruent deictic gesture pointing from the listener towards the perceptual sound source, then:

H10. the estimated time needed to understand Experiment 1 is less than 2 seconds if participants consider that they quickly understand it;

H11. participants need a time of less than 1 second to control sound in Experiment 1 if they more likely consider that they feel immediate control over it.

3.6.4.3 Experiment 2 - Questions and Hypotheses

This experiment, whose description can be found in section 3.6, addresses the following questions:

Q12. Is there any relation between considering that Experiment 2 is quickly understood and the estimated time needed to understand it?

Q13. Is there any relation between admitting that it is easy to define the origin of sound, assuming that the gesture does not coincide with the origin of sound, and presuming that the proposed experiment does not confuse anybody in Experiment 2?

Q14. Is there any relation between imagining that immediate control over sound is felt and the estimated time needed to control sound in Experiment 2?

Thus, we hypothesized that if participants have the control over the perceptual sound source's direction in the horizontal plane, whose sound arises from behind towards the listener without his or her prior knowledge, with a deictic gesture pointing forwards, then:

H12. the estimated time needed to understand Experiment 2 is less than 2 seconds if participants consider that they quickly understand it;

H13. participants most likely reckon that their gesture does not coincide with the origin of sound in Experiment 2 and that the proposed experiment does not confuse them if they are convinced that it is easy to define the origin of sound;

H14. participants estimate a time of less than 1 second to control sound in Experiment 2 if they find that they feel immediate control over it.

3.6.4.4 Experiment 3 - Questions and Hypotheses

This experiment, whose description can be found in section 3.6, addresses the following questions:

Q15. Is there any relation between assuming that Experiment 3 is quickly understood and the estimated time needed to understand it?

Q16. Is there any relation between considering that sound is quickly located and the estimated time needed to locate sound in Experiment 3?

Q17. Is there any relation between admitting that it is easy to define the origin of sound, considering that locating sound is achieved, and presuming that the gesture coincides with the origin of sound in Experiment 3?

Thus, we hypothesized that if participants are trying to localize sound in the horizontal plane, whose direction is predetermined by software, using their deictic gesture, then:

H15. the estimated time needed to understand Experiment 3 is less than 2 seconds if participants consider that they quickly understand it;

H16. participants estimate a time of less than 3 seconds to locate sound in Experiment 3 if they admit that they quickly locate it;

H17. participants are most likely to admit that their gesture coincides with the origin of sound in Experiment 3 if they mean that it is easy to define the origin of sound and assume that they manage to locate sound.

3.6.4.5 Experiments 1 and 2 - Questions and Hypotheses

These experiments, whose descriptions can be found in section 3.6, address the following questions:

Q18. In which of Experiments 1 and 2 will a participant feel a more immediate control over sound?

Q19. In which of Experiments 1 and 2 will a participant estimate a lower time needed to control sound?

Thus, we hypothesized that:

H18. participants feel more immediate control over sound in Experiment 1 than in Experiment 2;

H19. participants estimate a lower time needed to control sound in Experiment 1 than in Experiment 2.

3.7 Software Architecture

3.7.1 Computer Software Used in the Research

The programming language Processing 1.5.1[24] and the Cycling'74's visual programming MAX MSP Jitter 6.0.1 (50928)[25] software are used in our interactive installation. This software was chosen for practical reasons, since we already knew it reasonably well at the time when we began to develop the experiments. However, before we decided on this final solution, we had initially tried out other approaches to hand tracking, first in MAX MSP Jitter and then in Processing, too, in order to test their respective capacities and performances. Techniques, such as frame differencing, colour detection, brightness detection, background subtraction, blob extraction, movement estimation, and face detection, the descriptions of which are beyond the scope of our thesis, were tested with a Sony Handycam DCR-PC330E PAL camera[26], but the results were not robust enough for us and did not please us at all, essentially because of the environment conditions, which had to be highly controlled, such as the lighting.

Thus, based on the idea of the face detection technique, we tried to find a more reliable technique that would eventually enable the detection of a hand or finger. The Haar Feature-based Cascade Classifiers technique, a machine-learning-based technique used for object detection that was proposed by Viola & Jones (2001), where a lot of images of an object (positive images) and a lot of images without that object (negative images) are used to train a certain classifier, was tested accordingly, but we could not get any positive results as well. Other software, such as the open source C++ toolkit for creative coding called openFrameworks, was also explored, but its complexity would lead us to a more time-consuming work, so it was rejected. Fortunately, all these difficulties were overcome with relative ease with the discovery of the SimpleOpenNI library for Processing.

Thus, in our research, Processing makes use of the: 1) Open Natural Interaction (OpenNI) Application Programming Interface (API), version 1.5.4.0, which enables communication with vision and audio sensors and perception *middleware*, that is, "... software components that analyze the audio and visual data that is recorded from the scene, and comprehend it" ("OpenNI_UserGuide.pdf", installed with the software into the

[24] Retrieved 02/12/2011, from `http://processing.org`
[25] `http://cycling74.com`
[26] Retrieved 12/12/2016, from urlhttp://pdf.crse.com/manuals/3084864131.pdf

`Documentation` folder, p. 4); 2) PrimeSense's Natural Interaction Technology for End-user (NITE), version 1.5.2.21, which "... is the middleware that perceives the world in 3D, based on the PrimeSensor depth images, and translates these perceptions into meaningful data in the same way that people do" ("NITE Controls 1.3.1 User Guide.pdf", installed with the software into the `Documentation` folder, p. 6); and 3) SimpleOpenNI library[27], version 0.27, which is a software extension for Processing that provides access to all of the data from the Kinect depth camera, in our particular case for left or right hand motion capture.

Although the orientation data of the hands is more difficult to gather by OpenNI, it is this skeleton joint which is actually tracked with the help of the depth camera. According to Borenstein (2012, pp. 229–230, 236), it is much easier for OpenNI to determine the orientation of the inner joints on the skeleton that have two neighbours than the outer joints, such as the hands. The same occurs with the head and with the feet. Therefore, a vector is defined and the three-dimensional position of one of the hands, which is the hand previously selected by a user to be tracked, is stored into it by using `kinect.getJointPositionSkeleton`. In addition, the three-dimensional orientation of the respective elbow is also stored into a matrix by using `kinect.getJointOrientationSkeleton`. As the joints are represented by vectors, the hand position vector is transformed into a unit vector by calling `normalize`, so that its orientation becomes more meaningful. If the confidence, which OpenNI has in telling where it thinks the requested joint is currently positioned, is over 50%, then a green arrow corresponding to the direction of the deictic (pointing) gesture, described in section 3.5, is shown on an Open Graphics Library (OpenGL) computer window.

Furthermore, Processing provides other visual information on the computer screen for the researcher, such as the identification of the present practical experiment, the depth image of the participant, the horizontal angle and the coordinates of the orientation of the green arrow, and some testing movement recording and memory function short-cuts (see figure 3.30), the latter not being used during the experiments in order not to increase the overall latency time of the whole system even more, which is of approximately 470 milliseconds and that was determined with the help of Sony Vegas Pro 11.0 (build 700)[28] software (cf. appendix E.10). Although these testing movement recording and memory functions were not used during the experiments, we nevertheless developed two readers for the data recorded in files, which can be found in appendix E.5, one for experiments 1 and 2, and one for experiment 3. The deictic gesture was thought to trigger the audio signal, as explained in more detail in section 3.6.

In turn, MAX MSP Jitter 6.0.1 (50928) processes only the left channel of the stereo audio files presented in section 3.4 and calculates the signal's amplitude for each loudspeaker based on Ambisonics Equivalent Panning (AEP) (see section 2.3.4.7), according to the participant's left or right hand position and orientation, to the loudspeaker's position, and to the chosen number of loudspeakers, which in this case is equal to eight (cf. section 3.3). The signal's amplitude is usually calculated as follows:

$$\text{Signal Amplitude} = \left(0.5 + 0.5 \times \frac{x \times x_s + y \times y_s + z \times z_s}{r \times r_s}\right)^p ,$$

where x, y and z are the coordinates of a perceptual sound source's position, x_s, y_s, and z_s are the coordinates of a loudspeaker's position, r is the radius of the circumference containing the perceptual sound source's

[27] All three were initially retrieved 15/06/2012, from `http://simple-openni.googlecode.com/files/`; still recently (03/10/2016), they could be retrieved from `https://code.google.com/archive/p/simple-openni/downloads`

[28] http://www.sonycreativesoftware.com

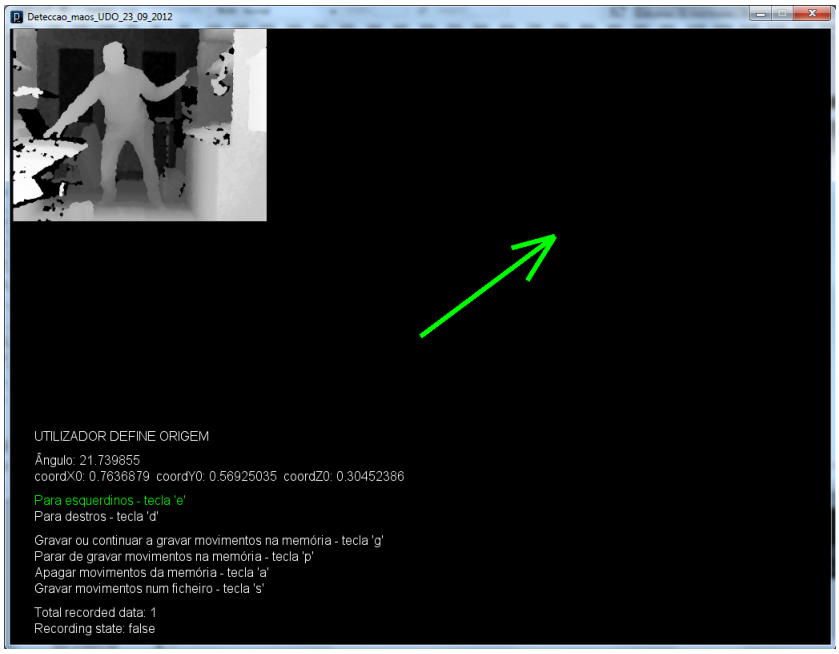


Figure 3.30: Processing visual information.

position, r_s is the radius of the circumference containing the loudspeaker's position, and p corresponds to the AEP's order (Neukom & Schacher, 2008). However, in Processing, the x-axis corresponds to the width, the y-axis to the height, and the z-axis to the depth measured by the Kinect camera. Thus, considering r and r_s equal to one, only the x, x_s and z, z_s coordinates are taken into account in the MAX MSP Jitter "level.maxpat" sub-patch (see figure 3.31), because these define the desired horizontal hearing plane:

$$\text{Signal Amplitude} = (0.5 + 0.5 \times (x \times x_s + z \times z_s))^p \ .$$

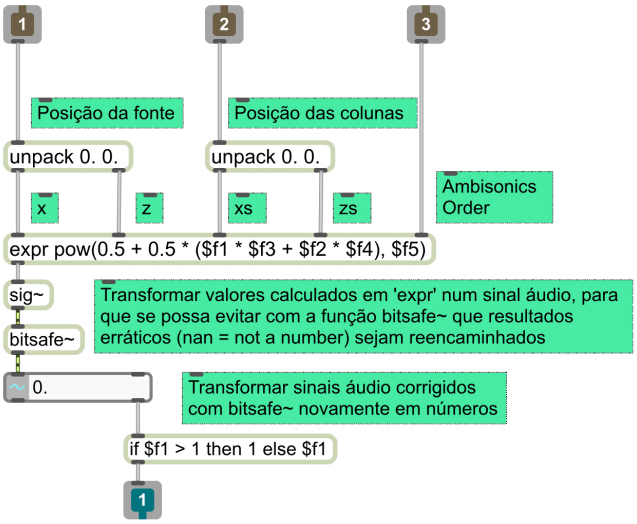


Figure 3.31: The MAX MSP Jitter "level.maxpat" sub-patch, which calculates the signal amplitude of a loudspeaker based on Ambisonics Equivalent Panning.

Since localization improves at higher ambisonics' orders (Frank et al., 2008) and given that in AEP the order p in the signal amplitude calculation can be an arbitrary positive number (Neukom & Schacher, 2008), 4.64 is

the order which was considered in the practical experiments (see sections 3.3 and 3.6).

In order to establish communication between Processing and MAX MSP Jitter, the MaxLink 0.36 library[29], developed by Jesse Kriss, is used. The parameters which are sent from Processing to MAX MSP Jitter are essentially the horizontal angle and the x and z coordinates of the left or right hand, and the '1' or '0' commands to turn audio and loop on or off, according to the respective experiments described in section 3.6.

Finally, Promethean ActiveInspire Studio Professional Edition 1.4.20411 was used to record all the on-screen information during the practical experiments at a rate of 5 frames per second.

All algorithms and patches used in the experiments are presented and explained in detail below, in order to support and justify what is intended with each of the experiments described in section 3.6.

3.7.1.1 Processing Algorithms of Experiment 1

In our Processing program concerning experiment 1 (see appendix E.3), the code is run in this order:

1. Import MaxLink, OpenGL, and SimpleOpenNI libraries (see pseudo-code in algorithm 3.1)

Algorithm 3.1 Experiment 1 - Import MaxLink, OpenGL, and SimpleOpenNI libraries

```
1: import maxlink.*
2: import processing.opengl.*
3: import SimpleOpenNI.*
```

As already mentioned in section 3.7.1, the MaxLink library is used to make it possible to send data from Processing to MAX MSP Jitter. In turn, OpenGL (Open Graphics Library) is used to create a window on the computer screen, where the depth image of the participant, the green arrow corresponding to the direction of the deictic (pointing) gesture, and other information are shown. SimpleOpenNI is used for left or right hand motion capture.

2. Declare global variables (Reas & Fry, 2010, p. 52), that is, symbolic names or identifiers of a computer storage location associated with values that can be changed and used anywhere in the program (Wikipedia, 2016d), including class instance or object variables used in MaxLink, SimpleOpenNI, and in item 6 of the code running order (cf. item 6) (see algorithm 3.2). "A class defines a group of methods (functions) and fields (variables)" (Reas & Fry, 2007, p. 395), that is, it is the specification, blueprint, or template for an object or instance (Reas & Fry, 2010, p. 130). Thus, an object is a single instance of a class (Reas & Fry, 2007, p. 395), that is, a collection of related variables and functions, the former also known as instance variables or fields and the latter called methods in the context of objects (Reas & Fry, 2010, p. 129). These global variables are declared outside of functions `setup` and `draw` (see items 3 and 4 of the code running order), in order to avoid repetitions in the code (Reas & Fry, 2010, p. 37). By declaring variables, space in the computer's memory is set aside in order to store information (Reas & Fry, 2010, p. 39).

   On line 1 of algorithm 3.2, a new connection object between Processing and MAX MSP Jitter is created or initialized (cf. item 6 of the code running order), so that the instance variable, or field, called `parameters`

[29] Retrieved 10/06/2012, from `http://jklabs.net/maxlink/distro/maxlink-0.36.zip`

Algorithm 3.2 Experiment 1 - Declare global and class variables

```
 1: MaxLink link = new MaxLink(this, "parameters")
 2: SimpleOpenNI kinect
 3: boolean autoCalib = true
 4: boolean maos = true
 5: boolean recording = false
 6: Data data
 7: int count = 1
 8: float angle
 9: float m0
10: float m1
11: float m2
12: float m3
13: float m4
14: float m5
15: float m6
16: float m7
17: float m8
18: float m9
19: float m10
20: float m11
21: float m12
22: float m13
23: float m14
24: float m15
25: String title = "UTILIZADOR DEFINE ORIGEM"
26: PFont font
```

can receive data in MAX MSP Jitter. On line 2, the instance variable `kinect` is declared, so that all data from the Kinect camera can be accessed later. In order to always perform the self-calibration of skeleton joints when a new user is detected, the boolean variable `autoCalib`, which can have two values, usually *true* or *false*, is defined on line 3 as `true` (cf. algorithm 3.14). The option for the detection of the left (*false*) or right (*true*) hand is achieved by the boolean variable `maos`, which is initialized at first with the value `true` (right hand) on line 4. On line 5, the boolean variable `recording` is initialized with the value `false`, making it possible to continuously store or reproduce the user's motion data respectively in or from memory, when it is altered to `true` by pressing the *g* key (cf. algorithm 3.13). Thus, data can be stored on the instance variable `data` through the class `Data`, declared on line 6. The amount of the user's motion data sets stored in the `data` instance variable is counted by the integer `count` variable, which is initialized with value 1 on line 7 (cf. algorithm 3.12).

On line 8, the float variable `angle` is declared, in order to contain real numbers representing the angle of the arrow related to the direction of the deictic (pointing) gesture of the right or left hand. Variables `m0` to `m15` are declared from line 9 to line 24, so that it is possible to continuously store the orientation data of the left or right arm on memory from a 4x4 `PMatrix3D` transformation matrix called `orientation`, if the *g* key is pressed. According to Borenstein (2012, p. 233), a matrix is a mathematical structure which stores "... multiple vectors in a single grid of numbers." In turn, recalling what has already been mentioned in section 2.3.4.8, a vector is a way of describing the position of a point in space as the combination of a distance and a direction, having therefore a certain magnitude, length, or size, and a

unit vector representing its direction, with a length of one unit (Borenstein, 2012, p. 221). Since the default or current coordinate system is usually changed by a set of transformations, such as rotations, translations, and scales, in order to match the orientation of the skeleton joint that is being tracked, the 4x4 matrix called `orientation` is a transformation matrix that stores the changes from one vector space into another. Thus, this matrix stores four vectors, three of which represent the x-, y-, and z-axes, and the fourth defines the position of the new vector space relatively to the previous space, that is, it holds the results of the performed translation (Borenstein, 2012, p. 233). Variables `m0`, `m1`, and `m2` represent therefore the x-axis orientation vector in the new vector space; `m4`, `m5`, and `m6` correspond to the y-axis orientation vector in the new vector space; `m8`, `m9`, and `m10` describe the z-axis orientation vector in the new vector space; `m12`, `m13`, and `m14` represent the translation vector, but in our case all these three components are automatically defined as equal to 0.0 by calling `pushMatrix` and `popMatrix` in algorithms 3.11 and 3.12, because we want the origin of our coordinate system to always be maintained at the same position, so that the green arrow related to the direction of the deictic gesture of the right or left hand also has its origin at that same position on the OpenGL window; `m15` is the homogeneous coordinate[30] used for projective transformations in OpenGL, which in our case will be always equal to 1.0; and `m3`, `m7`, and `m11` are all equal to 0.0 (see figure 3.32).

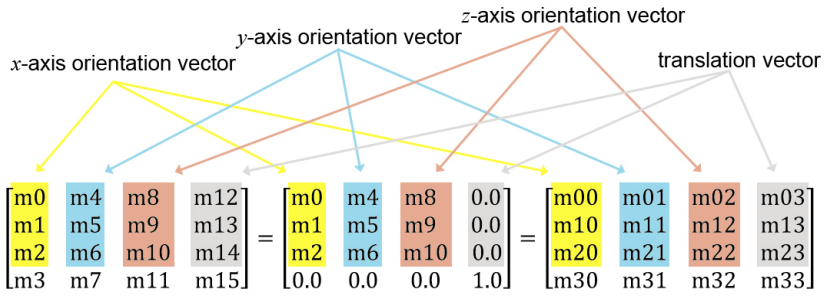


Figure 3.32: The 4x4 transformation matrix: our declared variables are on the left and centre matrices; the equivalent variables defined by `PMatrix3D` are on the right matrix.

On line 25, the string variable `title` is initialized with a three words long text indicative of the experiment being performed, in this case of experiment 1: "User Defines Origin", which appears on the OpenGL window. Finally, on line 26, the variable `font` is declared, so that the type of text font, which will be used in the OpenGL window, can be defined later in the program.

3. Call function `setup` (see algorithm 3.3), a basic building block of a Processing program (Reas & Fry, 2010, p. 15) containing code that runs just once when the program starts, in order to define starting values (Reas & Fry, 2010, p. 52). This function does not return any value, reason why the keyword `void` is used before the name of the function `setup` (Reas & Fry, 2010, p. 131).

On line 1 of algorithm 3.3, the previously mentioned OpenGL window, where the depth image, the green arrow, and other information are displayed, is created on the computer screen with the dimensions of

[30] For further explanations on this subject, which is beyond the scope of our thesis, please visit `https://en.wikipedia.org/wiki/Homogeneous_coordinates`

Algorithm 3.3 Experiment 1 - Function `setup`

```
1: size(1024, 768, OPENGL)
2: kinect = new SimpleOpenNI(this)
3: kinect.enableDepth()
4: kinect.enableUser(SimpleOpenNI.SKEL_PROFILE_ALL)
5: kinect.setMirror(true)
6: data = new Data()
7: data.beginSave()
8: link.output(4, 0)
9: link.output(5, 0)
```

1024x768 pixels. On line 2, the SimpleOpenNI instance `kinect`, declared on line 2 in algorithm 3.2, is created in order to access all of the depth camera's information. Thus, the `enableDepth` function, rather called method in this context because it is a function within an object (Reas & Fry, 2007, p. 107), is called on line 3 to access the depth image, the `enableUser` method is called on line 4 to be able to track all of the joints of a user's skeleton, and the `setMirror` method is called on line 5 to activate the mirror effect, so that the depth camera's data match those of the actual motions of the user.

Whereas on line 6 the instance `data` is created for the system to be able to save information related to the user's motions, on line 7 the `beginSave` method of class `Data` is called to create the instance variable `datalist` as a data array without any limited number of these data elements.

The last two lines of function `setup()` call the `output` method of the instance variable `parameters`, defined on line 1 in algorithm 3.2, in order to send the command `0` through outlets 4 and 5 of `parameters` to MAX MSP Jitter, so that a predetermined sound bank for this experiment can be selected (line 8) and no looping sounds are generated (line 9).

4. Call function `draw` (algorithms 3.4 to 3.7 and 3.9 to 3.12 are within it), another basic building block of a Processing program, but that updates the code within it continuously in a loop, since it repeats itself until the program is quit (Reas & Fry, 2010, p. 51). By default, Processing tries to run this function 60 times per second, but this depends on its contents and on the computer resources (Borenstein, 2012, p. 60). This function does not return any value as well, reason why the keyword `void` is also used before the name of the function `draw`.

Algorithm 3.4 Experiment 1 - Function `draw` - Part 1

```
1: kinect.update()
2: background(0)
3: image(kinect.depthImage(), 0, 0, 320, 240)
4: translate(width/2, height/2, 0)
5: rotateX(radians(180))
6: carry out algorithm 3.5
```

On line 1 of algorithm 3.4, the `update` method is called for the depth camera to permanently collect new data at a rate of 30 frames per second or every 1/30 of a second (Borenstein, 2012, p. 60). In order to get a black OpenGL window on the computer screen, the `background()` function is called on line 2, where `0` stands for black. The depth image is drawn on the OpenGL window from the top left to the bottom with the dimensions of 320x240 pixels by using the `image` function on line 3. On lines 4 and 5, the coordinate

system, used as a reference for the green arrow corresponding to the direction of the deictic gesture, is shifted to the middle of the OpenGL window and rotated by 180º around the x-axis, so that the positive side of the y-axis, usually pointing downwards in Processing, now points upwards as in SimpleOpenNI (Borenstein, 2012, p. 120), and the positive side of the z-axis, usually pointing forwards, now points backwards. On line 6, algorithm 3.5 is executed.

Algorithm 3.5 Experiment 1 - Function `draw` - Part 2

```
1: IntVector userList = new IntVector()
2: kinect.getUsers(userList)
3: if userList.size() > 0 then
4:    carry out algorithm 3.6
5: end if
```

On line 1 of algorithm 3.5, a vector of integers called `userList` is created to store the list of all detected users to whom a unique integer is associated (user identifiers or IDs). On line 2, all detected user IDs are put into this vector. If any user is detected (line 3), then algorithm 3.6 is carried out (line 4).

Algorithm 3.6 Experiment 1 - Function `draw` - Part 3

```
1: int userId = userList.get(0)
2: if kinect.isTrackingSkeleton(userId) then
3:    carry out algorithm 3.7
4: end if
```

On line 1 of algorithm 3.6, the first user from the list of all detected users is selected. If the user has been successfully calibrated by performing the ”Psi” detection posture (line 2), then algorithm 3.7 is executed (line 3).

On line 1 of algorithm 3.7, a vector named `position` is created to store the three-dimensional position of the right or left hand in the new coordinate system defined in algorithm 3.4. The float variable `confidencePos` is declared on line 2, in order to store the confidence with which the right or left hand is detected (lines 4 and 6, respectively). In turn, on line 8, the three-dimensional orientation of the respective elbow is stored into a matrix called `orientation`. The float variable `confidenceOrient` is declared on line 9, so that the confidence with which the right or left elbow is detected can be stored as well (lines 11 and 13, respectively). From line 15 to line 17, the float variables `coordX0`, `coordY0`, and `coordZ0` are declared to store the direction of the deictic gesture. Its vertical angle can be defined based on `coordX0` and `coordY0` and its horizontal angle can be determined by `coordX0` and `coordZ0`. Depending on which hand is being tracked, the coordinates `coordZ0` or `coordX0` are multiplied by -1, so that the respective forward/backward or left/right relationship remains correct. Finally, on line 27, algorithm 3.9 is called. Algorithm 3.7 could have been written in a shorter form, such as in the alternative algorithm 3.8, but we let it stay as elementary as possible, so that it could be more easily understood by us at the time when we were carrying out the practical experiments.

From line 1 to line 16 of algorithm 3.9, the matrix elements corresponding to the direction of the deictic gesture are assigned to the float variables `m0` to `m15`, declared in algorithm 3.2. On line 17, algorithm 3.10 is called.

From line 1 to line 7 of algorithm 3.10, the horizontal angle of the right or left hand is determined according to its `coordX0` and `coordZ0` coordinates in the horizontal plane, so that the angle is defined as being: 1) on the first quadrant[31] if `coordX0` and `coordZ0` are both greater than zero; 2) on the second quadrant if `coordX0` is less than zero and `coordZ0` is greater than zero; 3) on the third quadrant if `coordX0` and `coordZ0` are both less than zero; and 4) on the fourth quadrant if `coordX0` is greater than zero and `coordZ0` is less than zero. On line 8, algorithm 3.11 is called.

Algorithm 3.7 Experiment 1 - Function `draw` - Part 4

```
1: PVector position = new PVector()
2: float confidencePos
3: if maos is true then
4:   confidencePos = kinect.getJointPositionSkeleton(userId, SimpleOpenNI.SKEL_RIGHT_HAND,
     position)
5: else if maos is false then
6:   confidencePos = kinect.getJointPositionSkeleton(userId, SimpleOpenNI.SKEL_LEFT_HAND,
     position)
7: end if
8: PMatrix3D orientation = new PMatrix3D()
9: float confidenceOrient
10: if maos is true then
11:   confidenceOrient = kinect.getJointOrientationSkeleton(userId, SimpleOpenNI.SKEL_RIGHT
     _ELBOW, orientation)
12: else if maos is false then
13:   confidenceOrient = kinect.getJointOrientationSkeleton(userId, SimpleOpenNI.SKEL_LEFT
     _ELBOW, orientation)
14: end if
15: float coordX0
16: float coordY0
17: float coordZ0
18: if maos is true then
19:   coordX0 = orientation.m00
20:   coordY0 = orientation.m10
21:   coordZ0 = (orientation.m20)*(-1)
22: else if maos is false then
23:   coordX0 = (orientation.m00)*(-1)
24:   coordY0 = orientation.m10
25:   coordZ0 = orientation.m20
26: end if
27: carry out algorithm 3.9
```

Algorithm 3.11 has essentially to do with written information on the OpenGL window for the researcher. Thus, on line 1 and line 2, the text font `ArialMT-16` is loaded. In order not to interfere with the coordinate system of the green arrow corresponding to the direction of the deictic gesture, so that information can be displayed on the OpenGL window using its own coordinate system, the pair of functions `pushMatrix()` (line 3) and `popMatrix` (line 31) is used (Borenstein, 2012, p. 120). The coordinate system associated with the text is rotated by 180° around the x-axis on line 4, so that the positive side of the y-axis is pointing downwards again, as usually in Processing. On line 5, the white colour is selected for the text which comprises the title of the experiment, the angle and the coordinates `coordX0`, `coordY0`, and

[31] A quadrant is any of four infinite regions that are obtained when a Cartesian plane is divided by two infinite axes that are perpendicular to each other, where each of these regions is bounded by two half-axes. "When the axes are drawn according to the mathematical custom, the numbering [of the quadrants] goes counter-clockwise starting from the upper right ("northeast") quadrant" (Wikipedia, 2016c).

Algorithm 3.8 Experiment 1 - Function `draw` - Part 4 (alternative)

```
1: PVector position = new PVector()
2: float confidencePos
3: PMatrix3D orientation = new PMatrix3D()
4: float confidenceOrient
5: float coordX0
6: float coordY0
7: float coordZ0
8: if maos is true then
9:   confidencePos = kinect.getJointPositionSkeleton(userId, SimpleOpenNI.SKEL_RIGHT_HAND,
     position)
10:  confidenceOrient = kinect.getJointOrientationSkeleton(userId, SimpleOpenNI.SKEL_RIGHT
     _ELBOW, orientation)
11:  coordX0 = orientation.m00
12:  coordY0 = orientation.m10
13:  coordZ0 = (orientation.m20)*(-1)
14: else if maos is false then
15:  confidencePos = kinect.getJointPositionSkeleton(userId, SimpleOpenNI.SKEL_LEFT_HAND,
     position)
16:  confidenceOrient = kinect.getJointOrientationSkeleton(userId, SimpleOpenNI.SKEL_LEFT
     _ELBOW, orientation)
17:  coordX0 = (orientation.m00)*(-1)
18:  coordY0 = orientation.m10
19:  coordZ0 = orientation.m20
20: end if
21: carry out algorithm 3.9
```

Algorithm 3.9 Experiment 1 - Function `draw` - Part 5

```
1: m0 = orientation.m00
2: m1 = orientation.m10
3: m2 = orientation.m20
4: m3 = orientation.m30
5: m4 = orientation.m01
6: m5 = orientation.m11
7: m6 = orientation.m21
8: m7 = orientation.m31
9: m8 = orientation.m02
10: m9 = orientation.m12
11: m10 = orientation.m22
12: m11 = orientation.m32
13: m12 = orientation.m03
14: m13 = orientation.m13
15: m14 = orientation.m23
16: m15 = orientation.m33
17: carry out algorithm 3.10
```

Algorithm 3.10 Experiment 1 - Function `draw` - Part 6

```
1: if coordX0 > 0 and coordZ0 > 0 then
2:   angle = degrees(atan(coordZ0/coordX0))
3: else if (coordX0 < 0 and coordZ0 > 0) or (coordX0 < 0 and coordZ0 < 0) then
4:   angle = 180 + (degrees(atan(coordZ0/coordX0)))
5: else if coordX0 > 0 and coordZ0 < 0 then
6:   angle = 360 + (degrees(atan(coordZ0/coordX0)))
7: end if
8: carry out algorithm 3.11
```

Algorithm 3.11 Experiment 1 - Function `draw` - Part 7

```
font = loadFont("ArialMT-16.vlw")
textFont(font)
pushMatrix()
rotateX(radians(180))
fill(255, 255, 255)
text(title, -(width/2)+30, 130)
text("Ângulo:  " + angle, -(width/2)+30, 160)
text("coordX0:  " + coordX0 + " coordY0:  " + coordY0 + " coordZ0:  " + coordZ0,
-(width/2)+30, 180)
text("Gravar ou continuar a gravar movimentos na memória - tecla 'g'", -(width/2)+30,
260)
text("Parar de gravar movimentos na memória - tecla 'p'", -(width/2)+30, 280)
text("Apagar movimentos da memória - tecla 'a'", -(width/2)+30, 300)
text("Gravar movimentos num ficheiro - tecla 's'", -(width/2)+30, 320)
if recording is true then
  fill(0, 255, 0)
else if recording is false then
  fill(255, 255, 255)
end if
text("Total recorded data:  " + count, -(width/2)+30, 350)
text("Recording state:  " + recording, -(width/2)+30, 370)
if maos is true then
  fill(0, 255, 0)
  text("Para esquerdinos - tecla 'e'", -(width/2)+30, 210)
  fill(255, 255, 255)
  text("Para destros - tecla 'd'", -(width/2)+30, 230)
else if maos is false then
  fill(0, 255, 0)
  text("Para destros - tecla 'd'", -(width/2)+30, 230)
  fill(255, 255, 255)
  text("Para esquerdinos - tecla 'e'", -(width/2)+30, 210)
end if
popMatrix()
carry out algorithm 3.12
```

Algorithm 3.12 Experiment 1 - Function `draw` - Part 8

```
link.output(0, angle)
link.output(1, coordX0)
link.output(2, coordZ0)
position.normalize()
if confidencePos > 0.5 then
  println(confidencePos)
  pushMatrix()
  translate(position.x, position.y, position.z)
  applyMatrix(orientation)
  stroke(0, 255, 0)
  strokeWeight(6)
  if maos is true then
    line(0, 0, 0, 200, 0, 0)
    line(200, 0, 0, 150, 50, 0)
    line(200, 0, 0, 150, -50, 0)
  else if maos is false then
    line(0, 0, 0, -200, 0, 0)
    line(-200, 0, 0, -150, 50, 0)
    line(-200, 0, 0, -150, -50, 0)
  end if
  popMatrix()
  if recording is true then
    data.add("Title:  "+title)
    data.add("Count:  "+count)
    data.add("userId:  "+userId)
    data.add("confidencePos:  "+confidencePos)
    data.add("confidenceOrient:  "+confidenceOrient)
    data.add("maos:  "+maos)
    data.add("coordX0:  "+coordX0)
    data.add("coordZ0:  "+coordZ0)
    data.add("angle:  "+angle)
    data.add("position.x:  "+position.x)
    data.add("position.y:  "+position.y)
    data.add("position.z:  "+position.z)
    data.add("m0:  "+m0)
    data.add("m1:  "+m1)
    data.add("m2:  "+m2)
    data.add("m3:  "+m3)
    data.add("m4:  "+m4)
    data.add("m5:  "+m5)
    data.add("m6:  "+m6)
    data.add("m7:  "+m7)
    data.add("m8:  "+m8)
    data.add("m9:  "+m9)
    data.add("m10:  "+m10)
    data.add("m11:  "+m11)
    data.add("m12:  "+m12)
    data.add("m13:  "+m13)
    data.add("m14:  "+m14)
    data.add("m15:  "+m15)
    data.add("-------")
    count +=1
  end if
  if coordY0 >= -0.5 and coordY0 <= 0.6 then
    link.output(3, 1)
  else
    link.output(3, 0)
  end if
end if
```

coordZ0 of the direction of the deictic gesture, and the instructions to store, stop storing, or delete motion data in or from memory, as well as to store motion data in a file (line 6 to line 12).

If the researcher is recording data in memory (line 13), then the amount of total recorded data and the recording state are displayed in green on the OpenGL window (line 14). Otherwise, they are shown in white (line 16). Whereas the instructions on the selection of the left or right hand are respectively written in green and white if the right hand is being used (line 20), the same instructions are displayed respectively in white and green if the left hand is instead chosen (line 25). Finally, algorithm 3.12 is called on line 32.

From line 1 to line 3 of algorithm 3.12, the angle and the coordX0 and coordZ0 coordinates are sent by the link object, initialized in algorithm 3.2, to the respective outlets 0, 1, and 2 of instance variable parameters in MAX MSP Jitter. On line 4, the position vector is normalized, so that it is transformed into a unit vector with a magnitude of 1 unit. If the confidence with which the right or left hand position is detected is greater than 0.5 (line 5), that is, if it is greater than 50%, then the value of the confidencePos variable is printed on the Processing's console or bottom area of the Processing Environment (line 6) (Reas & Fry, 2010, p. 181) as mere information for us during the course of the experiments that the participant's deictic gesture is being tracked. In addition, a green (line 10) six-pixels-thick (line 11) arrow (line 13 to line 15 if the right hand is selected or line 17 to line 19 if the left hand is used), representing the direction of the right or left hand deictic gesture (line 9), is drawn in the centre of the OpenGL window with its origin matching the hand's joint (line 8), isolated from any other coordinate systems used in the program by applying the pushMatrix() (line 7) and popMatrix() (line 21) functions again.

From line 23 to line 51, a set of values of the variables title, count, userId, confidencePos, confidence Orient, maos, coordX0, coordZ0, angle, position.x, position.y, position.z, m0 to m15, and a separator formed by seven hyphens is added to the instance variable data if recording of these data is desired. On line 52, the count variable, which counts the amount of added data sets to data, is incremented by one.

Finally, if the deictic gesture of the participant is used in the horizontal plane, a calibration task which has been defined experimentally by us in line 54, that is, if the value of the vertical coordinate of the deictic gesture coordY0 is between -0.5 and 0.6, then a 'turn on audio' command is sent to outlet 3 of instance variable parameters in MAX MSP Jitter (line 55), so that a random audio file is triggered. Otherwise, the audio is turned off (line 57).

5. Define other functions outside setup and draw, without returning a value to the main program (reason why the keyword void is used before the names of the respective functions), such as keyPressed, onNewUser, onLostUser, onExitUser, onReEnterUser, onStartCalibration, onEndCalibration, onStartPose, and onEndPose.

   On line 1 of algorithm 3.13, if the "e" key is pressed, then the left hand is tracked by the system (variable maos equal to false means that the left hand is taken into account). On line 4, if the "d" key is pressed, then the right hand is tracked by the system (variable maos equal to true means that the right hand is considered). On line 7, if the "g" key is pressed and motion data are still not being recorded, then the

program starts to write them to memory. On line 10, if the ”p” key is pressed and motion data are being recorded, then the program stops to store them in the memory. On line 13, if the ”a” key is pressed, motion data are not being recorded, and the amount of data sets already written to memory is at least equal to one, then memory is erased, that is, all elements of the instance variable `data` are deleted by calling the `restart` method in the `Data` class (see item 6 of the code running order). Finally, if the ”s” key is pressed, motion data are not being recorded, and the amount of data sets already stored in memory is at least equal to one, then data are written to a text file with a name consisting of five digits starting at zero, which is incremented by one whenever a new file is recorded.

Algorithm 3.13 Experiment 1 - Function `keyPressed`

```
1: if key = 'e' then
2:   maos is false
3: end if
4: if key = 'd' then
5:   maos is true
6: end if
7: if key = 'g' and recording is false then
8:   recording is true
9: end if
10: if key = 'p' and recording is true then
11:   recording is false
12: end if
13: if key = 'a' and recording is false and count > 1 then
14:   data.restart()
15: end if
16: if key = 's' and recording is false and count > 1 then
17:   data.endSave(data.getIncrementalFilename(sketchPath("save" + java.io.File.separator + "data #####.txt")))
18: end if
```

Algorithm 3.14 Experiment 1 - Function `onNewUser`

```
1: if autoCalib is true then
2:   println("onNewUser - requestCalibrationSkeleton - userId:  " + userId)
3:   kinect.requestCalibrationSkeleton(userId, true)
4: else
5:   println("onNewUser - startPoseDetection 'Psi' - userId:  " + userId)
6:   kinect.startPoseDetection("Psi", userId)
7: end if
```

On line 1 of algorithm 3.14, when a new user is detected by the system, then the skeleton calibration routine is called if the variable `autoCalib` is `true` and that fact is made known to us during the course of the experiments on the Processing’s console. Otherwise, the detection of the ”Psi” posture is started and that information is displayed to us in the Processing’s console.

If the system no longer detects the user (`onLostUser` function), if the user leaves the scene (`onExitUser` function), if the user re-enters the scene (`onReEnterUser` function), or if the calibration is started (`onStart Calibration` function), then this information is given to us on the Processing’s console.

On line 1 of algorithm 3.15, information is given to us during the course of the experiments on the Processing’s console that the `onEndCalibration` function is being performed. If the calibration is carried

Algorithm 3.15 Experiment 1 - Function `onEndCalibration`

```
1: println("onEndCalibration - userId:  " + userId + ", successful:  " + successful)
2: if successful is true then
3:   println("onEndCalibration - startTrackingSkeleton - userId:  " + userId)
4:   kinect.startTrackingSkeleton(userId)
5: else
6:   println("onEndCalibration - startPoseDetection 'Psi' - userId:  " + userId)
7:   kinect.startPoseDetection("Psi", userId)
8: end if
```

out successfully on line 2, then the system starts tracking the user's skeleton and that information is given to us on the Processing's console. Otherwise, the detection of the "Psi" posture is started and that situation is reported to us on the Processing's console.

Algorithm 3.16 Experiment 1 - Function `onStartPose`

```
1: println("onStartPose - stopPoseDetection - userId:  " + userId)
2: kinect.stopPoseDetection(userId)
3: kinect.requestCalibrationSkeleton(userId, true)
```

As soon as the user's "Psi" posture detection starts (algorithm 3.16), this information is given to us on the Processing's console and the system stops the detection of this posture, calling the skeleton calibration function afterwards.

If the detection of the "Psi" posture is finished (`onEndPose` function), then this fact is made known to us on the Processing's console.

6. Define the `Data` class, which is used to store motion data in memory and to files, based on the Norwegian artist Marius Watz's example[32] showing how to use the Data utility class to save and load data from text files.

   On line 1 of algorithm 3.17, the array `datalist`, which will contain a list of motion data, is declared. The number of elements it will hold is unlimited, depending only on the memory and the storage capacity of the computer. On lines 2 and 3, the string variables `filename` and `data`, and the integer variable `datalineId` are respectively declared. The `beginSave` method on line 4 creates a new `datalist` array. The four next methods, from lines 6 to 13, allow the addition of integers (line 6), real numbers (line 8), boolean values (line 10) and characters (line 12) to the new `datalist` array. From line 14 to 18, the `endSave` method is defined, so that in algorithm 3.18 a file containing all the motion data written here to memory can be saved in the computer when the "s" key is pressed, as described in algorithm 3.13. The `restart` method is presented on line 19, in order to clear the memory, that is, to delete all elements of the instance variable `data` when the "a" key is pressed, as described in algorithm 3.13 as well. Lastly, algorithm 3.18 is called on line 22.

   Algorithm 3.18 is a utility function which saves files, containing the motion data kept in memory, into a directory, whose name "save" is defined on line 17 of algorithm 3.13. This directory will in turn appear inside the directory where the Processing experiment program is stored (line 12 of algorithm 3.18). In addition, algorithm 3.18 auto-increments file names based on the template "data #####.txt", defined in

[32] Retrieved 28/07/2012, from `http://wiki.processing.org/w/Saving_data_to_text_files`

Algorithm 3.17 Experiment 1 - Class `Data` - Part 1

```
 1: ArrayList datalist
 2: String filename, data[]
 3: int datalineId
 4: void beginSave()
 5:   datalist = new ArrayList()
 6: void add(int val)
 7:   datalist.add("" + val)
 8: void add(float val)
 9:   datalist.add("" + val)
10: void add(boolean val)
11:   datalist.add("" + val)
12: void add(String s)
13:   datalist.add(s)
14: void endSave(String _filename)
15:   filename = _filename
16:   data = new String[datalist.size()]
17:   data = (String [])datalist.toArray(data)
18:   saveStrings(filename, data)
19: void restart()
20:   count = 1
21:   datalist.clear()
22: carry out algorithm 3.18
```

algorithm 3.13 on line 17, where each # stands for a number. The code contained in the `while` loop, starting on line 16 of algorithm 3.18, is responsible for that action.

Thus, on line 1 of algorithm 3.18, the variable `s`, which will hold the name of the new file to record (line 21), is initialized with the empty string default value and the string variables `prefix`, `suffix`, `padstr`, and `numstr` are declared. Whereas the variable `prefix` saves the first word from the template name (line 9), that is, ”data ”, `suffix` stores the text ”.txt” (line 10). In turn, since in our case it is possible to record up to 100000 files in the computer (from 00000 to 99999), `numstr` stores the integer number of the file to be written (line 18) and `padstr`, declared with the empty string default value on line 17, stores as many zeros as necessary before `numstr`, until both have together the length equal to the number of # that appear in the template name, that is, a length of five digits (line 20). On line 2, the integer variable `index`, which will hold the number of files already recorded, is initialized with zero and the integer variables `first`, `last`, and `count` are declared, so that the number of # occurring in the template file name can be counted. In our case, `first` is initially equal to zero, because the first # which occurs in the group of # is in position zero (line 5); `last` is equal to four, because the last # that occurs in the group of # is in position four (line 6). Thus, # occurs five times and `count` is therefore equal to five (line 7).

On line 22, the file instance `f`, declared on line 3, is created with the name of variable `s` if it does not exist yet. This condition is controlled by the boolean `ok` variable on line 23. If the number of recorded files is greater than 100000, then an error is presented to us on the Processing’s console and returned to the main program (line 30 to line 34). Finally, if there are no errors, on line 36 the `s` value is returned to the main program by the `getIncrementalFilename` utility function.

Algorithm 3.18 Experiment 1 - Class `Data` - Part 2 - Function `public String getIncrementalFilename(String templ)`

```
 1: String s = "", prefix, suffix, padstr, numstr
 2: int index = 0, first, last, count
 3: File f
 4: boolean ok
 5: first = templ.indexOf('#')
 6: last = templ.lastIndexOf('#')
 7: count = last - first + 1
 8: if first != -1 and last - first > 0 then
 9:   prefix = templ.substring(0, first)
10:   suffix = templ.substring(last+1)
11:   if sketchPath != null then
12:     prefix = savePath(prefix)
13:   end if
14:   index = 0
15:   ok = false
16:   while !ok do
17:     padstr = ""
18:     numstr = "" + index
19:     for (int i = 0; i < count - numstr.length(); i++) do
20:       padstr += "0"
21:       s = prefix + padstr + numstr + suffix
22:       f = new File(s)
23:       ok =!  f.exists()
24:       index++
25:     end for
26:     if index > 100000 then
27:       ok = true
28:     end if
29:   end while
30:   if index > 100000 then
31:     println("getIncrementalFilename thinks there is a problem - " + "Are there more than
        100000 files already in the sequence " + " or is the filename invalid?")
32:     println("Returning " + prefix + "ERR" + suffix)
33:     return prefix + "ERR" + suffix
34:   end if
35: end if
36: return s
```

3.7.1.2 Processing Algorithms of Experiment 2

In experiment 2, all algorithms are equal to those of experiment 1, except algorithm 3.7, which presents a slight difference from line 19 to line 25 (see algorithm 3.19 as a replacement for algorithm 3.7). Depending on which hand is being tracked, the coordinates `coordX0` or `coordZ0` are multiplied by -1, so that the respective left/right or the forward/backward relationship is reversed relatively to the direction of the deictic gesture, as described in section 3.6. Algorithm 3.19 could have also been written in a shorter form, such as in the alternative algorithm 3.20, but we let it stay as elementary as possible for it to be more easily understood by us at the time when we were carrying out the practical experiments.

Algorithm 3.19 Experiment 2 - Function `draw` - Part 4

```
1: PVector position = new PVector()
2: float confidencePos
3: if maos is true then
4:   confidencePos = kinect.getJointPositionSkeleton(userId, SimpleOpenNI.SKEL_RIGHT_HAND,
     position)
5: else if maos is false then
6:   confidencePos = kinect.getJointPositionSkeleton(userId, SimpleOpenNI.SKEL_LEFT_HAND,
     position)
7: end if
8: PMatrix3D orientation = new PMatrix3D()
9: float confidenceOrient
10: if maos is true then
11:   confidenceOrient = kinect.getJointOrientationSkeleton(userId, SimpleOpenNI.SKEL_RIGHT_ELBOW,
     orientation)
12: else if maos is false then
13:   confidenceOrient = kinect.getJointOrientationSkeleton(userId, SimpleOpenNI.SKEL_LEFT_ELBOW,
     orientation)
14: end if
15: float coordX0
16: float coordY0
17: float coordZ0
18: if maos is true then
19:   coordX0 = (orientation.m00)*(-1)
20:   coordY0 = orientation.m10
21:   coordZ0 = orientation.m20
22: else if maos is false then
23:   coordX0 = orientation.m00
24:   coordY0 = orientation.m10
25:   coordZ0 = (orientation.m20)*(-1)
26: end if
27: carry out algorithm 3.9
```

3.7.1.3 Processing Algorithms of Experiment 3

In experiment 3, the code is run in an identical order as in experiments 1 and 2, but presents some differences. Thus, algorithms 3.1, 3.4 to 3.10, and 3.14 to 3.18, which were already introduced in section 3.7.1.1, are used as in experiment 1. Algorithm 3.2 in experiment 1 is used in experiment 3 as well, but some changes are introduced: the global float variables `coordX0Soft`, `coordZ0Soft`, and `angleSoft` are also declared, in order to hold the coordinates and the angle which Processing determines randomly for the production of a sound.

Algorithm 3.20 Experiment 2 - Function `draw` - Part 4 (alternative)

```
PVector position = new PVector()
float confidencePos
PMatrix3D orientation = new PMatrix3D()
float confidenceOrient
float coordX0
float coordY0
float coordZ0
if maos is true then
  confidencePos = kinect.getJointPositionSkeleton(userId, SimpleOpenNI.SKEL_RIGHT_HAND,
  position)
  confidenceOrient = kinect.getJointOrientationSkeleton(userId, SimpleOpenNI.SKEL_RIGHT_ELBOW,
  orientation)
  coordX0 = (orientation.m00)*(-1)
  coordY0 = orientation.m10
  coordZ0 = orientation.m20
else if maos is false then
  confidencePos = kinect.getJointPositionSkeleton(userId, SimpleOpenNI.SKEL_LEFT_HAND,
  position)
  confidenceOrient = kinect.getJointOrientationSkeleton(userId, SimpleOpenNI.SKEL_LEFT_ELBOW,
  orientation)
  coordX0 = orientation.m00
  coordY0 = orientation.m10
  coordZ0 = (orientation.m20)*(-1)
end if
carry out algorithm 3.9
```

Algorithm 3.21 Experiment 3 - Function `setup`

```
size(1024, 768, OPENGL)
kinect = new SimpleOpenNI(this)
kinect.enableDepth()
kinect.enableUser(SimpleOpenNI.SKEL_PROFILE_ALL)
kinect.setMirror(true)
angleSoft = random(-40, 221)
if angleSoft < 0 then
  angleSoft = 360 + angleSoft
end if
if (angleSoft >= 0 and angleSoft < 90) or (angleSoft >= 270 and angleSoft < 360) then
  coordX0Soft = random(0.1, 1)
else if (angleSoft >= 90 and angleSoft < 180) or (angleSoft >= 180 and angleSoft < 270)
then
  coordX0Soft = random(-0.1, -1)
end if
coordZ0Soft = coordX0Soft*tan(radians(angleSoft))
data = new Data()
data.beginSave()
link.output(4, 1)
link.output(5, 1)
link.output(3, 1)
```

In addition, the boolean variable `delayTime`, used to determine if a waiting time of two seconds between the end of a sound that has been correctly localized and the beginning of another sound is performed, is defined as equal to `false` (see appendix E.3).

Algorithm 3.3 in experiment 1 is replaced by algorithm 3.21 in experiment 3. Thus, on line 6 of algorithm 3.21, a random value determined by Processing between -40 and 220, as explained in section 3.6, is set in `angleSoft`, a variable that holds the angle of the direction of the sound that will be produced by MAX MSP Jitter. If `angleSoft` is negative, it becomes positive on line 8. If `angleSoft` is in the range of 0º to less than 90º or in the range of 270º to less than 360º, then `coordX0Soft` is set to a random positive value between 0.1 and 1 (line 11). Nevertheless, if `angleSoft` is in the range of 90º to less than 180º or in the range of 180º to less than 270º, then `coordX0Soft` is set to a random negative value between -0.1 and -1 (line 13). At the end of the algorithm, the `output` method of the instance variable `parameters` is called, so that the command 1 is sent through outlets 4, 5, and 3 of `parameters` to MAX MSP Jitter, in order to select a predetermined sound bank for this experiment (line 18), to generate looping sounds (line 19), and to turn audio on (line 20).

Whereas the `draw` function in experiment 1 starts with the `update` method of the depth camera (cf. algorithm 3.4), in experiment 3 it begins with a first part presented in algorithm 3.22, which takes into account the option of a two-second pause (line 2) between the end of a sound that has been correctly localized and the beginning of another reproduced sound if the variable `delayTime` is `true` (line 1). Furthermore, if the latter variable is `true`, then the command 1 is sent through outlets 5, 3, and 6 of `parameters` to MAX MSP Jitter, so that looping sounds are reproduced again (line 4), audio is turned on (line 5), and audio processing is once more started in our MAX MSP Jitter patch and sub-patches only (line 6). Afterwards, the `draw` function in experiment 3 is followed by algorithm 3.4 as a second part. Finally, at the end of the latter (line 6), algorithm 3.23 is called as the third part of the `draw` function, instead of algorithm 3.5 as in experiment 1.

Algorithm 3.22 Experiment 3 - Function `draw` - Part 0

```
1: if delayTime is true then
2:   delay(2000)
3:   delayTime = false
4:   link.output(5, 1)
5:   link.output(3, 1)
6:   link.output(6, 1)
7: end if
```

Algorithm 3.23 Experiment 3 - Function `draw` - Part 1A

```
1: link.output(0, angleSoft)
2: link.output(1, coordX0Soft)
3: link.output(2, coordZ0Soft)
4: carry out algorithm 3.5
```

Algorithm 3.23 sends the values of the angle `angleSoft` and of the respective coordinates `coordX0Soft` and `coordZ0Soft` to MAX MSP Jitter through outlets 0, 1, and 2 of the instance variable `parameters`, so that sound is reproduced in the direction determined by Processing. Algorithm 3.5 is performed next (line 4).

In experiment 3, algorithm 3.11 is used as in experiment 1, but the written information on the OpenGL window is somehow distinct, because of the nature of the experiments. The ”n” key instruction is now available in experiment 3, in order for the researcher to be able to select a new random reproduction angle and sound

if desired. Furthermore, the angle `angleSoft` and the respective coordinates `coordX0Soft`, `coordY0Soft`, and `coordZ0Soft` of the sound's origin, defined in Processing, are presented in the OpenGL window in a red colour.

The first three lines of code of algorithm 3.12 in experiment 1, which have to do with the transmission of the angle and coordinates of the direction of the deictic gesture of a participant to MAX MSP Jitter, are not available in experiment 3, because it is now intended that MAX MSP Jitter receives the angle and coordinates defined by Processing in order to produce a sound with a random direction, which is already performed by algorithm 3.23. By pressing the "g" key, motion data is written to memory as in algorithm 3.12 in experiment 1. However, the values of `angleSoft`, `coordX0Soft`, and `coordZ0Soft` are also written to memory.

Algorithm 3.24 Experiment 3 - Function `draw` - Part 9

```
1: if (angle >= angleSoft - 15 and angle <= angleSoft + 15) and (coordY0 >= -0.4 and coordY0
   <= 0.4) then
2:   pushMatrix()
3:   rotateX(radians(180))
4:   fill(255, 0, 0)
5:   text("SOM CORRECTAMENTE LOCALIZADO!!!", -(width/2)+330, 110)
6:   popMatrix()
7:   link.output(5, 0)
8:   link.output(3, 0)
9:   link.output(6, 0)
10:  delayTime = true
11:  angleSoft = random(-40, 221)
12:  if angleSoft < 0 then
13:    angleSoft = 360 + angleSoft
14:  end if
15:  if (angleSoft >= 0 and angleSoft < 90) or (angleSoft >= 270 and angleSoft < 360) then
16:    coordX0Soft = random(0.1, 1)
17:  else if (angleSoft >= 90 and angleSoft < 180) or (angleSoft >= 180 and angleSoft < 270)
     then
18:    coordX0Soft = random(-0.1, -1)
19:  end if
20:  coordZ0Soft = coordX0Soft*tan(radians(angleSoft))
21: end if
```

The final pseudo-code of algorithm 3.12 in experiment 1 (lines 54 to 58) is different from that of experiment 3, presented in algorithm 3.24. Thus, in experiment 3, if the participant uses his deictic gesture and manages to identify the sound's localization within a margin of $\pm$15 degrees (line 1), then the researcher is informed of that fact (line 5) with a red coloured text (line 4) in the OpenGL window. This margin allows the researcher to analyse the error committed by the participant when both the horizontal deictic gesture's `angle` and the Processing's determined sound angle `angleSoft` variables are compared (cf. section 3.6).

Furthermore, the looping sounds are stopped (line 7), audio is turned off (line 8), audio processing is suspended (line 9), and `delayTime` becomes equal to `true` (line 10), so that a pause of two seconds is performed when the function `draw` is read again from the beginning (cf. algorithm 3.22). From line 11 to line 20, the code is the same as in lines 6 to 15 in algorithm 3.21.

Finally, the function `keyPressed` in experiment 3 is globally identical to that of experiment 1 (cf. algorithm 3.13), but also allows the researcher to press the "n" key to choose another random reproduction angle and sound, as explained above. The pseudo-code which is run when that key is pressed is equal to that of lines 7 to 20 of algorithm 3.24.

3.7.1.4 MAX MSP Jitter Patches

The MAX MSP Jitter ”Gesture Detection Mono to Octophonic - 23-09-2012.maxpat” patch used in the three experiments (see appendix E.4) consists of several parts. The first one has to do with the number of loudspeakers which are actually used, ranging from one to eight. Since we use eight loudspeakers, the number 8 is immediately loaded into the `ncol` argument of the `send` or `s` object (see figure 3.33) and sent to a `receive` or `r` object named by the same argument (see figure 3.44) as soon as the patch is opened.

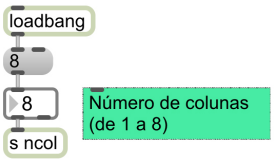


Figure 3.33: Number of loudspeakers used in MAX MSP Jitter.

The panning cross-fade time between loudspeakers can be changed, but in our case it is set to zero when the patch is loaded (see figure 3.34), in order to carry out the actual Ambisonics Equivalent Panning (AEP) function described in section 3.7.1. In figure 3.43, it can be observed that the cross-fade time is received through `r time`. This option was introduced for testing purposes only.

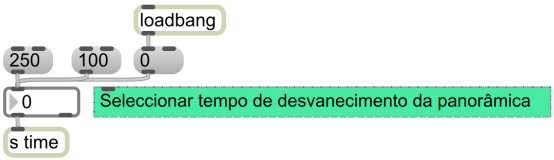


Figure 3.34: Selection of the panning cross-fade time between loudspeakers.

If we want to hear sound coming out of the loudspeakers, audio has to be turned on (see figure 3.35) after processing of audio is automatically activated in this patch and in its sub-patches only, when they are loaded for the first time (see figure 3.36).

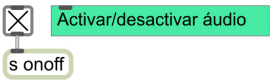


Figure 3.35: Turn audio on or off.

The Ambisonics’ order of 4.64, explained in sections 3.3, 3.6, and 3.7.1, is automatically loaded as well and sent to the ”level.maxpat” sub-patch, that is, to the sub-patch where the AEP function is defined (see figure 3.31 in section 3.7.1), although it can be modified if wanted, as can be seen in the part of the patch shown in figure 3.37.

Furthermore, the position of each loudspeaker is also defined in the same part of the patch presented in figure 3.37. Since we are using eight loudspeakers, the angles corresponding to their positions in the horizontal plane are automatically assigned, when the patch is loaded for the first time. Nevertheless, any integer value can be inserted for the angle of a loudspeaker in the horizontal plane if desired.

When we started to develop the experiments, we naturally considered that the angle of 0^{o} would be as-

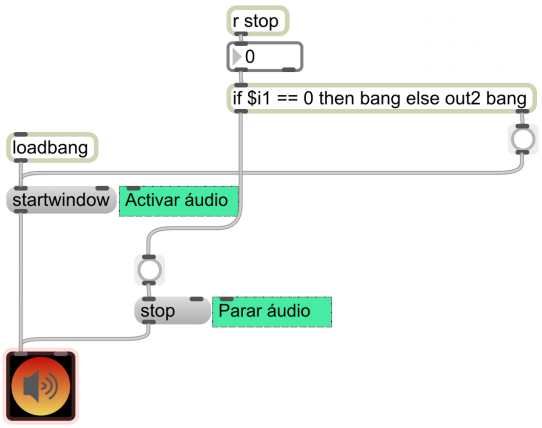


Figure 3.36: Activation of audio processing.

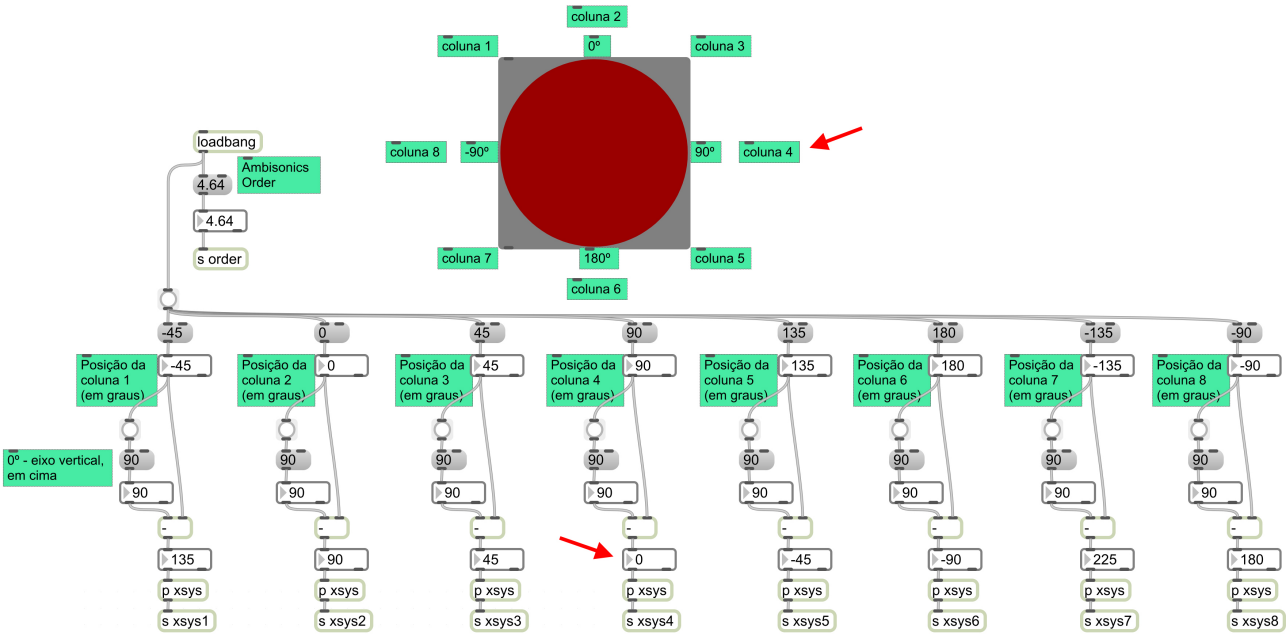


Figure 3.37: Loudspeaker positions and Ambisonics Equivalent Panning order.

signed to loudspeaker 2 (L2 in figure 3.2, in section 3.1), because this direction would coincide with that of a person facing the Kinect depth camera as well. Loudspeaker 1 (L1 in figure 3.2) would therefore have an associated angle of -45º, loudspeaker 3 (L3 in figure 3.2, too) an angle of 45º, and so on, as can be observed in the representation in the upper area of figure 3.37. Since in Processing the angle of 0º of the direction of the deictic gesture in the horizontal plane in experiments 1 and 3 follows the mathematical custom, so that the angle increases from there anticlockwise when dividing that plane into four quadrants (see figure 3.38), as explained in section 3.7.1.1, and we want it to coincide with that of the perceptual sound source, in the AEP function defined in MAX MSP Jitter the actual angle of 0º is however considered to be assigned to loudspeaker 4 and not to loudspeaker 2 (shown by the red arrows in figure 3.37).

Thus, the x_s (cosine value) and z_s (sine value) coordinates of each loudspeaker's position in the circumference, as already explained in section 3.7.1, are then computed by the sub-patch called "xsys" (see figure 3.39) and sent to the corresponding "level.maxpat" sub-patch (see figure 3.31 in section 3.7.1).

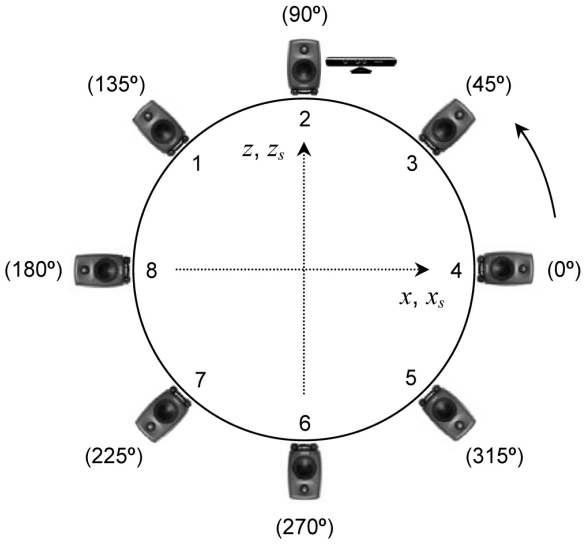


Figure 3.38: Relationship between angles, loudspeakers, and perceptual sound sources in the Ambisonics Equivalent Panning function in MAX MSP Jitter and in Processing.

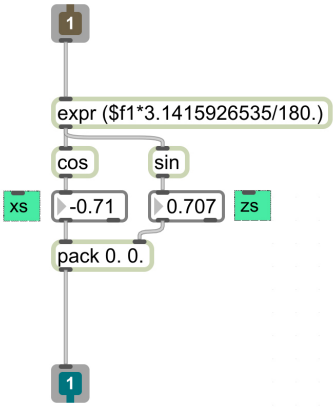


Figure 3.39: The MAX MSP Jitter "xsys" sub-patch, which computes the x_s and the z_s coordinates of each loudspeaker's position.

In turn, the x and z coordinates of the deictic gesture, which we want to be equal to those of the perceptual sound source's position in the circumference in the horizontal plane in experiments 1 and 3, are given by Processing through MaxLink's instance variable `parameters` to MAX MSP Jitter's object called `mxj` (see figure 3.40). This object is created in MAX MSP Jitter, in order to execute Java MaxLink's `link` method of class `jk`, whose arguments `parameters 0 7` mean that no inlets and seven outlets are created for this object and that `parameters` is the instance variable receiving data from Processing.

The received x and z coordinates are then packed together as a list and sent to the "level.maxpat" sub-patch (cf. figure 3.31 in section 3.7.1), so that they can be further processed (see figure 3.41).

Other parameters, such as the angle of the direction of the deictic gesture, the number of the experiment being performed (zero for experiments 1 and 2; one for experiment 3), the sound looping option (zero for no loop; one for loop), and the option for turning on or off the sound, are also received in MAX MSP Jitter from Processing.

In the part of the patch shown in figure 3.42, the researcher can observe from which direction the perceptual

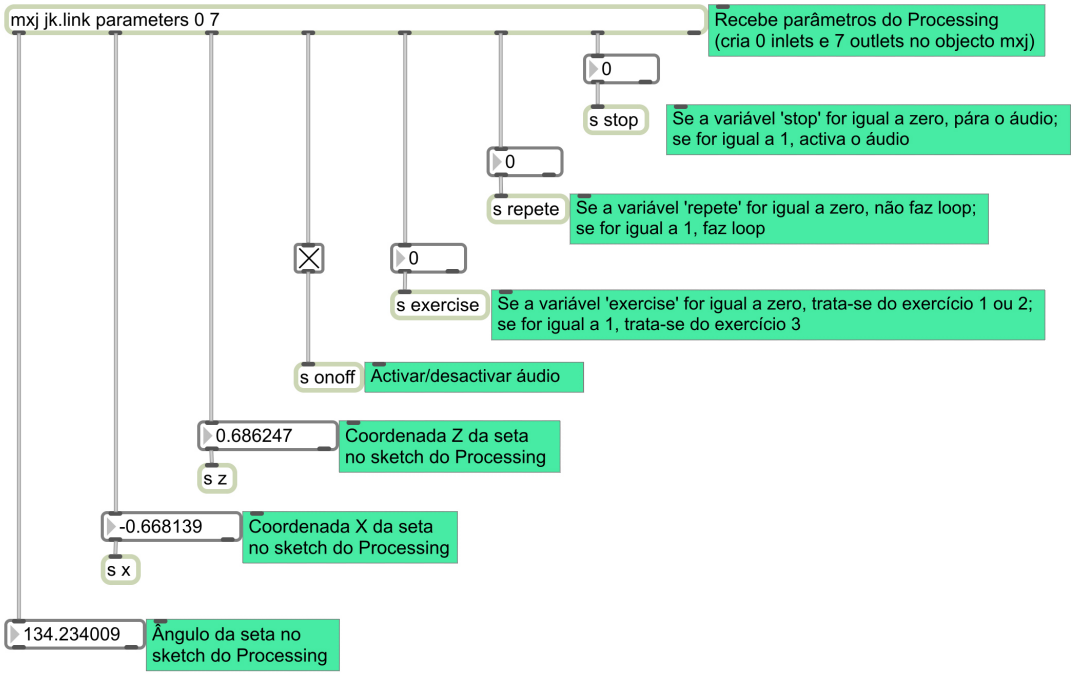


Figure 3.40: The MAX MSP Jitter connection object from Processing to MAX MSP Jitter.

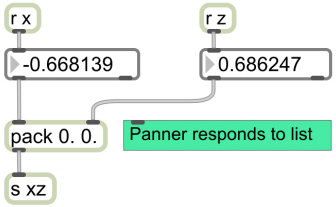


Figure 3.41: The x and z coordinates are packed together as a list.

sound source is being reproduced in the horizontal plane during the course of the experiments.

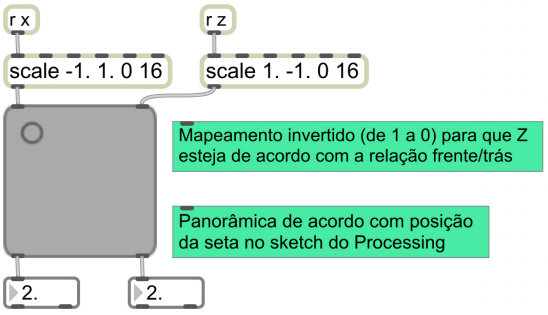


Figure 3.42: Observation of the direction from which the perceptual sound source is being reproduced in the horizontal plane.

From the moment that audio is turned on, the sub-patch named ”testsound” (see figure 3.45) is called from the part of the main patch shown in figure 3.43 and the reproduction of a randomly selected sound file out of a total of eighteen sound files, described in section 3.4, is ready to start with the correspondent AEP output amplitudes for each loudspeaker, in turn readjusted according to the previously calibrated loudspeaker main output amplitudes, as explained in section 3.2.

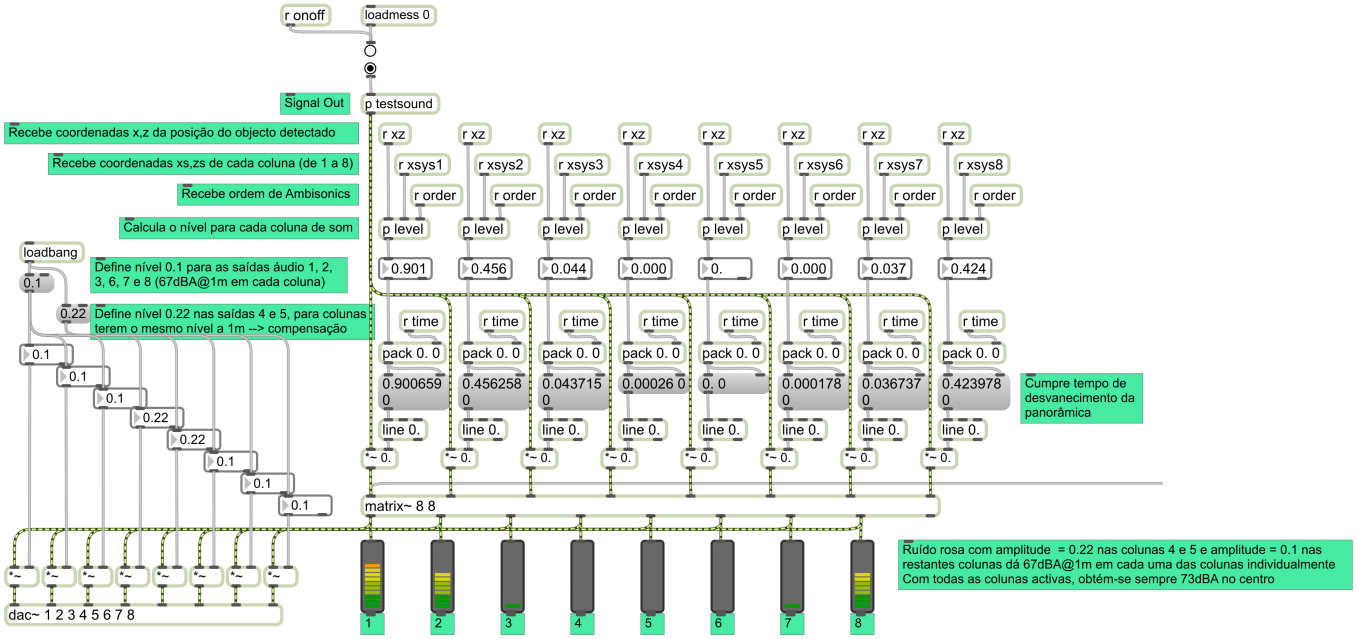


Figure 3.43: The reproduction of a randomly selected sound and the correspondent computed output amplitudes.

Figure 3.44 shows the matrix that opens a number of audio channels equal to the number of selected loudspeakers sent by the `ncol` argument of the `send` or `s` object displayed in figure 3.33.

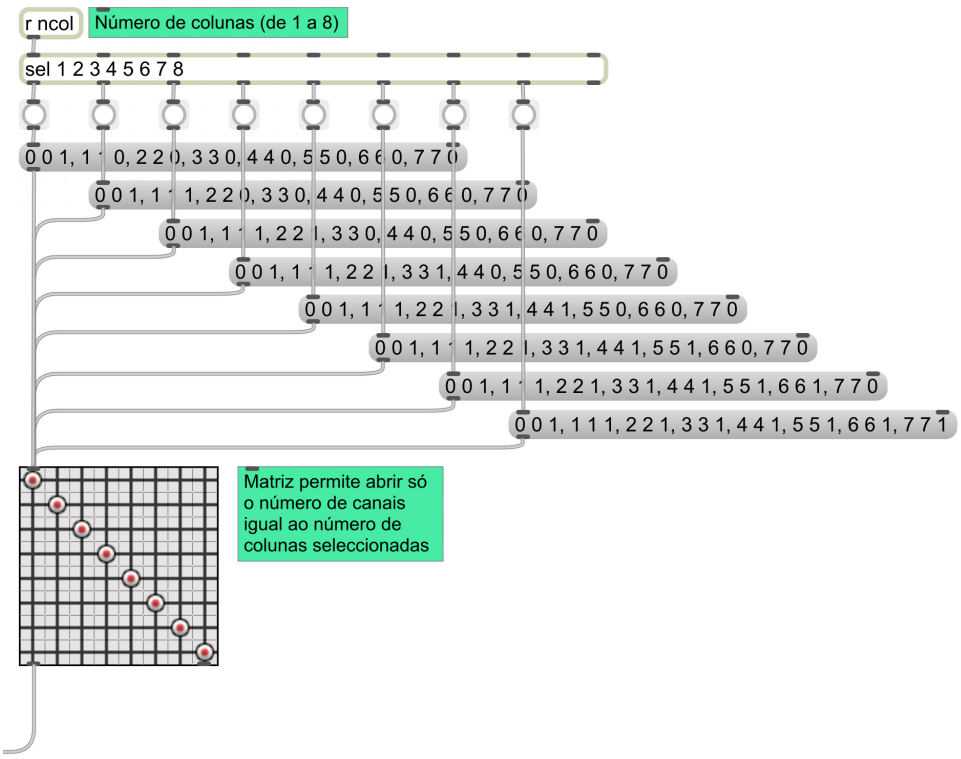


Figure 3.44: The matrix which controls the total number of output channels used in the reproduction of a sound.

Finally, the number of the current randomly selected sound file is shown to the researcher during the course of the experiments (see figure 3.46).

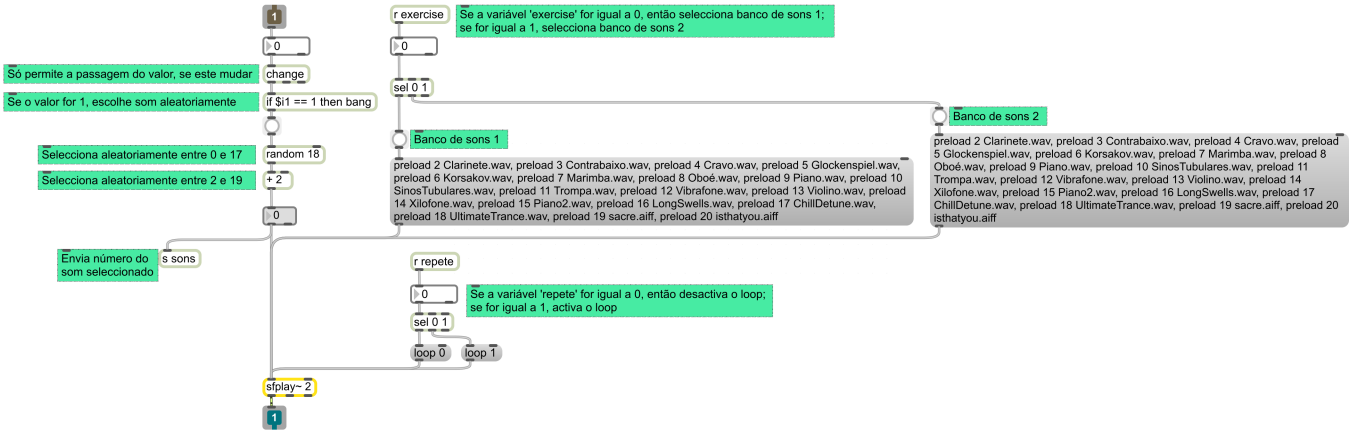


Figure 3.45: Selection of a single sound file out of eighteen to be reproduced.

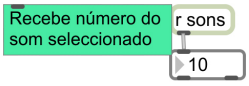


Figure 3.46: Number of the randomly selected sound file as an information for the researcher only.

3.8 Chapter Conclusions

In this chapter, we thoroughly describe the technical and material conditions under which three practical experiments were performed, as well as the equipment and the sound reproduction system used. In addition, we present and justify the selection of sounds for our research purposes, in order to conclude that the chosen musical items meet the conditions to be used in these experiments.

With regard to the type of gesture used in the investigation, it is the relatively fast downward vertical to horizontal deictic empty-handed gesture (cf. section 3.5) that we select to trigger and control the motion of sound in space. The chief advantage of this gesture is that it provides freedom of movements to the performer.

The main hypotheses of the research are presented in section 3.6.4. Furthermore, the computer software used in the research, their respective capacities and performances, and a set of relevant visual and audio information that they provide to the researcher are also described by us.

In addition, both MaxLink 0.36 library, which allows us to establish communication between Processing and MAX MSP Jitter, as well as the system for recording all on-screen information during the tests, are taken into account, too.

Finally, sections 3.7.1.1, 3.7.1.2, 3.7.1.3, and 3.7.1.4 respectively describe the processing algorithms and patches of the experiments in detail, that is, the special series of instructions carried out in a particular order in these experiments, to ensure reliable, valid, and justifiable results. Whereas reliability means to reproduce consistent findings if they are repeated on another occasion or if they are replicated by another researcher (Field, 2009, pp. 11–12, 792–793), validity has to do with the extent to which our research tools measure what they claim to measure (Field, 2009, pp. 11–12, 795).

Chapter 4

Data Results and Analysis

According to Field (2009, p. 16), in experimental research, where variables (see section 2.1.4) are manipulated so that their effects on other variables can be observed, ”the role of statistics is to discover how much variation there is in performance, and then to work out how much of this is systematic and how much is unsystematic”, that is, to work out how much of this variation is respectively due to ”differences in performance created by a specific experimental manipulation” by the experimenter to all of the participants in one experimental condition, but not in another one, or due to ”... differences in performance created by unknown factors” that exist between the experimental conditions, not attributable to the effect that is being studied (cf. section 3.6).

Thus, in the next sections we will statistically analyse the collected data of our experimental work, aiming to find some meaning in them, and present the results thereof, using the IBM SPSS software, version 22, originally known as Statistical Package for the Social Sciences (SPSS), although it is also currently used in other fields (Wikipedia, 2017b). However, even before that, we will briefly describe the method we have followed.

4.1 Method of Analysis

All questions in our Inquiry Mode Questionnaire (InQ), which can be found in appendix A, are of the category type, which means that qualitative variables, such as binary, nominal, or ordinal variables, are involved. Whereas a binary or dichotomous variable has just two independent categories that can be selected individually as a response, such as the binary variable *gender* (question number 2, in part 1) from which one can usually only choose *male* or *female*, a nominal variable like *educational qualifications* (question number 4, in part 1) has more than two categories (Field, 2009, p. 8). However, there is no intrinsic order between the categories in binary and nominal variables. With respect to ordinal variables, categories are logically ordered, although the size of the difference between categories or the distances between the rankings is not known (Field, 2009, pp. 8–9). Here we point out that nominal and ordinal scales are two of the four scales of measurement proposed by Stanley Smith Stevens (Stevens, 1946; Freixo, 2012, p. 234; Harpe, 2015, p. 837) (cf. section 2.1.3.2).

Thus, firstly and whenever possible, we have coded or assigned a number to each possible response, in order to proceed with the abbreviated identification of information units with common characteristics (Reis,

2012, p. 19). For instance, since participants were asked to indicate their age in ordered categories in question number 1, in part 1, which makes *age* an ordinal variable, we have assigned a *0* for a non-answered question (”Did not answer”), a *1* for the ”15 to 24” age group, a *2* for the ”25 to 34” age group, and so on.

Most items or questions about experiments 1, 2, and 3 in part 2 of the InQ (see appendix A and figures 4.1, 4.2, and 4.3) should be answered using five response alternatives numbered from 1 to 5 (1 – I totally disagree, 2 – I disagree, 3 – Not always / Sometimes, 4 – I agree, and 5 – I totally agree), known as 5-point *Likert* alternatives (Likert, 1932; Boone, Jr. & Boone, 2012), which caused the variables involved to be ordinal, because of the ”greater than” relationship expressed by the assigned numbers. Since we wanted to essentially compare opinions of participants on specific issues between the experiments and study their correlations, and had therefore no intention to combine the responses from the questions into a composite scale, as Likert did originally to create an attitudinal measurement scale, these questions were considered as individual or single, also called Likert-type, questions (Boone, Jr. & Boone, 2012).

Experiment 1	1.	You quickly understood the experiment
	2.	It was easy to interact with the installation
	3.	It was easy to define the origin of sound
	4.	The suggested gesture is adequate to the experiment
	5.	You felt immediate control over sound
	6.	The system's response to gesture was immediate
	7.	The system's response to gesture was slow
	8.	You felt surrounded by sound in the installation
	9.	Your gesture coincided with the origin of sound
	10.	Estimate the time needed to understand the experiment
	11.	Estimate the time needed to control sound
	12.	Estimate the time of the system's response to gesture

Figure 4.1: Questions of part 2 of the Inquiry Mode Questionnaire (InQ), related to experiment 1.

Experiment 2	1.	You quickly understood the experiment
	2.	It was easy to interact with the installation
	3.	The proposed experiment confused me
	4.	It was easy to define the origin of sound
	5.	The suggested gesture is adequate to the experiment
	6.	You felt immediate control over sound
	7.	The system's response to gesture was immediate
	8.	The system's response to gesture was slow
	9.	You felt surrounded by sound in the installation
	10.	Your gesture coincided with the origin of sound
	11.	Estimate the time needed to understand the experiment
	12.	Estimate the time needed to control sound
	13.	Estimate the time of the system's response to gesture

Figure 4.2: Questions of part 2 of the Inquiry Mode Questionnaire (InQ), related to experiment 2.

In questions number 10, 11, and 12 in experiment 1, and in questions number 11, 12, and 13 in experiments 2 and 3, where the participants were asked to freely estimate times related to the respective experiments, the variables were expected to be of the ordinal type as well, rather than of the interval or ratio type, because neither a unit nor a scale or range to be used to estimate these times was given. Thus, these variables consisted of elements provided by the participants themselves. Whereas in interval variables equal intervals

Experiment 3	1.	You quickly understood the experiment
	2.	It was easy to interact with the installation
	3.	It was easy to define the origin of sound
	4.	The suggested gesture is adequate to the experiment
	5.	Your gesture coincided with the origin of sound
	6.	The system's response to gesture was immediate
	7.	The system's response to gesture was slow
	8.	You felt surrounded by sound in the installation
	9.	You managed to locate sound
	10.	You quickly located sound
	11.	Estimate the time needed to understand the experiment
	12.	Estimate the time of the system's response to gesture
	13.	Estimate the time needed to locate sound

Figure 4.3: Questions of part 2 of the Inquiry Mode Questionnaire (InQ), related to experiment 3.

between points on a "... scale represent equal differences in the property being measured" (Field, 2009, p. 9), in a ratio variable ratios of scores along a scale should make sense and a zero value should exist (Field, 2009, p. 9). We point out here that the interval and ratio scales are the remaining two scales which, together with the aforementioned ordinal and nominal scales, were also proposed by Stanley Smith Stevens (Stevens, 1946; Freixo, 2012, p. 234; Harpe, 2015, p. 837) (see section 2.1.3.2).

The estimated times should help us getting a sense of each participant's time perception in each experiment and, at the same time, detecting inconsistencies in responses. Questions 1 and 10 of experiment 1 have to do with each other, as well as questions 5 and 11 and questions 6, 7, and 12. The same can be observed in questions 1 and 11, questions 6 and 12, and questions 7, 8, and 13 in experiment 2. In turn, in experiment 3, questions 1 and 11, questions 6, 7, and 12, and questions 10 and 13 have also to do with each other.

Afterwards, we transferred the information from the filled questionnaires (see appendix E.11) to SPSS (cf. appendix E.12.1), as well as their respective coding or assigned numbers, as described above.

Questions 6 and 7 relating to experiments 1 and 3, and questions 7 and 8 concerning experiment 2 had purposely inverted senses, as a means of determining inconsistencies in responses. In order to standardize the scales of these questions, so that statistical evaluation could be performed, the last question of each of these pairs was inverted and each possible response was recoded in SPSS, that is, "The system's response to gesture was slow" was inverted to "The system's response to gesture was fast", becoming analogous to the previous question "The system's response to gesture was immediate."

In order to select and fit a statistical model to the data as accurately as possible, so that we could test our original hypotheses or predictions, and with the purpose of getting a better overall idea of the collected data, raw data or responses of all participants were individually summarized and described firstly in the form of frequency tables or frequency distributions (Field, 2009, p. 18), and in the form of bar charts, which is typical of an univariate analysis, where one question or variable is examined at a time (Field, 2009, p. 585) (see appendix E.12.2). Frequency tables or frequency distributions consist of at least a column with all the observed modalities or values that a variable presents and another column with the corresponding number of occurrences for each modality or value of the variable (Spiegel, 2000, p. 6; Reis, 2012, p. 46, 48–51), although percentages are most commonly used expressing relative frequencies (Spiegel, 2000, p. 9; Reis, 2012, p. 47). Bar charts are graphs in which frequency distributions are plotted using rectangles, whose length

is proportional to the observed frequencies or relative frequencies, so that the x-axis represents the observed modalities or values of a variable and the y-axis holds the corresponding number or percentage of occurrences for each modality or value of the variable (Reis, 2012, p. 28).

Then descriptive statistical measures were computed for every question (Reis, 2012, pp. 15, 63) (see appendix E.12.3), even without assessing whether they had any meaning or not, a matter that would be evaluated afterwards case by case, according to the nature of the involved variables: 1) central tendency measures of localization of data, that is, measures that attempt to describe a dataset by determining the central or middle position of that dataset: *mean*, *standard error of mean*, *median*, and *mode*; 2) non-central tendency measures of localization of data: *minimum*, *maximum*, *quartiles*, and *percentiles*; and 3) data dispersion measures, that is, measures that show "the degree to which data tend to disperse around a central value" (Spiegel, 2000, p. 16) and that are useful to verify the representativeness of localization measures (Reis, 2012, p. 97): *range*, *interquartile range*, *variance*, *standard deviation*, *coefficient of variation* or *relative standard deviation*, *skewness*, and *kurtosis*. It is naturally beyond the scope of our thesis to explain how all these measures are calculated in detail. In this sense, we suggest reading Hines et al. (2003), Spiegel (2000), Field (2009), Reis (2012), and Pestana & Gageiro (2014), among many other authors. However, we will not abstain from briefly explaining the meaning of each measure.

Therefore, with respect to 1), the *mean* is an average score or representative value of a dataset, which tends to be in the middle of the set (Spiegel, 2000, p. 12; Field, 2009, p. 22), although it may not be the case if the distribution of data is highly asymmetric or skewed due to some extreme values (Reis, 2012, p. 84). In addition, the mean may not make any sense if the data is not numeric or if the mean value does not match a particular value of the variable (Field, 2009, p. 22; Reis, 2012, p. 71), which is the case in our study because our variables are of the category type. The *standard error of mean* is an estimate of the average variability or deviation of the data distribution off the mean (Field, 2009, p. 794), which also does not make any sense in our investigation because it depends on the mean. In turn, the *median* of a dataset containing an odd amount of ordered numbers is the central value or middle score of the set which divides the set into a lower and a higher half, so that the extreme values do not affect its value, which is the case in strongly asymmetric or skewed distributions (Reis, 2012, p. 85). If the dataset is made up of an even amount of ordered numbers, the median is the arithmetic mean of the two central values (Spiegel, 2000, p. 14; Field, 2009, p. 21). The *mode* is the value or score that occurs most frequently in a dataset, that is, the value of the largest number of observations (Spiegel, 2000, p. 15; Field, 2009, p. 21). It is therefore not influenced by extreme values (Reis, 2012, p. 83). However, it may not exist or may not be unique (Spiegel, 2000, p. 15; Reis, 2012, p. 83).

Regarding 2), the *minimum* is the smallest score or value and the *maximum* is the largest score or value within a dataset of a given variable (Field, 2009, p. 23). *Quartiles* "are the three values that split the sorted data into four equal parts" (Field, 2009, p. 23) and *percentiles* are values that divide data into 100 equal parts (Reis, 2012, p. 87).

Finally, concerning 3), the *range* is the difference between the maximum and the minimum values of a given variable, being therefore insensitive to the intermediate values, but extremely sensitive to the extreme scores (Field, 2009, p. 23; Reis, 2012, p. 98), and the *interquartile range* is the difference between the upper and lower quartiles, corresponding to the range covering 50% of the central observations (Field, 2009, p. 24; Reis,

2012, p. 99). Thus, it is not influenced by the extreme values, which correspond to half of the observed values, and it can be used in highly asymmetric or skewed distributions (Reis, 2012, p. 109). In turn, the *variance* corresponds to the sum of the squared errors or squared deviations or distances of the observations made from the mean, "... divided by the number of values on which the sum of squares is based minus 1", so that the error can be estimated for a whole population and not only for a sample of it (Field, 2009, pp. 36–37, 796). In other words, the variance represents the average error that exists between the mean and the observations made (Field, 2009, p. 37). Since the variance has the disadvantage of representing the square of the units in which the variable is defined, the *standard deviation* is more commonly used, which is defined as the square root of the variance (Field, 2009, p. 37). The standard deviation thus shows in a better way the degree of variation or dispersion that exists with respect to the mean. The *coefficient of variation* or *relative standard deviation* is a measure which can be used to compare the dispersion between two distributions defined in different units or having different means, so that it is determined by the percentage ratio between the standard deviation and the mean (Reis, 2012, p. 107). Variance, standard deviation, and coefficient of variation were obviously not used in our study due to their dependence on the mean. The *skewness* is a measure of the symmetry of a frequency distribution. The *kurtosis* "... measures the degree to which scores cluster in the tails of a frequency distribution" (Field, 2009, p. 788).

To be more precise in our interpretations, inferential statistics, used to make inferences or generalizations about a population from the collected data for a sample, was performed at the end to test our hypotheses, that is, to "... help us to confirm or reject our predictions", using probability to verify if values or scores were obtained by chance (Field, 2009, p. 49). In cases where they were obtained by chance, they were said to be non-significant and the hypotheses were not generally accepted as true. This means that "... the effect is not big enough to be anything other than a chance finding..." (Field, 2009, p. 53). Thus, we used a probability of 5% = .05, known as the α-level, of something occurring by chance for all statistical tests, which means that a criterion of 95% confidence, suggested by Sir Ronald Aylmer Fisher (February 17, 1890 - July 29, 1962)[1], was applied, so that "... only when we are 95% certain that a result is genuine (i.e. not a chance finding) should we accept it as being true" (Field, 2009, p. 50).

If interval or ratio variables had been involved in our hypotheses, then we would have applied parametric techniques based on the normal distribution (Field, 2009, p. 132), such as the Analysis of Variance (ANOVA) and t-tests, in order to obtain results as accurate as possible. Parametric techniques assume "... that sample data comes from a population that follows a probability distribution based on a fixed set of parameters" (Wikipedia, 2017a) described by statisticians. Thus, four assumptions would have to be met (Field, 2009, pp. 132–133): 1) Data would have to be normally distributed, that is, data would have to be "... distributed symmetrically around the centre of all scores" (Field, 2009, p. 18), presenting therefore a skew and a kurtosis equal to zero and a graph "... characterized by the bell-shaped curve..." (Field, 2009, p. 18); 2) Variance would have to be homogeneous or the same throughout the data; 3) Data would have to be of the interval or ratio type; and 4) Independence of the behaviour between different participants would have to be satisfied.

Nevertheless, non-parametric techniques were the most suitable ones in our study in order to get the most accurate results possible, such as the Friedman's ANOVA test and the Wilcoxon signed-rank test, because we

[1] Retrieved 22/08/2017, from `https://en.wikipedia.org/wiki/Ronald_Fisher`

were using ordinal variables and some of the above mentioned parametric assumptions were not met, although many authors and researchers argue that ordinal variables can be treated as if they were of the quantitative type, that is, of the interval or ratio type, as it is common in educational research, so that parametric techniques could have been applied anyway due to their robustness to violations of their assumptions, but in our case with less accurate results (Norman, 2010; Sullivan & Artino Jr, 2013).

Finally, whenever possible, bivariate correlation analysis was also carried out. Thus, pairs of questions were analysed with the Kendall's and Spearman's correlation coefficient tests, so that we could compare them in order to see how they interacted with each other or differed from each other.

4.2 Participants and Demographic Information

In order to carry out the three practical experiments described in section 3.6, which have given support to the study of the correlation between gesture and localization of sound sources in space, it was necessary to ask a large number of people (with or without any musical background) to participate voluntarily in the research without constraints. This was mainly accomplished by sending emails to friends, to students from different subject areas studying, and to colleagues working, at the School of Arts at the Portuguese Catholic University (EA-UCP), at the School of Music and Performing Arts at the Polytechnic Institute of Porto (ESMAE-IPP), and at the Music Academy of Espinho (AME). A brief explanation about the principal aim of the research and about the procedure of the practical experiments was included in the emails (see appendix B). Furthermore, it was also crucial to talk to friends and relatives, and to inform as many people as possible about the research by placing bills on the walls of the EA-UCP a few days before the experiments were carried out in the Motion Capture Laboratory (cf. chapter 3), so that a random sample of participants could be gathered.

Eventually, 43 volunteers were counted up. According to the 43 filled in Inquiry Mode Questionnaires (InQ) (see appendix E.11) and to the respective demographic information (see figure 4.4), there were at first 8 female and 34 male participants, and 1, who did not fill in the corresponding gender field. But after checking the names on the previously arranged timetable of the participants (see appendix C, where names have been replaced by numbers to ensure anonymity) and after analysing all videos, recorded during the practical sessions, the conclusion is that the person who did not fill in the gender field had to be actually female. So, the end result is that 9 (20.9%) female and 34 (79.1%) male volunteers contributed to the study (see figure 4.5). In this binary variable (cf. section 4.1), the mode of 2 (see appendix E.12.3), taking into account that 2 represents the male gender, confirms that the majority of participants was of the male gender.

Demographic Information	1.	Age
	2.	Gender
	3.	Which hand do you use for writing?
	4.	Educational qualifications
	5.	Do you have any musical knowledge?
	6.	Do you have hearing handicaps?

Figure 4.4: Questions of part 1 of the Inquiry Mode Questionnaire (InQ).

Figure 4.6 shows that from a total of 43 participants, 18.6% were aged between 15 and 24, 39.5% were

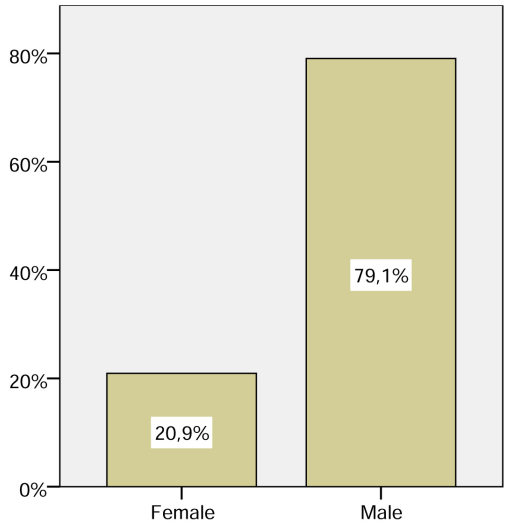


Figure 4.5: Gender, taking into account all participants.

aged between 25 and 34, 16.3% were aged between 35 and 44, 16.3% were aged between 45 and 54, 7.0% were aged between 55 and 64, and 2.3% were 65 or over 65 years old. Since age is an ordinal variable, as already referred to in section 4.1, the median of 2 and the mode of 2 (cf. appendix E.12.3) are meaningful, so that they confirm respectively that the 25 to 34 years' category is both the central and the most frequent one of the dataset.

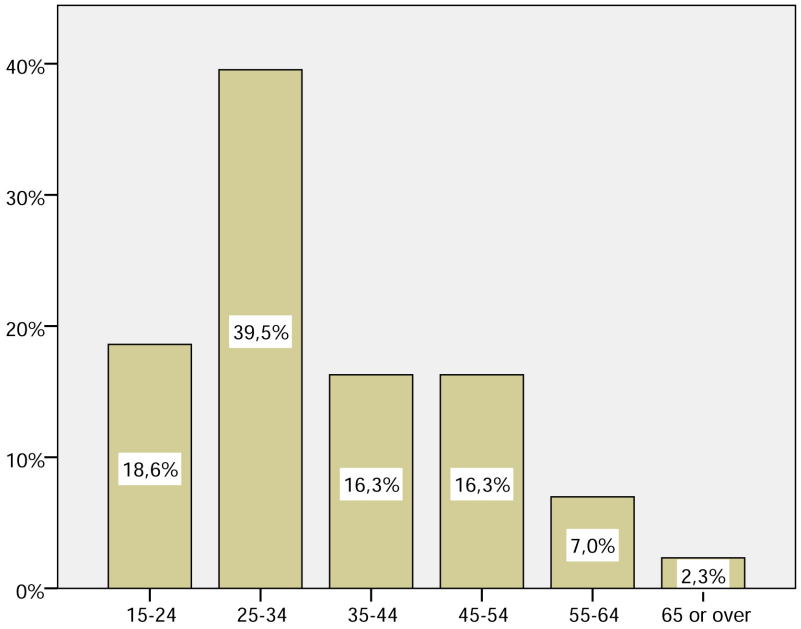


Figure 4.6: Age, taking into account all participants.

Taking account of the age by gender and relating it to the female gender, 22.2% of the volunteers were aged between 15 and 24, 22.2% were aged between 25 and 34, 22.2% were aged between 35 and 44, and 33.3% were aged between 45 and 54 (see figure 4.7). The median of 3 in table 4.1 shows that the central value of this dataset corresponds to the 35 to 44 years' category and the mode of 4 strengthens the idea that the 45 to 54 years' category for the female gender occurs most frequently.

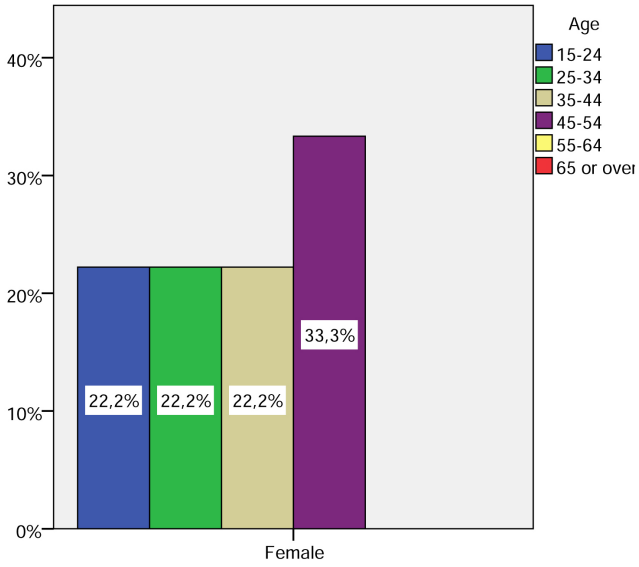


Figure 4.7: Age of all participants by gender: female.

Table 4.1: Median and mode of age by gender, taking into account all participants.

Gender of participants			Age
1 Female	N	Valid	9
	Median		3,00
	Mode		4
2 Male	N	Valid	34
	Median		2,00
	Mode		2

Concerning the male gender, 17.6% were aged between 15 and 24, 44.1% were aged between 25 and 34, 14.7% were aged between 35 and 44, 11.8% were aged between 45 and 54, 8.8% were aged between 55 and 64, and 2.9% were 65 or over 65 years old (see figure 4.8). Both the median and mode of 2 in table 4.1 confirm respectively that the 25 to 34 years' category is the central range of this dataset and that it occurs most frequently for the male gender.

Since: 1) the 45 to 54 years' category for the female gender occurred most frequently; 2) the remaining women ($22.2\% + 22.2\% + 22.2\% = 66.6\%$) were younger than 45; 3) a vast majority of men was aged between 25 and 34 years or less; and 4) the percentage of men aged over 34 years was increasingly small, we consider that the decrease of the auditory sensibility with age, more intensively felt by men than by women, essentially at higher frequencies, as referred to in section 2.1.3.2, is by itself not a problem to take into account in the data analysis.

From a total of 43 individuals, 6 did not answer the question: "Which hand do you use for writing?". It appears that they did not fill in the corresponding field, because it is found on the right side of the form, which might have caused them to ignore it (see appendix A, question 3 in part 1). Consequently, leaving out the missing values and considering the valid answers only, 6 participants (16.2%) replied that they use the left hand and 31 (83.8%), therefore the majority, responded that they use the right hand for writing (see figure 4.9). This result is supported by the mode of 2 (cf. appendix E.12.3), where 2 represents the right hand. After observing the videos of the experiments of all participants, we concluded that only 4 (9,3%) of them used the

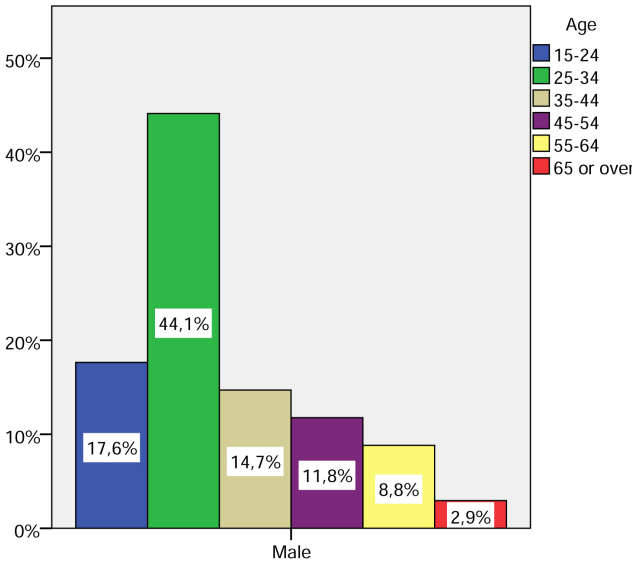


Figure 4.8: Age of all participants by gender: male.

left hand's deictic gesture (cf. section 2.6), although one of these four volunteers started and changed from the right to the left hand still during the beginning of the first experiment, and 39 (90,7%) participants used the right hand's deictic gesture. Consequently, there were at least 2 volunteers (out of 6 left-handed valid answers) who did not use the hand they usually use for writing in the experiments.

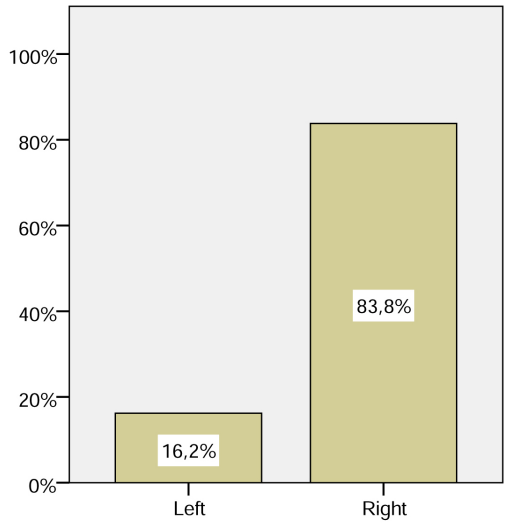


Figure 4.9: Taking into account all participants: "Which hand do you use for writing?"

Question number 4 of part 1, concerning the educational qualifications, was answered by 42 participants and 1 did not answer correctly, so it was considered invalid. Thus, 11.9% said that they had A levels (12th grade), 45.2% had a graduation, 4.8% had a postgraduate qualification, 23.8% answered that they had a master's degree, 11.9% had a PhD, and 2.4% replied that they had another qualification, that is, a Diploma of Higher Studies in English by the University of Cambridge (see figure 4.10). The mode of 3 (see appendix E.12.3), representing the group of participants with a graduation, confirmed that this group was the most frequent among all participants.

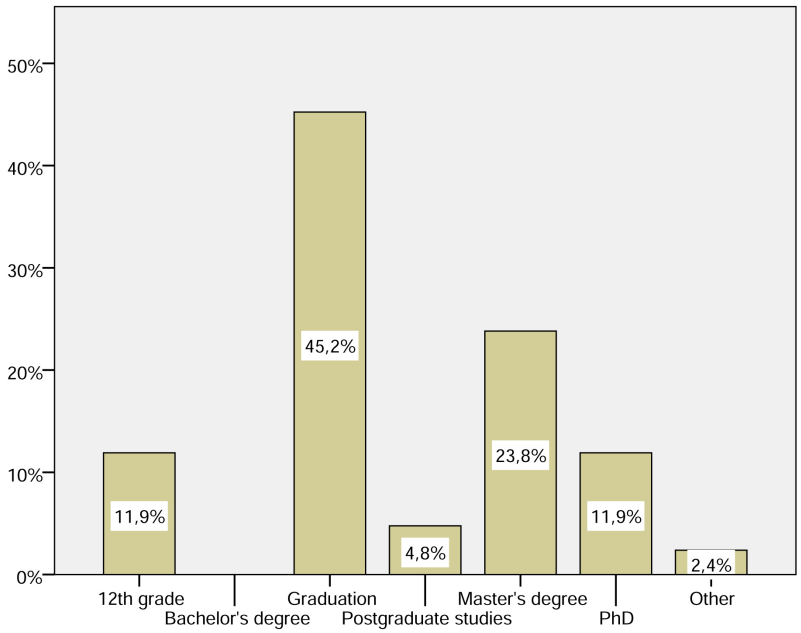


Figure 4.10: Educational qualifications of all participants.

67.4% of all volunteers answered ”yes” and 32.6% ”no” to the question: “Do you have any musical knowledge?” (see figure 4.11). The majority had therefore musical knowledge, which is also confirmed by the mode of 1 (cf. appendix E.12.3), where 1 represents the affirmative answer to this question. This was a very important question for our investigation, because it would allow us to test hypothetical deviations in the results based on musical knowledge, as explained in section 4.2.1. Musical knowledge indicated by the participants derived from the fact that they had: 1) studied music in a Music Conservatory; 2) studied Acoustics, Audio Technology, Classical Guitar, Musical Composition, Musical Education, Organ, Piano, Singing, Sound Design, Sound for Cinema; 3) a Higher Education in Music, were graduated in Piano, were graduated in Singing, had a postgraduate degree in Musical Sciences, a Master’s degree in Musical Education, a Master’s degree in Musical Interpretation; or 4) were music composers or musicians, such as pianists or amateur percussionists.

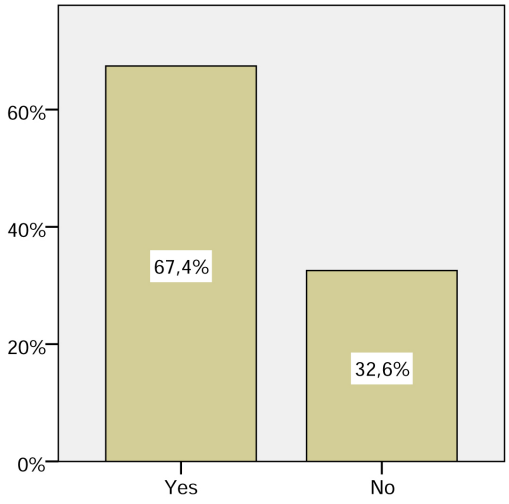


Figure 4.11: Taking into account all participants: ”Do you have any musical knowledge?”

The last question of part 1 was: “Do you have hearing handicaps?”. This question was also very important for the research, because the validation of the data obtained from the questions in the second part of the inquiry mode questionnaire would be fully dependent on the listening conditions of the participants themselves. Thus, 7.0% of all participants answered ”yes” and 93.0% replied ”no” to this question (see figure 4.12). This result is substantiated by the mode of 2, where 2 represents a negative answer. With regard to the affirmative answers, one of the volunteers, a male participant, did not mention the kind of difficulty he had. The types of hearing difficulties pointed out by the other participants were partial loss of hearing (*hypoacusis*) in the right ear and difficulties in the identification of sounds. Since the majority of participants answered ”no” to this question, it is reasonable to consider that the remaining data are reliable for analysis, although a standard audiometric analysis of each participant could have supported this assumption more consistently, which we did not have opportunity to carry out.

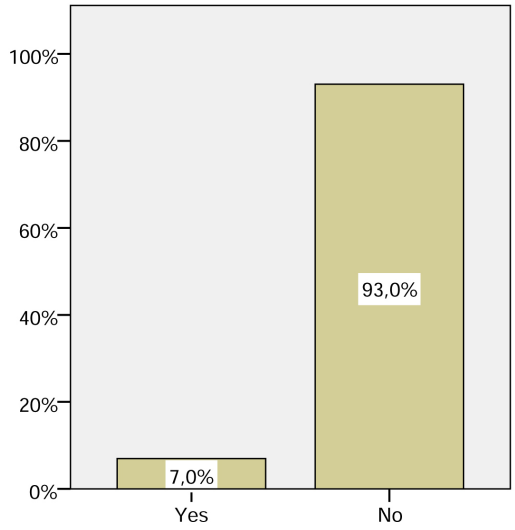


Figure 4.12: Taking into account all participants: ”Do you have hearing handicaps?”

4.2.1 Groups of Participants

At the time when the research plan was presented and approved by the Scientific Board of the School of Arts of the Portuguese Catholic University, we intended to gather at least 20 participants with no musical knowledge and 20 volunteers with musical knowledge from the sample containing all participants, with the purpose of later testing hypothetical deviations in the results of the analysed data based on musical knowledge. Although our participants with musical knowledge might not be, even so, specialized listeners, we hypothesized that they were more likely to be focused on listening when compared to those without such knowledge. The suggested number of participants in each group was chosen due to statistical demand. According to Spiegel (2000, pp. 44 and 61), Field (2009, p. 42), and Laureano (2013, p. 25), the number of samples n in small sample data sets is considered to be less than 30, and it should be as high as possible to guarantee better statistical results. So we considered that it would be reasonable for us to choose a number above half 30, that is, 20.

Unfortunately this objective could not be totally achieved, but we managed to form one group of 14 participants (32.6%) who answered that they had no musical knowledge and another group of 29 participants

(67.4%) who replied that they had musical knowledge, as already mentioned in section 4.2.

We will now describe these two groups separately.

4.2.1.1 Group of Participants without any Musical Knowledge

In the group of 14 participants without any musical knowledge, 21.4% of the volunteers were female and 78.6% were male (see figure 4.13). In this binary variable (see section 4.1), the mode of 2 (cf. appendix E.12.5), taking into account that 2 represents the male gender, confirms that the majority of participants without any musical knowledge was of the male gender. This result is similar to that in which we considered all participants (see section 4.2).

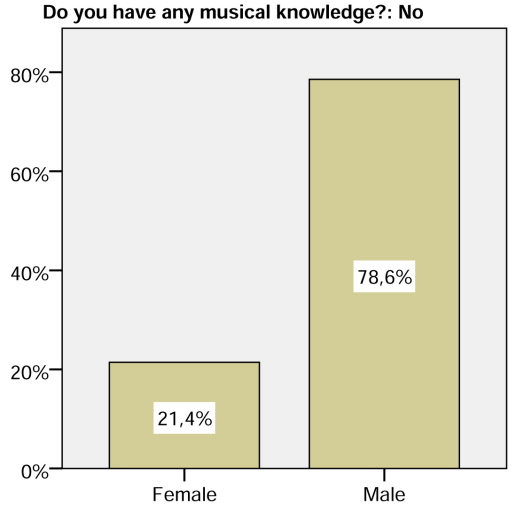


Figure 4.13: Gender (group without any musical knowledge).

Figure 4.14 shows that 21.4% were aged between 15 and 24, 42.9% were aged between 25 and 34, 14.3% were aged between 35 and 44, 14.3% were aged between 55 and 64, and 7.1% were 65 or over 65 years old. Since age is an ordinal variable (cf. section 4.1), the median of 2 and the mode of 2 (see appendix E.12.5) are meaningful, so that they confirm respectively that the 25 to 34 years' category is both the central and the most frequent one of the dataset, just as when we took into account all participants in section 4.2.

Considering the age by gender and relating it to the female gender, 33.3% of the volunteers were aged between 15 and 24, 33.3% were aged between 25 and 34, and 33.3% were aged between 35 and 44 (see figure 4.15). The median of 2 in table 4.2 shows that the central value of this dataset corresponds to the 25 to 34 years' category and the multiple modes reinforce the idea that there is no age category that occurs most frequently than others, that is, the age categories occur equally.

With reference to the male gender, 18.2% were aged between 15 and 24, 45.5% were aged between 25 and 34, 9.1% were aged between 35 and 44, 18.2% were aged between 55 and 64, and 9.1% were 65 or over 65 years old (see figure 4.16). Both the median and mode of 2 in table 4.2 confirm respectively that the 25 to 34 years' category is the central range of this dataset and that it occurs most frequently for the male gender. This outcome is identical to that in which we considered all participants (cf. section 4.2).

Once again, the decrease of the auditory sensibility with age, mentioned in section 2.1.3.2, is by itself not

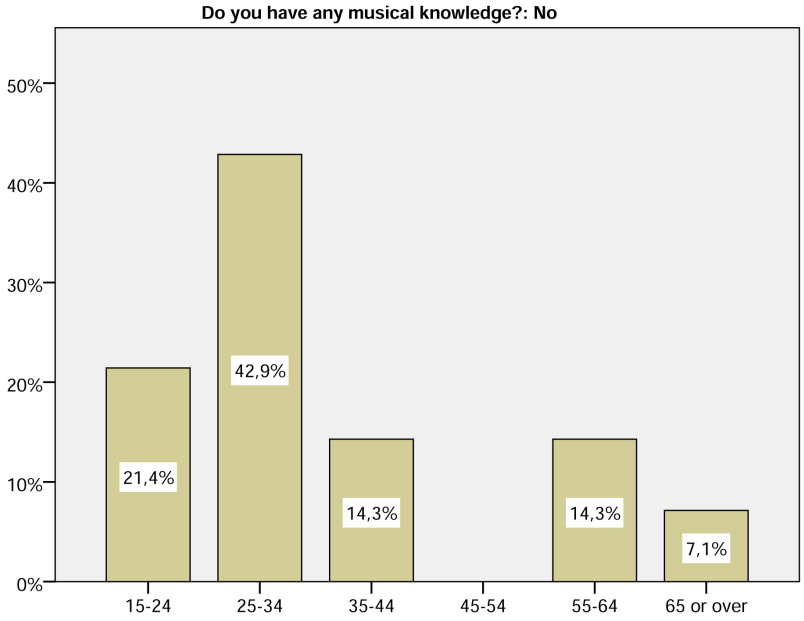


Figure 4.14: Age (group without any musical knowledge).

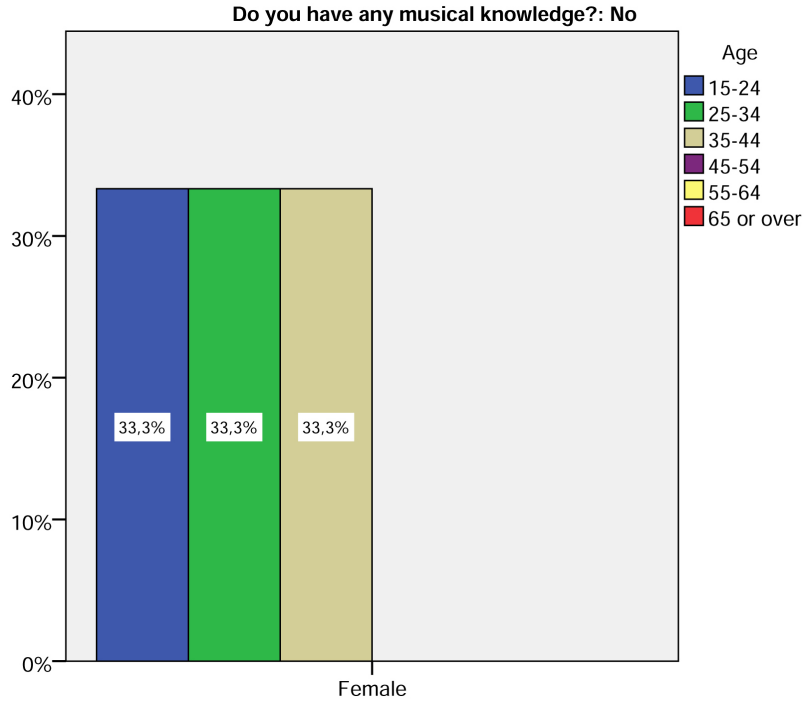


Figure 4.15: Age by gender: female (group without any musical knowledge).

a problem to consider in the data analysis of this group of participants, because the majority of women (33.3% + 33.3% = 66.6%) and men (18.2% + 45.5% = 63.7%) was under 35 years old, as can also be reinforced by the mode of 2 in table 4.3, representing the range of ages between 25 and 34 years, for the entire group of participants without any musical knowledge.

In terms of the question: “Which hand do you use for writing?”, 1 volunteer did not reply. As a result, 15.4% answered that they used the left hand and 84.6% answered that they used the right hand for writing (see figure 4.17). These outcomes are very similar to those in section 4.2, when considering all participants,

Table 4.2: Median and mode of age by gender, taking into account participants by musical knowledge.

Do you have any musical knowledge?	Gender of participants			Age
1 Yes	1 Female	N	Valid	6
		Median		3,50
		Mode		4
	2 Male	N	Valid	23
		Median		2,00
		Mode		2
2 No	1 Female	N	Valid	3
		Median		2,00
		Mode		1[a]
	2 Male	N	Valid	11
		Median		2,00
		Mode		2

a. Multiple modes exist. The smallest value is shown

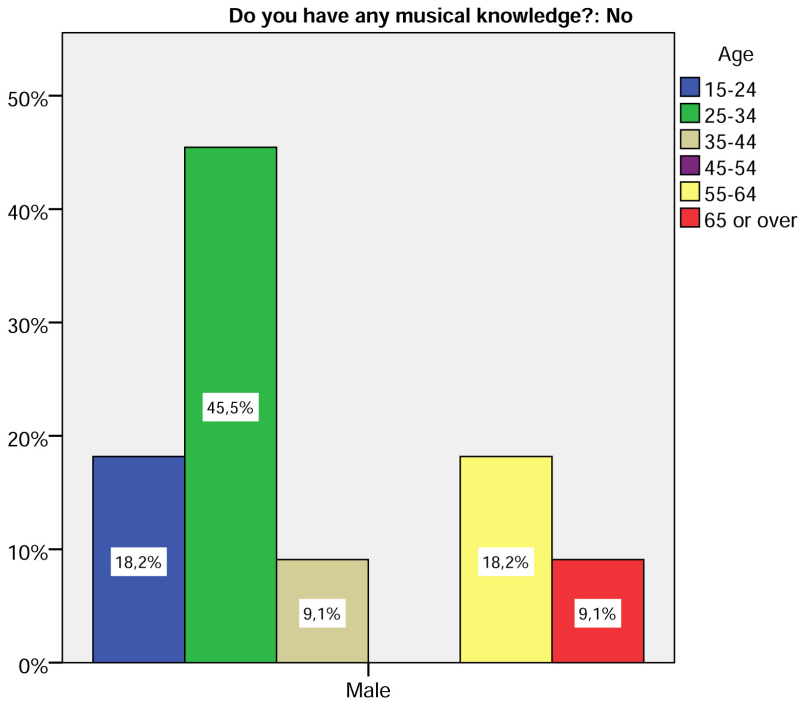


Figure 4.16: Age by gender: male (group without any musical knowledge).

and supported by the mode of 2 (see appendix E.12.5), where 2 represents the right hand.

Figure 4.18 shows that 21.4% said that they had A levels (12th grade), 35.7% had a graduation, 28.6% answered that they had a master's degree, 7.1% had a PhD, and 7.1% replied that they had another qualification, that is, a Diploma of Higher Studies in English by the University of Cambridge. The mode of 3 (cf. appendix E.12.5), representing the group of participants with a graduation, just like in section 4.2 when considering all participants, confirms that this group was the most frequent among the participants without any knowledge.

The last question of part 1 was: "Do you have hearing handicaps?". 7.1% of the participants answered "yes" and 92.9% replied "no" (see figure 4.19). The volunteer, a male participant, who did not mention what kind of hearing difficulties he had, was still included in this group. The mode of 2, where 2 represents a negative answer, substantiates these results, which are very similar to those obtained in section 4.2 when considering

Table 4.3: Mode of age in the group of participants without any musical knowledge

Do you have any musical knowledge?			Age
No	N	Valid	14
	Mode		2

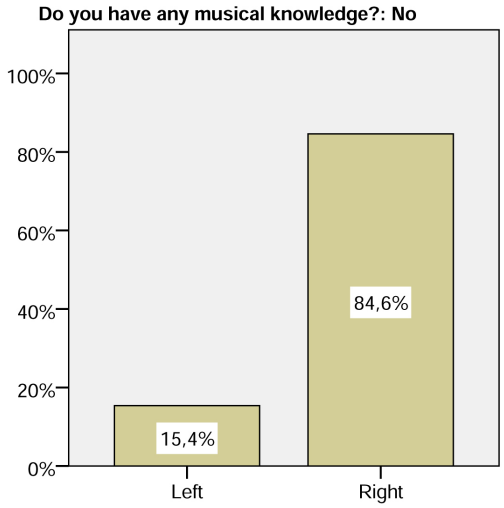


Figure 4.17: Which hand do you use for writing? (group without any musical knowledge)

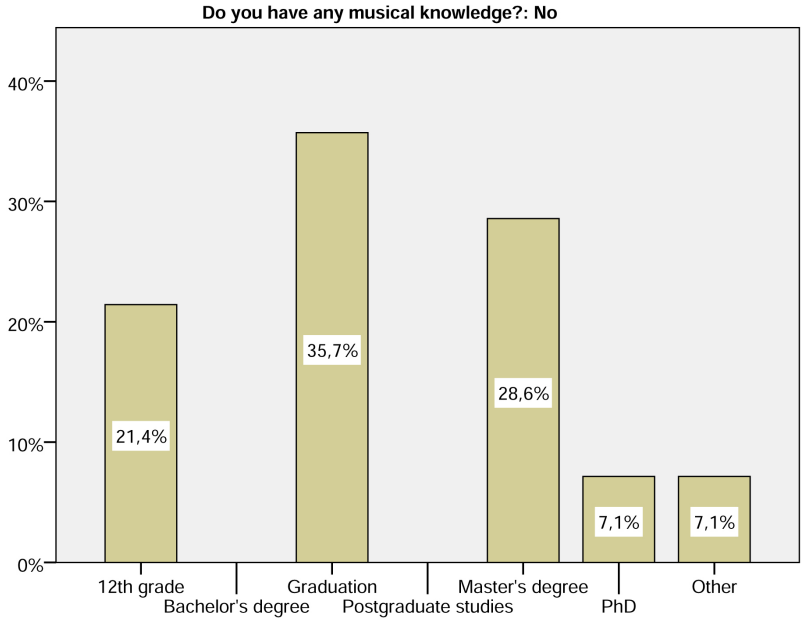


Figure 4.18: Educational qualifications (group without any musical knowledge).

all participants, so that it is reasonable to consider that the remaining data of this group of participants are reliable for analysis.

4.2.1.2 Group of Participants with Musical Knowledge

Figure 4.20 shows that 20.7% of the volunteers of the group of 29 participants with musical knowledge were female and 79.3% were male. These percentages are very similar to those in the group without any musical

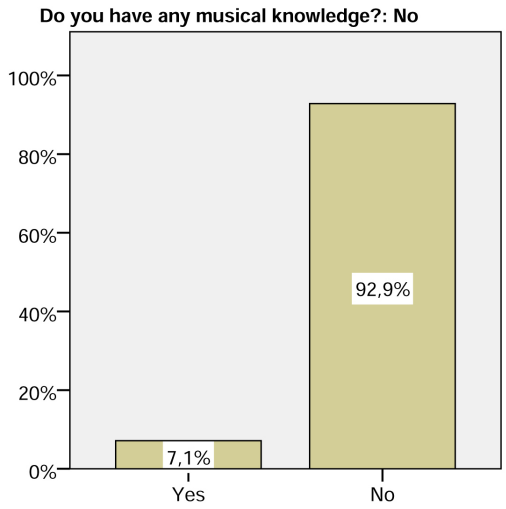


Figure 4.19: Do you have hearing handicaps? (group without any musical knowledge)

knowledge (see section 4.2.1.1) and also to those in which all participants were considered (cf. section 4.2). In this sense, the mode of 2 (see appendix E.12.5), where 2 represents the male gender, confirms that the majority of participants with musical knowledge was of the male gender.

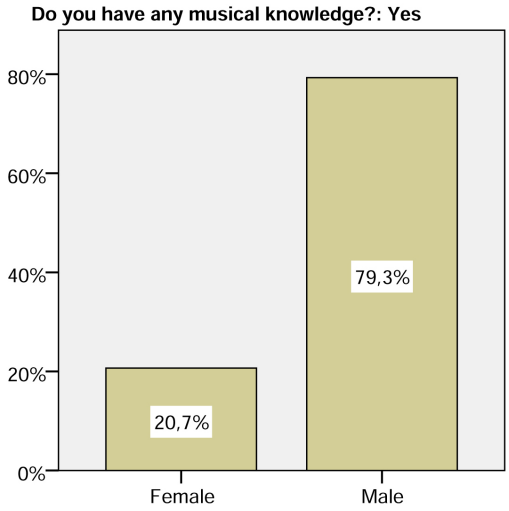


Figure 4.20: Gender (group with musical knowledge).

Regarding age, 17.2% were aged between 15 and 24, 37.9% were aged between 25 and 34, 17.2% were aged between 35 and 44, 24.1% were aged between 45 and 54, and 3.4% were aged between 55 and 64 (see figure 4.21). In this case, the median of 2 and the mode of 2 (cf. appendix E.12.5) are meaningful, confirming that the 25 to 34 years category is both the central and the most frequent one of the dataset. These results are similar to those presented in section 4.2, where all participants were taken into account, and also in section 4.2.1.1, where the group of participants without any musical knowledge was considered.

With regard to the age by gender and relating it to the female gender, 16.7% of the volunteers were aged between 15 and 24, 16.7% were aged between 25 and 34, 16.7% were aged between 35 and 44, and 50.0% were aged between 45 and 54 (see figure 4.22). Whereas the median of 3.50 in table 4.2 demonstrates that in this case the central value of this dataset, consisting of an even amount of ordered scores, is between the

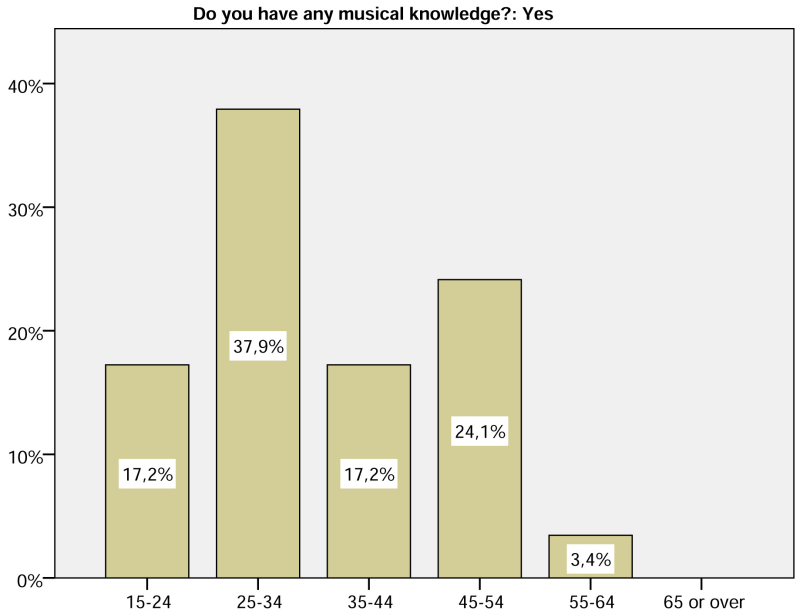


Figure 4.21: Age (group with musical knowledge).

35 to 44 years' and the 45 to 54 years' categories, the mode of 4 strengthens the idea that the 45 to 54 years' category for the female gender occurs most frequently.

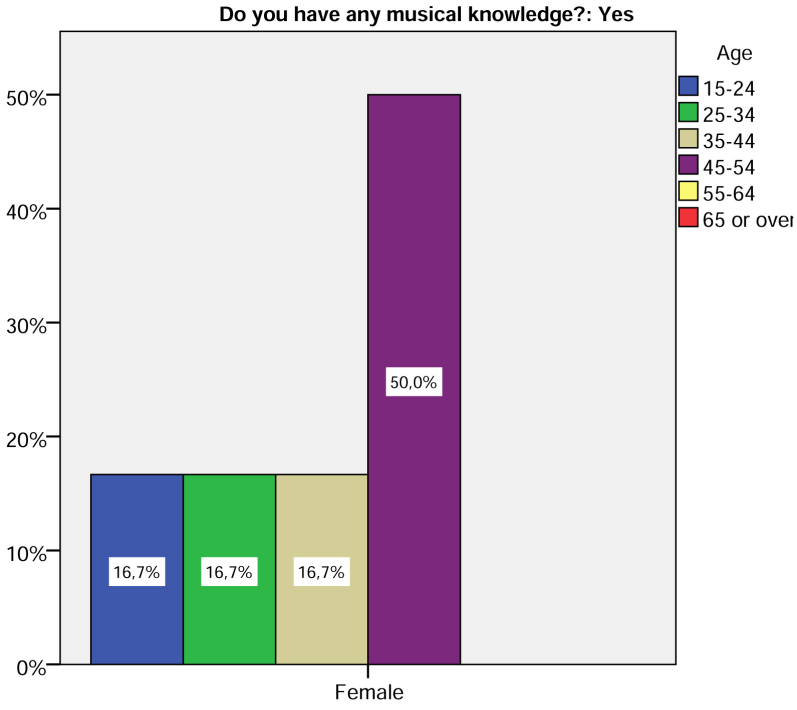


Figure 4.22: Age by gender: female (group with musical knowledge).

Concerning the male gender, 17.4% were aged between 15 and 24, 43.5% were aged between 25 and 34, 17.4% were aged between 35 and 44, 17.4% were aged between 45 and 54, and 4.3% were aged between 55 and 64 (see figure 4.23). Both the median and mode of 2 in table 4.2 corroborate respectively that the 25 to 34 years' category is the central range of this dataset and that it occurs most frequently for the male gender,

just like in sections 4.2 and 4.2.1.1, where all participants and the group of participants without any musical knowledge were respectively considered.

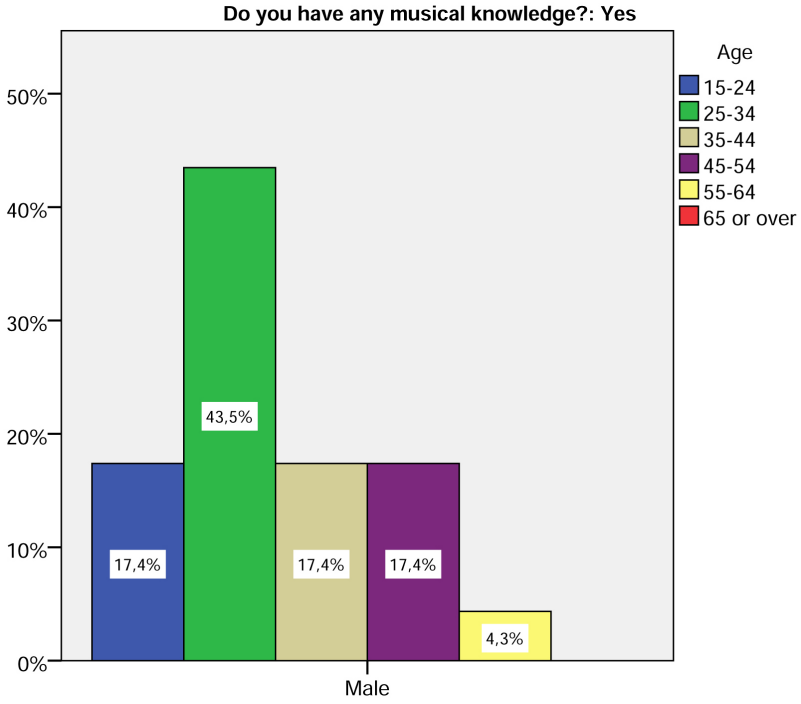


Figure 4.23: Age by gender: male (group with musical knowledge).

The decrease of the auditory sensibility with age, mentioned in section 2.1.3.2, is by itself not a problem to consider in the data analysis of this group of participants, because most women were under 55 years old and the majority of men ($17.4\% + 43.5\% = 60.9\%$) was under 35 years old. This result is similar to that obtained in section 4.2 when considering all participants.

With reference to the question: “Which hand do you use for writing?”, 5 volunteers did not reply. As a result and taking only the valid answers into account, 16.7% answered that they used the left hand and 83.3% answered that they used the right hand for writing (see figure 4.24). These outcomes are also very similar to those in section 4.2, when considering all participants, and to those of the group of participants without any musical knowledge in section 4.2.1.1. They are confirmed by the mode of 2 (cf. appendix E.12.5), where 2 represents the right hand.

Figure 4.25 shows that 7.1% said that they had A levels (12th grade), 50.0% said that they had a graduation, 7.1% said that they had a postgraduate qualification, 21.4% said that they had a master’s degree, and 14.3% said that they had a PhD. The mode of 3 (see appendix E.12.5), representing the group of participants with a graduation, confirms that this group was the most frequent among the participants with musical knowledge.

The last question of part 1 was: “Do you have hearing handicaps?”. 6.9% of the participants answered yes and 93.1% replied no (see figure 4.26). The types of hearing difficulties pointed out by the participants were partial loss of hearing (*hypoacusis*) in the right ear and difficulties in the identification of sounds. These results are also very similar to those obtained in section 4.2, when considering all participants, and to those of the group of participants without any musical knowledge in section 4.2.1.1, so that it is reasonable to consider that the remaining data of this group of participants with musical knowledge are reliable for analysis as well. The

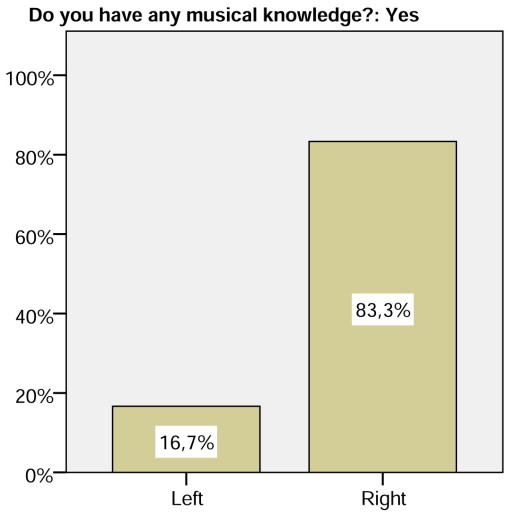


Figure 4.24: Which hand do you use for writing? (group with musical knowledge)

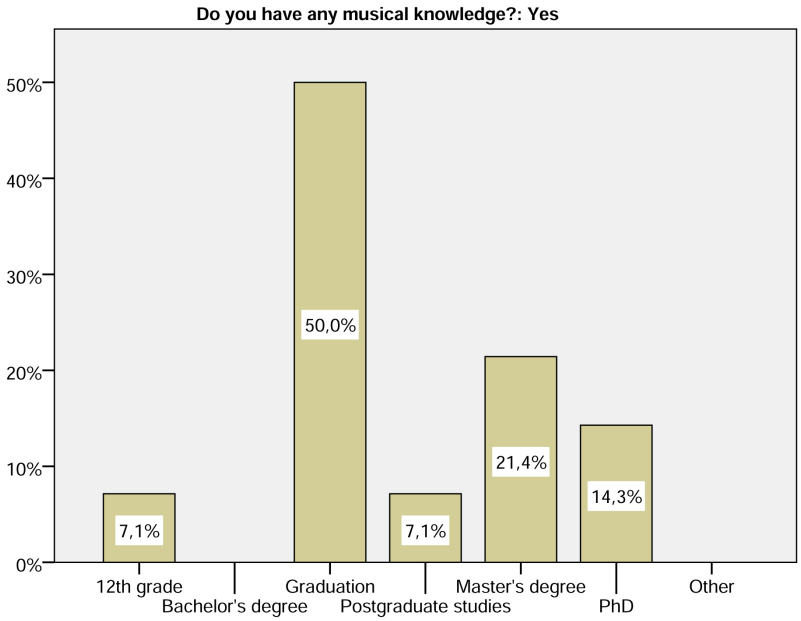


Figure 4.25: Educational qualifications (group with musical knowledge).

mode of 2, where 2 represents a negative answer, corroborates this statement.

4.3 Experiment Data Results - All Participants

As already mentioned in section 4.1, for each experimental condition item or question in part 2 of the Inquiry Mode Questionnaire (InQ) (see appendix A and figures 4.1, 4.2, and 4.3) we firstly checked four particular assumptions before deciding which statistical tests would be more appropriate for each hypothesis' test related to the experiments, since we followed the statistical procedures based on the normal distribution of data, in which data are considered as being symmetrically distributed around the centre of all scores (Field, 2009, p. 18).

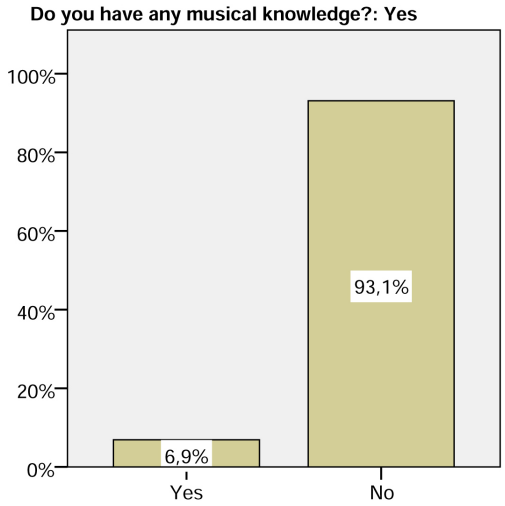


Figure 4.26: Do you have hearing handicaps? (group with musical knowledge)

Thus, in order to test whether any of the distributions of scores of each item would be normal, we calculated their descriptive statistics to give us an insight into the central tendency and dispersion of data, created and reviewed their histograms to give us an idea of the general shape of the frequency distributions and the possibility of outlier scores, and carried out the Kolmogorov-Smirnov (with Lilliefors significance correction in order to detect deviations from the normal distribution in a more powerful way than the standard Kolmogorov-Smirnov test) and Shapiro-Wilk normality tests, the latter having more power to detect differences from normality (Field, 2009, p. 148) (see Appendix E.12.6). The explanation of how these tests are calculated is naturally beyond the scope of our study and can be found in Field (2009) and Pestana & Gageiro (2014).

We therefore concluded that there was a significant test with a significance value p less than .05 for almost every item, meaning that almost every distribution was significantly different from a normal distribution, that is, they were non-normal, which violated the normality assumption (Field, 2009, p. 144):

1. In experiment 1, the scores of item 1, $D(43) = .48, p < .001$ and $W(43) = .53, p < .001$, item 2, $D(43) = .29, p < .001$ and $W(43) = .62, p < .001$, item 3, $D(43) = .28, p < .001$ and $W(43) = .82, p < .001$, item 4, $D(43) = .32, p < .001$ and $W(43) = .74, p < .001$, item 5, $D(43) = .33, p < .001$ and $W(43) = .82, p < .001$, item 6, $D(43) = .31, p < .001$ and $W(43) = .83, p < .001$, item 7, $D(43) = .30, p < .001$ and $W(43) = .84, p < .001$, item 8, $D(43) = .37, p < .001$ and $W(43) = .63, p < .001$, item 9, $D(43) = .29, p < .001$ and $W(43) = .79, p < .001$, item 10, $D(42) = .14, p < .05$ and $W(42) = .92, p < .05$, and item 12, $D(40) = .14, p > .05$ but $W(40) = .94, p < .05$, were significantly non-normal;

2. In experiment 2, the scores of item 1, $D(43) = .45, p < .001$ and $W(43) = .58, p < .001$, item 2, $D(43) = .30, p < .001$ and $W(43) = .78, p < .001$, item 3, $D(42) = .33, p < .001$ and $W(42) = .72, p < .001$, item 4, $D(43) = .36, p < .001$ and $W(43) = .76, p < .001$, item 5, $D(42) = .29, p < .001$ and $W(42) = .76, p < .001$, item 6, $D(43) = .29, p < .001$ and $W(43) = .84, p < .001$, item 7, $D(43) = .29, p < .001$ and $W(43) = .80, p < .001$, item 8, $D(43) = .34, p < .001$ and $W(43) = .72, p < .001$, item 9, $D(42) = .34, p < .001$ and $W(42) = .69, p < .001$, item 10, $D(43) = .28, p < .001$ and $W(43) = .83, p < .001$, item 11, $D(41) = .14, p < .05$ and $W(41) = .92, p < .05$, item 12, $D(40) = .13, p > .05$ but $W(40) = .94, p < .05$,

and item 13, $D(40) = .15, p < .05$ and $W(40) = .93, p < .05$, were also significantly non-normal;

3. In experiment 3, the scores of item 1, $D(43) = .45, p < .001$ and $W(43) = .58, p < .001$, item 2, $D(42) = .23, p < .001$ and $W(42) = .83, p < .001$, item 3, $D(42) = .20, p < .001$ and $W(42) = .91, p < .05$, item 4, $D(43) = .29, p < .001$ and $W(43) = .75, p < .001$, item 5, $D(43) = .29, p < .001$ and $W(43) = .87, p < .001$, item 6, $D(42) = .29, p < .001$ and $W(42) = .85, p < .001$, item 7, $D(43) = .30, p < .001$ and $W(43) = .84, p < .001$, item 8, $D(42) = .37, p < .001$ and $W(42) = .57, p < .001$, item 9, $D(43) = .35, p < .001$ and $W(43) = .80, p < .001$, item 10, $D(43) = .20, p < .001$ and $W(43) = .90, p < .001$, item 11, $D(42) = .12, p > .05$ but $W(42) = .94, p < .05$, and item 12, $D(38) = .18, p < .05$ and $W(38) = .93, p < .05$, were significantly non-normal as well.

The respective descriptive statistics of mean, median, range, interquartile range, standard deviation, and essentially of skewness and kurtosis, and the respective histograms supported the idea that the assumption of normality was violated for all these items. For instance, the values of skewness and kurtosis should be both very close to, or equal to, zero, as is expected in a normal distribution (Field, 2009, p. 19) (skewness of approximately $\pm.2$ and at the same time kurtosis around $\pm.7$).

The only exceptions were those scores of item 11 in experiment 1, $D(40) = .11, p > .05$ and $W(40) = .95, p > .05$, and of item 13 in experiment 3, $D(39) = .14, p > .05$ and $W(39) = .97, p > .05$, which were approximately normally distributed.

Since we developed a within-subjects or repeated-measures design, where the same participants were used in all three experiments (cf. section 3.6), the assumption of homogeneity of variance, that is, the assumption that variances should be the same throughout the data taken from several groups of participants, where scores in different experimental conditions should be independent, was violated (Field, 2009, p. 459). In our case, "... scores taken under different experimental conditions are likely to be related because they come from the same participants" (Field, 2009, p. 459).

The assumption that data should be measured at least at the interval level (Field, 2009, p. 133) was considered to be non-valid, since the items using five response alternatives numbered from 1 to 5 (cf. section 4.1) were ordinal. The remaining items, associated with time estimates, were also ordinal.

The assumption of independence was also violated, because scores are expected to be non-independent for a given participant in repeated-measures design. Nevertheless, "... behaviour between different participants should be independent" (Field, 2009, p. 133).

4.3.1 Experiments 1, 2, and 3 Data Results

Since assumptions of normality (for almost all items), homogeneity of variance, interval data, and independence were violated (cf. section 4.3), we performed non-parametric Friedman's ANOVA tests on hypotheses H1 to H9, presented in section 3.6.4.1, in which differences between the three experimental conditions, where the same participants had been used, were examined (Field, 2009, pp. 573, 581) (see appendix E.12.7). Thus, we had one categorical independent variable (*experimental condition*) with three levels, because it was manipulated in three ways, and one ordinal dependent variable (*answer on a 5-point scale or time estimates*).

Table 4.4: Friedman's ANOVA test on hypothesis H1.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Experiment 1 - Quickly understood the experiment	43	4,77	,480	3	5	5,00	5,00	5,00
Experiment 2 - Quickly understood the experiment	43	4,72	,504	3	5	4,00	5,00	5,00
Experiment 3 - Quickly understood the experiment	43	4,70	,558	3	5	4,00	5,00	5,00

Ranks

	Mean Rank
Experiment 1 - Quickly understood the experiment	2,07
Experiment 2 - Quickly understood the experiment	1,98
Experiment 3 - Quickly understood the experiment	1,95

Test Statistics[a]

N	43
Chi-Square	1,514
df	2
Asymp. Sig.	,469
Exact Sig.	,501
Point Probability	,052

a. Friedman Test

4.3.1.1 Hypothesis Test H1

The Friedman's ANOVA test is based on ranks and not the actual scores (Field, 2009, p. 576). Thus, when applied to hypothesis H1 that "there are no significant differences in how quickly participants understand the three experiments, but that they rate their quick understanding highly", this test showed that (see table 4.4): 1) the median levels for the three experimental conditions were of 5 (5 to 5), 5 (4 to 5), and 5 (4 to 5), respectively; and 2) the significance value was equal to .47, or .50 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in how quickly participants would understand the three experiments, $\chi^2(2) = 1.51, p > .05$. The descriptive statistic median of 5 in all experiments, standing for "I totally agree", supports the idea that participants firmly estimated that they had understood the three experiments quickly (cf. appendix E.12.3). No *post hoc* procedures, that is, pairwise comparisons in order to "... compare all different combinations of the treatment groups" (Field, 2009, p. 372), were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.3.1.2 Hypothesis Test H2

The Friedman's ANOVA test on the hypothesis H2 that "there are no significant differences in the estimation of time it takes for participants to understand the three experimental conditions, but that they take less than 2 seconds to understand them" showed that (see table 4.5): 1) the median levels for the three experimental conditions were of 22 (17 to 28), 21 (14 to 26.5), and 17 (13 to 27), respectively; and 2) the significance or the exact significance values were both equal to .002, which was well under .05. Therefore, we can conclude that there was a statistically significant difference in the estimation of time it took for participants to understand the three experiments, $\chi^2(2) = 12.36, p < .05$.

Nevertheless, because the Friedman's ANOVA test was significant (Field, 2009, p. 577) and because it does

Table 4.5: Friedman's ANOVA test on hypothesis H2.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Experiment 1 - Estimated time needed to understand the experiment	41	20,93	8,302	1	32	17,00	22,00	28,00
Experiment 2 - Estimated time needed to understand the experiment	41	18,51	8,692	1	31	14,00	21,00	26,50
Experiment 3 - Estimated time needed to understand the experiment	41	18,27	9,000	1	32	13,00	17,00	27,00

Ranks

	Mean Rank
Experiment 1 - Estimated time needed to understand the experiment	2,34
Experiment 2 - Estimated time needed to understand the experiment	1,90
Experiment 3 - Estimated time needed to understand the experiment	1,76

Test Statistics[a]

N	41
Chi-Square	12,356
df	2
Asymp. Sig.	,002
Exact Sig.	,002
Point Probability	,000

a. Friedman Test

not pinpoint which experiments in particular differ from each other, *post hoc* procedures were performed. Thus, we used three separate Wilcoxon signed-rank tests based on the different combinations of our experimental conditions, in which experiment 1 was compared with experiment 2, experiment 1 with experiment 3, and experiment 2 with experiment 3 (see appendix E.12.7). In order to ensure that Type I errors, which occur "... when we believe that there is a genuine effect in our population, when in fact there isn't" (Field, 2009, p. 56), would not build up to more than .05 because we were making multiple comparisons, we used a Bonferroni correction. This means that we used a critical value for significance of .05 divided by the number of tests we carried out, that is, .05 / 3 = .017, instead of .05, in each test (Field, 2009, p. 577).

The results showed that there was a statistically significant reduction of the estimated time in experiment 2 (median of 21, standing for "20 s") relatively to experiment 1 (median of 22, standing for "30 s") ($Z = -2.51, p = .012 < .017$), $T = 77, p < .017, r = -.28$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.51}{\sqrt{41 \times 2}} = -.28$, which means that there was a medium effect or a moderate difference of estimation of time between experiments 1 and 2 (Field, 2009, pp. 57, 170) (see figure 4.27).

In addition, there was also a statistically significant reduction of the estimated time in experiment 3 (median of 17, standing for "10 s") relatively to experiment 1 (median of 22, standing for "30 s") ($Z = -2.93, p = .003 < .017$), $T = 74.5, p < .017, r = -.32$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.93}{\sqrt{41 \times 2}} = -.32$, which means that there was a medium effect or a moderate difference of estimation of time between experiments 1 and 3 as well.

Finally, there were no significant differences between experiment 2 and experiment 3, although there is an overall reduction of the estimated time in experiment 3 (median of 17, standing for "10 s") relatively to

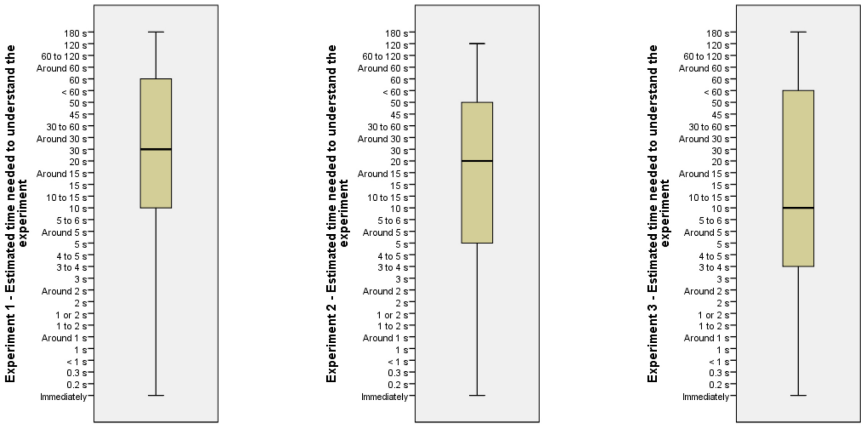


Figure 4.27: Boxplots of estimated times needed to understand the experiments (hypothesis tests H2).

experiment 2 (median of 21, standing for "20 s") ($Z = -.65, p = .51 > .017$), $T = 70.5, p > .017, r = -.07$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-.65}{\sqrt{41 \times 2}} = -.07$.

4.3.1.3 Hypothesis Test H3

The Friedman's ANOVA test on the hypothesis H3 that "there are no significant differences in how easily participants interact with the installation in the three experiments, but that they rate their ease of interaction highly" showed that (see table 4.6): 1) the median levels for the three experimental conditions were of 4 (4 to 5), 4.5 (4 to 5), and 4 (3.75 to 5), respectively; and 2) the significance or the exact significance values were both equal to .28, which was well above .05.

Therefore, we can conclude that there were no significant differences in how easily participants interacted with the installation in the three experiments, $\chi^2(2) = 2.52, p > .05$. The descriptive statistic median of 4 in all experiments, standing for "I agree", supports this idea and confirms that participants rated their ease of interaction highly (cf. appendix E.12.3). No *post hoc* procedures were performed (see section 4.3.1.1), because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.3.1.4 Hypothesis Test H4

The Friedman's ANOVA test on the hypothesis H4 that "participants define the origin of sound more easily in the first and then in the third experimental condition" showed that (see table 4.7): 1) the median levels for the three experimental conditions were of 4 (4 to 5), 4 (4 to 4.25), and 3 (3 to 4), respectively; and 2) the significance or the exact significance values were both equal to .001, which was well under .05. Therefore, we can conclude that participants admitted that they easily defined the origin of sound in significantly different ways, $\chi^2(2) = 13.79, p < .05$.

Post hoc procedures were performed because the Friedman's ANOVA test was significant (Field, 2009, p. 577). Thus, we used three Wilcoxon signed-rank tests, in which experiment 1 was compared with experiment 2, experiment 1 with experiment 3, and experiment 2 with experiment 3. We applied a Bonferroni correction, which means that we used a critical value for significance of .017, instead of .05, in each test (cf. section 4.3.1.1).

Table 4.6: Friedman's ANOVA test on hypothesis H3.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
Experiment 1 - It was easy to interact with the installation	42	4,33	,721	1	5	4,00	4,00	5,00
Experiment 2 - It was easy to interact with the installation	42	4,38	,697	3	5	4,00	4,50	5,00
Experiment 3 - It was easy to interact with the installation	42	4,10	,878	2	5	3,75	4,00	5,00

Ranks

	Mean Rank
Experiment 1 - It was easy to interact with the installation	2,05
Experiment 2 - It was easy to interact with the installation	2,10
Experiment 3 - It was easy to interact with the installation	1,86

Test Statistics[a]

N	42
Chi-Square	2,517
df	2
Asymp. Sig.	,284
Exact Sig.	,283
Point Probability	,008

a. Friedman Test

Table 4.7: Friedman's ANOVA test on hypothesis H4.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
Experiment 1 - It was easy to define the origin of sound	42	4,07	,808	2	5	4,00	4,00	5,00
Experiment 2 - It was easy to define the origin of sound	42	3,95	,909	1	5	4,00	4,00	4,25
Experiment 3 - It was easy to define the origin of sound	42	3,40	1,014	1	5	3,00	3,00	4,00

Ranks

	Mean Rank
Experiment 1 - It was easy to define the origin of sound	2,21
Experiment 2 - It was easy to define the origin of sound	2,17
Experiment 3 - It was easy to define the origin of sound	1,62

Test Statistics[a]

N	42
Chi-Square	13,786
df	2
Asymp. Sig.	,001
Exact Sig.	,001
Point Probability	,000

a. Friedman Test

The outcomes show that there were no significant differences between experiments 1 and 2 ($Z = -.69, p = .49 > .017$), $T = 87, p > .017, r = -.08$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-.69}{\sqrt{42 \times 2}} = -.08$, although it was a little bit easier to define the origin of sound in experiment 1 than in experiment 2 (the higher sum of ranks associated to the negative ranks indicates that). Nevertheless, there were statistically significant differences between experiments 1 and 3 ($Z = -3.59, p = .000 < .017$), $T = 52.5, p < .017, r = -.39$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-3.59}{\sqrt{42 \times 2}} = -.39$ means that there was a medium to large effect or a moderate to big difference in easily defining the origin of sound between experiments 1 and 3 (Field, 2009, pp. 57, 170). Thus, it was easier to define the origin of sound in experiment 1 than in experiment 3. There were also significant differences between experiments 2 and 3 ($Z = -2.52, p = .012 < .017$), $T = 95, p < .017, r = -.28$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.52}{\sqrt{42 \times 2}} = -.28$, which means that there was a medium effect or a moderate difference between experiments 2 and 3. Thus, according to the higher sum of ranks associated to the negative ranks, it was easier to define the origin of sound in experiment 2 than in experiment 3. The descriptive statistic median of 4 in experiments 1 and 2, standing for "I agree", and the median of 3, standing for "Not always / Sometimes", in experiment 3 (cf. appendix E.12.3), support these outcomes.

4.3.1.5 Hypothesis Test H5

The Friedman's ANOVA test on the hypothesis H5 that "there are no significant differences in how highly participants rate the suggested gesture as being adequate to any of the three experiments, but that they rate its adequacy highly" showed that (see table 4.8): 1) the median levels for the three experimental conditions were of 4.5 (4 to 5), 4 (4 to 5), and 4 (4 to 5), respectively; and 2) the significance value was equal to .54, or .58 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no significant differences in how highly participants rated the suggested gesture as being adequate to any of the three experiments, $\chi^2(2) = 1.22, p > .05$. The descriptive statistic median of 5 in experiment 1, standing for "I totally agree", and the median of 4 in experiments 2 and 3, standing for "I agree", support this idea and confirm that participants rated the gesture's adequacy highly (cf. appendix E.12.3). No *post hoc* procedures were performed (cf. section 4.3.1.1), because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.3.1.6 Hypothesis Test H6

The Friedman's ANOVA test on the hypothesis H6 that "there are no significant differences in the degree of appreciation whether the system's response to gesture is immediate in any of the three experiments, but that participants more likely consider that the system's response to gesture is immediate" showed that (see table 4.9): 1) the median levels for the three experimental conditions were of 4 (3 to 4), 4 (4 to 5), and 4 (3 to 4), respectively; and 2) the significance or the exact significance values were both equal to .02, which was under .05. Therefore, we can conclude that there were statistically significant differences in the degree of appreciation whether the system's response to gesture was immediate, $\chi^2(2) = 7.72, p < .05$.

Post hoc procedures were performed because the Friedman's ANOVA test was significant (Field, 2009, p. 577). Thus, we used three separate Wilcoxon signed-rank tests, in which experiment 1 was compared with

Table 4.8: Friedman's ANOVA test on hypothesis H5.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Experiment 1 - The suggested gesture is adequate to the experiment	42	4,43	,630	3	5	4,00	4,50	5,00
Experiment 2 - The suggested gesture is adequate to the experiment	42	4,36	,727	2	5	4,00	4,00	5,00
Experiment 3 - The suggested gesture is adequate to the experiment	42	4,36	,618	3	5	4,00	4,00	5,00

Ranks

	Mean Rank
Experiment 1 - The suggested gesture is adequate to the experiment	2,07
Experiment 2 - The suggested gesture is adequate to the experiment	1,98
Experiment 3 - The suggested gesture is adequate to the experiment	1,95

Test Statistics[a]

N	42
Chi-Square	1,217
df	2
Asymp. Sig.	,544
Exact Sig.	,576
Point Probability	,051

a. Friedman Test

Table 4.9: Friedman's ANOVA test on hypothesis H6.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Experiment 1 - The response of the system to the gesture was immediate	42	3,83	,853	2	5	3,00	4,00	4,00
Experiment 2 - The response of the system to the gesture was immediate	42	4,10	,692	2	5	4,00	4,00	5,00
Experiment 3 - The response of the system to the gesture was immediate	42	3,76	,906	2	5	3,00	4,00	4,00

Ranks

	Mean Rank
Experiment 1 - The response of the system to the gesture was immediate	1,92
Experiment 2 - The response of the system to the gesture was immediate	2,21
Experiment 3 - The response of the system to the gesture was immediate	1,87

Test Statistics[a]

N	42
Chi-Square	7,719
df	2
Asymp. Sig.	,021
Exact Sig.	,020
Point Probability	,002

a. Friedman Test

experiment 2, experiment 1 with experiment 3, and experiment 2 with experiment 3. We applied a Bonferroni correction, which means that we used a critical value for significance of .017, instead of .05, in each test (see section 4.3.1.1).

The outcomes show that there were no significant differences between experiments 1 and 3 ($Z = -.73, p = .47 > .017$), $T = 42, p > .017, r = -.08$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-.73}{\sqrt{42 \times 2}} = -.08$, although the higher sum of ranks associated to the negative ranks indicates that the rating tendency was slightly higher in experiment 1 than in experiment 3. There were also no significant differences between experiments 1 and 2 ($Z = -2.30, p = .022 > .017$), $T = 32, p > .017, r = -.25$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.30}{\sqrt{42 \times 2}} = -.25$, but the higher sum of ranks associated to the positive ranks suggests that the rating tendency was higher in experiment 2 than in experiment 1. Nevertheless, there were statistically significant differences between experiments 2 and 3 ($Z = -2.64, p = .008 < .017$), $T = 9, p < .017, r = -.29$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.64}{\sqrt{42 \times 2}} = -.29$, which means that there was a medium effect or a moderate difference in the degree of appreciation whether the system's response to gesture was immediate (Field, 2009, pp. 57, 170). The higher sum of ranks associated to the negative ranks indicates that the rating tendency was a little bit higher in experiment 2 than in experiment 3. Finally, the descriptive statistic median of 4 in all experiments, standing for "I agree" (see appendix E.12.3 as well), confirms that participants more likely considered that the system's response to gesture was immediate.

In order to determine inconsistencies in responses, questions 6 and 7 relating to experiments 1 and 3, and questions 7 and 8 concerning experiment 2, had purposely inverted senses, as already mentioned in section 4.1. Thus, we firstly inverted the last question of each of these pairs, so that we could compare their outcomes with those of the first question, transforming it into "The system's response to the gesture was fast." Then we conducted a Friedman's ANOVA test on them, which showed that (see table 4.10): 1) the median levels for the three experimental conditions were of 4 (3 to 4), 4 (4 to 5), and 4 (3 to 4), respectively; and 2) the significance or the exact significance values were both equal to .08, which was above .05.

Therefore, we can conclude that there were no significant differences in the degree of appreciation whether the system's response to gesture was fast in any of the three experiments, $\chi^2(2) = 5.182, p > .05$. The descriptive statistic median of 4 in the three experiments, standing for "I agree", supports this idea and confirms that participants more likely considered that the system's response to gesture was fast, just as in the first question of each of the inverted-sensed pair of questions (cf. appendix E.12.3). No *post hoc* procedures were performed (cf. section 4.3.1.1), because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.3.1.7 Hypothesis Test H7

The Friedman's ANOVA test on the hypothesis H7 that "there are no significant differences in the estimation of the shorter times of the system's response to gesture in the three experiments" showed that (see table 4.11): 1) the median levels for the three experimental conditions were of 13.5 (9 to 21.5), 12 (9 to 19.25), and 13 (9.75 to 24), respectively; and 2) the significance value was equal to .41, or .42 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no significant differences in the estimation of the shorter times of the system's response to gesture in the three experiments, $\chi^2(2) = 1.76, p > .05$. The descriptive statistic median of 12.5 in experiment 1, standing for between "1 s" and "Around 1 s", the median of 12 in experiment

Table 4.10: Friedman's ANOVA test on inverted questions related to hypothesis H6.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Experiment 1 - The response of the system to the gesture was fast	43	3,60	1,198	1	5	3,00	4,00	4,00
Experiment 2 - The response of the system to the gesture was fast	43	4,05	,899	1	5	4,00	4,00	5,00
Experiment 3 - The response of the system to the gesture was fast	43	3,79	1,036	1	5	3,00	4,00	4,00

Ranks

	Mean Rank
Experiment 1 - The response of the system to the gesture was fast	1,90
Experiment 2 - The response of the system to the gesture was fast	2,17
Experiment 3 - The response of the system to the gesture was fast	1,93

Test Statistics[a]

N	43
Chi-Square	5,182
df	2
Asymp. Sig.	,075
Exact Sig.	,075
Point Probability	,003

a. Friedman Test

Table 4.11: Friedman's ANOVA test on hypothesis H7.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Experiment 1 - Estimated time of the system's response to gesture	38	15,37	9,281	1	34	9,00	13,50	21,50
Experiment 2 - Estimated time of the system's response to gesture	38	14,39	9,084	1	34	9,00	12,00	19,25
Experiment 3 - Estimated time of the system's response to gesture	38	16,24	9,172	1	34	9,75	13,00	24,00

Ranks

	Mean Rank
Experiment 1 - Estimated time of the system's response to gesture	2,09
Experiment 2 - Estimated time of the system's response to gesture	1,88
Experiment 3 - Estimated time of the system's response to gesture	2,03

Test Statistics[a]

N	38
Chi-Square	1,763
df	2
Asymp. Sig.	,414
Exact Sig.	,422
Point Probability	,018

a. Friedman Test

Table 4.12: Friedman's ANOVA test on hypothesis H8.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Experiment 1 - Felt surrounded by sound in the installation	41	4,54	,809	1	5	4,00	5,00	5,00
Experiment 2 - Felt surrounded by sound in the installation	41	4,39	,972	1	5	4,00	5,00	5,00
Experiment 3 - Felt surrounded by sound in the installation	41	4,54	,840	1	5	4,00	5,00	5,00

Ranks

	Mean Rank
Experiment 1 - Felt surrounded by sound in the installation	2,04
Experiment 2 - Felt surrounded by sound in the installation	1,90
Experiment 3 - Felt surrounded by sound in the installation	2,06

Test Statistics[a]

N	41
Chi-Square	2,178
df	2
Asymp. Sig.	,337
Exact Sig.	,368
Point Probability	,045

a. Friedman Test

2, standing for "1 s", and the median of 13 in experiment 3, standing for "Around 1 s", support this idea and confirm that participants estimated a short time of the system's response to gesture (cf. appendix E.12.3).

4.3.1.8 Hypothesis Test H8

The Friedman's ANOVA test on the hypothesis H8 that "there are no significant differences in how strongly participants feel surrounded by sound in the installation, but that they rate this feeling highly" showed that (see table 4.12): 1) the median levels were of 5 (4 to 5) in the three experimental conditions; and 2) the significance value was equal to .34, or .37 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no significant differences in how strongly participants felt surrounded by sound in the installation, $\chi^2(2) = 2.18, p > .05$. The descriptive statistic median of 5 in all experiments supports this result and confirms that participants more firmly consider that they felt surrounded by sound in the installation (cf. appendix E.12.3).

4.3.1.9 Hypothesis Test H9

The Friedman's ANOVA test on the hypothesis H9 that "participants more firmly consider that their gesture coincides with the origin of sound in the first and then in the third experimental conditions, and that they do not coincide at all in the second" showed that (see table 4.13): 1) the median levels for the three experimental conditions were of 4 (4 to 5), 4 (2 to 4), and 4 (3 to 4), respectively; and 2) the significance value was equal

Table 4.13: Friedman's ANOVA test on hypothesis H9.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Experiment 1 - The gesture coincided with the origin of sound	43	4,19	,664	3	5	4,00	4,00	5,00
Experiment 2 - The gesture coincided with the origin of sound	43	3,16	1,413	1	5	2,00	4,00	4,00
Experiment 3 - The gesture coincided with the origin of sound	43	3,51	,910	1	5	3,00	4,00	4,00

Ranks

	Mean Rank
Experiment 1 - The gesture coincided with the origin of sound	2,35
Experiment 2 - The gesture coincided with the origin of sound	1,78
Experiment 3 - The gesture coincided with the origin of sound	1,87

Test Statistics[a]

N	43
Chi-Square	11,613
df	2
Asymp. Sig.	,003
Exact Sig.	,002
Point Probability	,000

a. Friedman Test

to .003, or .002 for the exact significance, which was well under .05. Therefore, we can conclude that there were statistically significant differences in how participants rated that their gesture coincided with the origin of sound, $\chi^2(2) = 11.61, p < .05$.

Post hoc procedures were performed because the Friedman's ANOVA test was significant (Field, 2009, p. 577). Thus, we used three separate Wilcoxon signed-rank tests, in which experiment 1 was compared with experiment 2, experiment 1 with experiment 3, and experiment 2 with experiment 3. We applied a Bonferroni correction, which means that we used a critical value for significance of .017, instead of .05, in each test (cf. section 4.3.1.1).

The outcomes show that there was a statistically significant reduction tendency (as can be read from the highest sum of ranks associated to the negative ranks) in considering that the gesture coincided with the origin of sound in experiment 2 relatively to experiment 1 ($Z = -3.36, p = .001 < .017$), $T = 72, p < .017, r = -.36$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-3.36}{\sqrt{43 \times 2}} = -.36$, which means that there was a medium effect or a moderate difference in this rating (Field, 2009, pp. 57, 170). There was also a statistically significant reduction tendency in considering that the gesture coincided with the origin of sound in experiment 3 relatively to experiment 1 ($Z = -3.61, p = .000 < .017$), $T = 40, p < .017, r = -.39$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-3.61}{\sqrt{43 \times 2}} = -.39$, which corresponds to a medium to large effect or a moderate to big difference in this rating.

Nevertheless, there was no significant difference in assuming that the gesture coincided with the origin of sound in experiments 2 and 3 ($Z = -1.35, p = .18 > .017$), $T = 156, p > .017, r = -.15$, where $r = \frac{Z}{\sqrt{N \times 2}} =$

Table 4.14: Kendall's and Spearman's correlation coefficient tests on hypothesis H10.

Correlations[a]

			Experiment 1 - Quickly understood the experiment	Experiment 1 - Estimated time needed to understand the experiment
Kendall's tau_b	Experiment 1 - Quickly understood the experiment	Correlation Coefficient	1,000	-,122
		Sig. (1-tailed)	.	,175
	Experiment 1 - Estimated time needed to understand the experiment	Correlation Coefficient	-,122	1,000
		Sig. (1-tailed)	,175	.
Spearman's rho	Experiment 1 - Quickly understood the experiment	Correlation Coefficient	1,000	-,147
		Sig. (1-tailed)	.	,177
	Experiment 1 - Estimated time needed to understand the experiment	Correlation Coefficient	-,147	1,000
		Sig. (1-tailed)	,177	.

a. Listwise N = 42

$\frac{-1.35}{\sqrt{43 \times 2}} = -.15$, although a small effect or difference could be still observed: the rating tendency was slightly lower in experiment 2 than in 3. Finally, the descriptive statistic median of 4 in all experiments, standing for "I agree" (cf. appendix E.12.3 as well), confirm that participants more likely assumed that the gesture coincided with the origin of sound.

4.3.2 Experiment 1 Data Results

Parametric assumptions, such as the assumption of normality, were violated, as already mentioned in section 4.3. Therefore, we conducted non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H10 and H11, presented in section 3.6.4.2, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.8).

4.3.2.1 Hypothesis Test H10

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H10 that "the estimated time needed to understand Experiment 1 is less than 2 seconds if participants consider that they quickly understand it" showed that there was no significant relationship between considering that participants quickly understood experiment 1 and the estimated time needed to understand it, $r_s = -.15, p > .05$ and $\tau = -.12, p > .05$, both one-tailed (see table 4.14). The descriptive statistic median of 5 in the first variable, standing for "I totally agree" (a high rating), and the median of 22 in the second variable, standing for "30 s", which is a long time when compared with the meaning of "quickly understood the experiment", support this outcome (cf. appendix E.12.3).

Table 4.15: Kendall's and Spearman's correlation coefficient tests on hypothesis H11.

Correlations[b]

			Experiment 1 - Felt immediate control over sound	Experiment 1 - Estimated time needed to control the sound
Kendall's tau_b	Experiment 1 - Felt immediate control over sound	Correlation Coefficient	1,000	-,256*
		Sig. (1-tailed)	.	,023
	Experiment 1 - Estimated time needed to control the sound	Correlation Coefficient	-,256*	1,000
		Sig. (1-tailed)	,023	.
Spearman's rho	Experiment 1 - Felt immediate control over sound	Correlation Coefficient	1,000	-,328*
		Sig. (1-tailed)	.	,019
	Experiment 1 - Estimated time needed to control the sound	Correlation Coefficient	-,328*	1,000
		Sig. (1-tailed)	,019	.

*. Correlation is significant at the 0.05 level (1-tailed).

b. Listwise N = 40

4.3.2.2 Hypothesis Test H11

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H11 that "participants need a time of less than 1 second to control sound in Experiment 1 if they more likely consider that they feel immediate control over it" showed that there was a statistically significant relationship between assuming that they felt immediate control over sound and the short estimated time to control sound in experiment 1, $r_s = -.33, p < .05$ and $\tau = -.26, p < .05$, both one-tailed (see table 4.15), where the negative correlation coefficients mean that the high rating contrasted with a low estimated time. The descriptive statistic median of 4 in the first variable, standing for "I agree" (a high rating), and the median of 17 in the second variable, standing for "10 s", support this outcome (cf. appendix E.12.3), although the estimated time was not less than 1 second as we hypothesized.

4.3.3 Experiment 2 Data Results

Parametric assumptions, such as the assumption of normality, were violated, as already mentioned in section 4.3. Therefore, we conducted non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H12 to H14, presented in section 3.6.4.3, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.9).

4.3.3.1 Hypothesis Test H12

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H12 that "the estimated time needed to understand Experiment 2 is less than 2 seconds if participants consider that they quickly understand it" showed that there was no significant relationship between admitting that participants quickly understood

Table 4.16: Kendall's and Spearman's correlation coefficient tests on hypothesis H13.

Correlations[b]

			Experiment 2 - It was easy to define the origin of sound	Experiment 2 - The gesture did not coincide with the origin of sound	Experiment 2 - The proposed experiment did not confuse me
Kendall's tau_b	Experiment 2 - It was easy to define the origin of sound	Correlation Coefficient	1,000	,059	,537**
		Sig. (1-tailed)	.	,330	,000
	Experiment 2 - The gesture did not coincide with the origin of sound	Correlation Coefficient	,059	1,000	,031
		Sig. (1-tailed)	,330	.	,410
	Experiment 2 - The proposed experiment did not confuse me	Correlation Coefficient	,537**	,031	1,000
		Sig. (1-tailed)	,000	,410	.
Spearman's rho	Experiment 2 - It was easy to define the origin of sound	Correlation Coefficient	1,000	,089	,575**
		Sig. (1-tailed)	.	,287	,000
	Experiment 2 - The gesture did not coincide with the origin of sound	Correlation Coefficient	,089	1,000	,045
		Sig. (1-tailed)	,287	.	,389
	Experiment 2 - The proposed experiment did not confuse me	Correlation Coefficient	,575**	,045	1,000
		Sig. (1-tailed)	,000	,389	.

**. Correlation is significant at the 0.01 level (1-tailed).

b. Listwise N = 42

experiment 2 and the estimated time needed to understand it, $r_s = -.04, p > .05$ and $\tau = -.03, p > .05$, both one-tailed (see appendix E.12.9). The descriptive statistic median of 5 in the first variable, standing for "I totally agree" (a high rating), and the median of 21 in the second variable, standing for "20 s", which is still a long time when compared with the meaning of "quickly understood the experiment", support this outcome (cf. appendix E.12.3).

4.3.3.2 Hypothesis Test H13

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H13 that "participants most likely reckon that their gesture does not coincide with the origin of sound in Experiment 2 and that the proposed experiment does not confuse them if they are convinced that it is easy to define the origin of sound" showed that there was no significant relationship between considering that it was easy to define the origin of sound in experiment 2 and admitting that their gesture did not coincide with the origin of sound, $r_s = .09, p > .05$ and $\tau = .06, p > .05$, both one-tailed (see table 4.16). In addition, there was no significant relationship between assuming that their gesture did not coincide with the origin of sound and considering that the proposed experiment did not confuse them, $r_s = .05, p > .05$ and $\tau = .03, p > .05$, both one-tailed. Nevertheless, there was a statistically significant relationship between admitting that it was easy to define the origin of sound in experiment 2 and that the proposed experiment did not confuse them, $r_s = .58, p < .01$ and $\tau = .54, p < .01$, both one-tailed. The descriptive statistic median of 4 in the first variable, standing for "I agree" (a high rating), and the median of 5 in the second variable, standing for "I totally agree [that the proposed experiment did not confuse me]", corroborate this conclusion.

4.3.3.3 Hypothesis Test H14

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H14 that "participants estimate a time of less than 1 second to control sound in Experiment 2 if they find that they feel immediate control over it" showed that there was a significant relationship between considering that participants felt immediate control over sound in experiment 2 and the estimated time needed to control it, $r_s = -.40, p < .05$ and $\tau = -.31, p < .05$, both one-tailed (see appendix E.12.9). The descriptive statistic median of 4 in the first variable, standing for "I agree" (a high rating), and the median of 13 in the second variable, standing for "5 s", support this outcome (cf. appendix E.12.3), although the estimated time was not less than 1 second as we hypothesized.

4.3.4 Experiment 3 Data Results

Parametric assumptions, such as the assumption of normality, were violated, as already mentioned in section 4.3. Therefore, we conducted non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H15 to H17, presented in section 3.6.4.4, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.10).

4.3.4.1 Hypothesis Test H15

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H15 that "the estimated time needed to understand Experiment 3 is less than 2 seconds if participants consider that they quickly understand it" showed that there was a statistically significant relationship between considering that participants quickly had understood experiment 3 and the estimated time needed to understand it, $r_s = -.34, p < .05$ and $\tau = -.28, p < .05$, both one-tailed (see appendix E.12.10). The descriptive statistic median of 5 in the first variable, standing for "I totally agree" (a high rating), and the median of 17 in the second variable, standing for "10 s", support this outcome (cf. appendix E.12.3), although the estimated time was not less than 2 seconds as we hypothesized.

4.3.4.2 Hypothesis Test H16

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H16 that "participants estimate a time of less than 3 seconds to locate sound in Experiment 3 if they admit that they quickly locate it" showed that there was a statistically significant relationship between considering that participants quickly located sound in experiment 3 and the estimated time needed to locate it, $r_s = -.60, p < .05$ and $\tau = -.50, p < .05$, both one-tailed (see appendix E.12.10). The descriptive statistic median of 3 in the first variable, standing for "Not always / Sometimes", and the median of 12 in the second variable, standing for "5 s", confirm this result (cf. appendix E.12.3), although the estimated time was not less than 3 seconds as we hypothesized.

4.3.4.3 Hypothesis Test H17

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H17 that "participants are most likely to admit that their gesture coincides with the origin of sound in Experiment 3 if they mean that it is

Table 4.17: Kendall's and Spearman's correlation coefficient tests on hypothesis H17.

Correlations[b]

			Experiment 3 - It was easy to define the origin of sound	Experiment 3 - Managed to localize the sound	Experiment 3 - The gesture coincided with the origin of sound
Kendall's tau_b	Experiment 3 - It was easy to define the origin of sound	Correlation Coefficient	1,000	,376**	,430**
		Sig. (1-tailed)	.	,003	,001
	Experiment 3 - Managed to localize the sound	Correlation Coefficient	,376**	1,000	,450**
		Sig. (1-tailed)	,003	.	,001
	Experiment 3 - The gesture coincided with the origin of sound	Correlation Coefficient	,430**	,450**	1,000
		Sig. (1-tailed)	,001	,001	.
Spearman's rho	Experiment 3 - It was easy to define the origin of sound	Correlation Coefficient	1,000	,411**	,472**
		Sig. (1-tailed)	.	,003	,001
	Experiment 3 - Managed to localize the sound	Correlation Coefficient	,411**	1,000	,504**
		Sig. (1-tailed)	,003	.	,000
	Experiment 3 - The gesture coincided with the origin of sound	Correlation Coefficient	,472**	,504**	1,000
		Sig. (1-tailed)	,001	,000	.

**. Correlation is significant at the 0.01 level (1-tailed).

b. Listwise N = 42

easy to define the origin of sound and assume that they manage to locate sound" showed that there was a statistically significant positive relationship between considering that it was easy to define the origin of sound in experiment 3 and assuming that participants managed to locate sound, $r_s = .41, p < .05$ and $\tau = .38, p < .05$, both one-tailed (see table 4.17). The descriptive statistic median of 3 in the first variable, standing for "Not always / Sometimes", and the median of 4 in the second variable, standing for "I agree", corroborate this result (cf. appendix E.12.3). In addition, there was also a statistically significant positive relationship between admitting that it was easy to define the origin of sound and that the gesture coincided with the origin of sound, $r_s = .47, p < .05$ and $\tau = .43, p < .05$, both one-tailed. The descriptive statistic median of 3 in the first variable, standing for "Not always / Sometimes", and the mode of 4 in the second variable, standing for "I agree", confirm this conclusion. Finally, there was once again a statistically significant positive relationship between considering that participants managed to locate sound and that the gesture coincided with the origin of sound, $r_s = .50, p < .05$ and $\tau = .45, p < .05$, both one-tailed. The descriptive statistic median of 4 in both variables, standing for "I agree" (a high rating), support this outcome.

4.3.5 Experiments 1 and 2 Data Results

Parametric assumptions, such as the assumption of interval data, were violated, as already mentioned in section 4.3. Therefore, we performed non-parametric Wilcoxon signed-rank tests on hypotheses H18 and H19, presented in section 3.6.4.5, in which differences between experiments 1 and 2, where the same participants have been used, were examined (Field, 2009, p. 552) (see appendix E.12.11). Thus, we had one categorical

independent variable (experimental condition) with two levels, because it was manipulated in two ways, and one ordinal dependent variable (answer on a 5-point scale or time estimates).

4.3.5.1 Hypothesis Test H18

The Wilcoxon signed-rank test on the hypothesis H18 that ”participants feel more immediate control over sound in Experiment 1 than in Experiment 2” showed that there was no statistically significant difference of the immediate control over sound between experiments 1 and 2 ($Z = -1.86, p = .062 > .05$), $T = 52, p > .05, r = -.21$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.86}{\sqrt{39 \times 2}} = -.21$ (see appendix E.12.11), although there was a slightly increase of immediate control in experiment 2 (median of 4 (4 to 5), standing for ”I agree”) relatively to experiment 1 (median of 4 (3 to 4), standing for ”I agree” as well).

4.3.5.2 Hypothesis Test H19

The Wilcoxon signed-rank test on the hypothesis H19 that ”participants estimate a lower time needed to control sound in Experiment 1 than in Experiment 2” showed that there was no statistically significant difference in estimating times needed to control sound between experiments 1 and 2 ($Z = -1.50, p = .13 > .05$), $T = 116.5, p > .05, r = -.17$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.50}{\sqrt{39 \times 2}} = -.17$ (see appendix E.12.11), although the estimated time needed to control the sound in experiment 1 (median of 17 (9 to 23), standing for ”10 s”) was a little bit higher than in experiment 2 (median of 13 (6 to 25), standing for ”5 s”) (see figure 4.28).

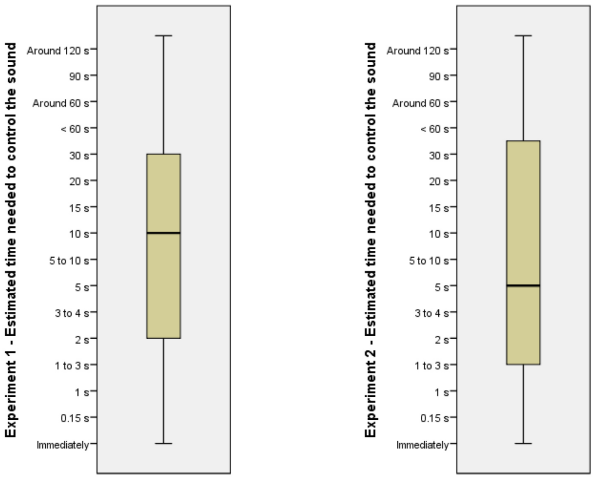


Figure 4.28: Boxplots of estimated times needed to control sound (hypothesis test H19).

4.4 Experiment Results By Musical Knowledge

Since we considered two groups of participants from the same sample containing all volunteers, one without and the other with musical knowledge (see section 4.2.1), with the purpose of testing hypothetical deviations in the results of the analysed data based on musical knowledge, and the same ordinal experimental condition items of part 2 of the Inquiry Mode Questionnaire (InQ) were used (see appendix A and figures 4.1, 4.2, and 4.3), at least one of the four assumptions required to be met in parametric tests based on the normal

distribution, such as the interval level assumption, was violated. Thus, we performed non-parametric tests (cf. section 4.1) on the same hypothesis H1 to H19, but this time based on musical knowledge.

4.4.1 Group Without Any Musical Knowledge

4.4.1.1 Experiments 1, 2, and 3 Data Results

After splitting the data into two groups by musical knowledge, we performed non-parametric Friedman's ANOVA tests on hypotheses H1 to H9, presented in section 3.6.4.1, because we wanted to examine differences between the two groups in the three experimental conditions (Field, 2009, pp. 573, 581). Thus, we had one categorical independent variable (*experimental condition*) with three levels and one ordinal dependent variable (*answer on a 5-point scale or time estimates*).

4.4.1.1.1 Hypothesis Test H1

The Friedman's ANOVA test on hypothesis H1 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 5 (4 to 5), 5 (4 to 5), and 5 (4.75 to 5), respectively; and 2) the significance value was equal to .17, or .25 for the exact significance, which was above .05.

Therefore, we can conclude that there were no statistically significant differences in how quickly participants without any musical knowledge would understand the three experiments, $\chi^2(2) = 3.50, p > .05$. The descriptive statistic median of 5 in all experiments, standing for "I totally agree", supports the idea that participants rated their quick understanding highly (see appendix E.12.5). No *post hoc* (cf. 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.1.2 Hypothesis Test H2

The Friedman's ANOVA test on hypothesis H2 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 24.5 (14.5 to 28.0), 19.5 (10.75 to 27.75), and 17.0 (9.75 to 27.75), respectively; and 2) the significance value was equal to .20, or .22 for the exact significance, which was above .05.

Therefore, we can conclude that there were no statistically significant differences in the estimation of time it took for participants without any musical knowledge to understand the three experiments, $\chi^2(2) = 3.27, p > .05$. However, the descriptive statistic median of 22 in experiment 1, standing for "30 s", the median of 19.5, standing for "15 s" or "Around 15 s", and the median of 17, standing for "10 s", reveal that the estimated times were not less than the 2 seconds we hypothesized (see appendix E.12.5). No *post hoc* (cf. 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.1.3 Hypothesis Test H3

The Friedman's ANOVA test on hypothesis H3 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 4 (4 to 5), 5 (4 to 5), and 4 (4 to 5), respectively; and 2) the significance value was equal to .61, or .74 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in how easily participants without any musical knowledge interacted with the installation in the three experiments, $\chi^2(2) = 1.00, p > .05$. The descriptive statistic median of 4.5 in experiment 1, standing for "I agree" or "I totally agree", the median of 5 in experiment 2, standing for "I totally agree", and the median of 4 in experiment 3, standing for "I agree", support this idea and confirm that participants rated their ease of interaction highly (cf. appendix E.12.5). No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.1.4 Hypothesis Test H4

The Friedman's ANOVA test on hypothesis H4 showed that (see appendix E.12.12): 1) the median levels were of 4 (4 to 5) in both experimental conditions 1 and 2, and of 3 (2 to 4.5) in experiment 3; and 2) the significance value was equal to .06, or .07 for the exact significance, which was slightly above .05.

Therefore, we can conclude that there were no statistically significant differences in how easily volunteers without any musical knowledge defined the origin of sound in the three experiments, $\chi^2(2) = 5.71, p > .05$. No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.1.5 Hypothesis Test H5

The Friedman's ANOVA test on hypothesis H5 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 5 (4 to 5), 4.5 (3.75 to 5), and 4 (4 to 5), respectively; and 2) the significance value was equal to .22, or .24 for the exact significance, which was above .05.

Therefore, we can conclude that there were no statistically significant differences in how highly participants without any musical knowledge rated the suggested gesture as being adequate to any of the three experiments, $\chi^2(2) = 3.07, p > .05$. The descriptive statistic median of 5 in experiment 1, standing for "I totally agree", the median of 4.5 in experiment 2, standing for "I agree" or "I totally agree", and the median of 4 in experiment 3, standing for "I agree", support this idea and confirm that participants rated its adequacy highly (cf. appendix E.12.5). No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.1.6 Hypothesis Test H6

The Friedman's ANOVA test on hypothesis H6 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 4 (4 to 5), 4 (3.5 to 5), and 4 (3 to 4.5), respectively; and 2) the significance value was equal to .37, or .52 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in the degree of appreciation whether the system's response to gesture was immediate in any of the three experiments, $\chi^2(2) = 2.00, p > .05$. The descriptive statistic median of 4 in the three experiments, standing for "I agree", corroborates this idea and confirms that volunteers without any musical knowledge more likely considered that the

system's response to gesture was immediate (cf. appendix E.12.5). No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

We also conducted a Friedman's ANOVA test on the inverted questions "The system's response to the gesture was fast", as already explained in section 4.3.1.6, which showed that: 1) the median levels for the three experimental conditions were of 4 (2 to 4), 4 (3.75 to 5), and 4 (3 to 4.25), respectively; and 2) the significance value was equal to .53, or .55 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in the degree of appreciation whether the system's response to gesture was fast in any of the three experiments, $\chi^2(2) = 1.28, p > .05$. The descriptive statistic median of 4 in the three experiments, standing for "I agree", supports this idea and confirms that volunteers without any musical knowledge more likely considered that the system's response to gesture was fast (cf. appendix E.12.5). Once again, no *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.1.7 Hypothesis Test H7

The Friedman's ANOVA test on hypothesis H6 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 17 (6.75 to 24.75), 15.5 (8.25 to 20.75), and 17 (7.75 to 28.5), respectively; and 2) the significance value was equal to .96, or .99 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in the estimation of the shorter times of the system's response to gesture in the three experiments, $\chi^2(2) = .087, p > .05$. The descriptive statistic median of 17 in experiments 1 and 3, standing for "2 s", and the median of 15.5, standing for "1 to 3 s" or "1 to 5 s", support this idea and confirm that participants without any musical knowledge estimated a short time of the system's response to gesture (cf. appendix E.12.5). No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.1.8 Hypothesis Test H8

The Friedman's ANOVA test on hypothesis H8 showed that (see appendix E.12.12): 1) the median levels were of 5 (5 to 5) in experiment 1, and 5 (4.25 to 5) in both experimental conditions 2 and 3; and 2) the significance value was equal to .45, or .62 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in how strongly volunteers without any musical knowledge felt surrounded by sound in the installation, $\chi^2(2) = 1.60, p > .05$. The descriptive statistic median of 5 in the three experiments, standing for "I totally agree", corroborates this idea and confirms that participants rated the feeling of being surrounded by sound in the installation highly (cf. appendix E.12.5). No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.1.9 Hypothesis Test H9

The Friedman's ANOVA test on hypothesis H9 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 4 (4 to 5), 4 (1.75 to 4), and 3 (2 to 4), respectively; and 2) the significance value was equal to .087, or .084 for the exact significance, which was above .05.

Therefore, we can conclude that there were no statistically significant differences in how volunteers without any musical knowledge considered that their gesture coincided with the origin of sound, $\chi^2(2) = 1.60, p > .05$. No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.1.2 Experiment 1 Data Results

Parametric assumptions, such as the interval level assumption, were violated, as already mentioned in section 4.4. Therefore, we conducted non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H10 and H11, presented in section 3.6.4.2, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.13).

4.4.1.2.1 Hypothesis Test H10

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H10 that "the estimated time needed to understand Experiment 1 is less than 2 seconds if participants consider that they quickly understand it" showed that there was a statistically significant relationship between admitting that they quickly understood experiment 1 and the estimated time needed to understand it, $r_s = -.49, p < .05$ and $\tau = -.43, p < .05$, both one-tailed (see appendix E.12.13), where the negative correlation coefficients mean that the high rating contrasted with a low estimated time, although the latter was not less than 2 seconds as we hypothesized. The descriptive statistic median of 5 in the first variable, standing for "I totally agree" (a high rating), and the median of 22 in the second variable, standing for "30 s", support this result. However, 30 s is still a long time when compared with the meaning of "quickly understood the experiment" (cf. appendix E.12.5).

4.4.1.2.2 Hypothesis Test H11

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H11 that "participants need a time of less than 1 second to control sound in Experiment 1 if they more likely consider that they feel immediate control over sound" showed that there was no significant relationship between admitting that they felt immediate control over sound and the short estimated time to control sound in experiment 1, $r_s = -.21, p > .05$ and $\tau = -.13, p > .05$, both one-tailed (see appendix E.12.13). The descriptive statistic median of 4 in the first variable, standing for "I agree" (a high rating), and the median of 21 in the second variable, standing for "20 s", confirm this outcome (cf. appendix E.12.5).

4.4.1.3 Experiment 2 Data Results

Non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H12, H13, and H14, presented in section 3.6.4.3, were conducted as well, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.14).

4.4.1.3.1 Hypothesis Test H12

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H12 that "the estimated time needed to understand Experiment 2 is less than 2 seconds if participants consider that they quickly understand it" showed that there was no significant relationship between considering that participants quickly understood experiment 2 and the estimated time needed to understand it, $r_s = -.11, p > .05$ and $\tau = -.081, p > .05$, both one-tailed (see appendix E.12.14). The descriptive statistic median of 5 in the first variable, standing for "I totally agree" (a high rating), and the median of 19.5 in the second variable, standing for "15 s" or "Around 15 s", which is still a long time when compared with the meaning of "quickly understood the experiment", support this outcome (cf. appendix E.12.5).

4.4.1.3.2 Hypothesis Test H13

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H13 that "participants most likely reckon that their gesture does not coincide with the origin of sound in Experiment 2 and that the proposed experiment does not confuse them if they are convinced that it is easy to define the origin of sound" showed that there was a statistically significant relationship between assuming that it was easy to define the origin of sound in experiment 2 and that their gesture did not coincide with the origin of sound, $r_s = .61, p < .05$ and $\tau = .53, p < .05$, both one-tailed (see appendix E.12.14), although the rating of the latter had a median of 2, standing for "I disagree". In addition, there was also a statistically significant relationship between admitting that it was easy to define the origin of sound in experiment 2 and that the proposed experiment did not confuse them, $r_s = .70, p < .01$ and $\tau = .67, p < .01$, both one-tailed. Nevertheless, there was no significant relationship between considering that their gesture did not coincide with the origin of sound and that the proposed experiment did not confuse them, $r_s = .090, p > .05$ and $\tau = .90, p > .05$, both one-tailed.

4.4.1.3.3 Hypothesis Test H14

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H14 that "participants estimate a time of less than 1 second to control sound in Experiment 2 if they find that they feel immediate control over sound" showed that there was no statistically significant relationship between considering that participants felt immediate control over sound in experiment 2 and the estimated time needed to control it, $r_s = -.45, p > .05$ and $\tau = -.32, p > .05$, both one-tailed (see appendix E.12.14). The descriptive statistic median of 4 in the first variable, standing for "I agree" (a high rating), and the median of 17 in the second variable, standing for "10 s", support this outcome (cf. appendix E.12.5).

4.4.1.4 Experiment 3 Data Results

Non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H15, H16, and H17, presented in section 3.6.4.4, were conducted as well, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.15).

4.4.1.4.1 Hypothesis Test H15

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H15 that "the estimated time needed to understand Experiment 3 is less than 2 seconds if participants consider that they quickly understand it" showed that there was no significant relationship between assuming that participants quickly understood experiment 3 and the estimated time needed to understand it, $r_s = -.44, p > .05$ and $\tau = -.38, p > .05$, both one-tailed (see appendix E.12.15). The descriptive statistic median of 5 in the first variable, standing for "I totally agree" (a high rating), and the median of 17 in the second variable, standing for "10 s", support this outcome (cf. appendix E.12.5).

4.4.1.4.2 Hypothesis Test H16

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H16 that "participants estimate a time of less than 3 seconds to locate sound in Experiment 3 if they admit that they quickly locate it" showed that there was a statistically significant relationship between considering that participants quickly located sound in experiment 3 and the estimated time needed to locate it, $r_s = -.68, p < .01$ and $\tau = -.58, p < .01$, both one-tailed (see appendix E.12.15). The descriptive statistic median of 3 in the first variable, standing for "Not always / Sometimes", and the median of 12 in the second variable, standing for "5 s", confirm this result (cf. appendix E.12.5), although the estimated time was not less than 3 seconds as we hypothesized.

4.4.1.4.3 Hypothesis Test H17

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H17 that "participants are most likely to admit that their gesture coincides with the origin of sound in Experiment 3 if they mean that it is easy to define the origin of sound and assume that they manage to locate sound" showed that there was a statistically significant positive relationship between admitting that it was easy to define the origin of sound in experiment 3 and that participants managed to locate sound, $r_s = .52, p < .05$ and $\tau = .48, p < .05$, both one-tailed (see appendix E.12.15). In addition, there was also a statistically significant positive relationship between assuming that it was easy to define the origin of sound and that the gesture coincided with the origin of sound, $r_s = .49, p < .05$ and $\tau = .45, p < .05$, both one-tailed. Finally, there was once again a statistically significant positive relationship between considering that participants managed to locate sound and that the gesture coincided with the origin of sound, $r_s = .59, p < .05$ and $\tau = .51, p < .05$, both one-tailed.

4.4.1.5 Experiments 1 and 2 Data Results

We performed non-parametric Wilcoxon signed-rank tests on hypotheses H18 and H19, presented in section 3.6.4.5, in which differences between experiments 1 and 2, where the same participants had been involved, were examined (Field, 2009, p. 552) (see appendix E.12.16).

4.4.1.5.1 Hypothesis Test H18

The Wilcoxon signed-rank test on the hypothesis H18 that "participants feel more immediate control over sound in Experiment 1 than in Experiment 2" showed that there was no statistically significant difference of the immediate control over sound between experiments 1 and 2 ($Z = -1.67, p = .096 > .05$), $T = 3, p > .05, r = -.33$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.67}{\sqrt{13 \times 2}} = -.33$ (see appendix E.12.16), although there was a slightly increase of immediate control in experiment 2 (median of 4 (3.5 to 5), standing for "I agree") relatively to experiment 1 (median of 4 (3 to 4), standing for "I agree" as well).

4.4.1.5.2 Hypothesis Test H19

The Wilcoxon signed-rank test on the hypothesis H19 that "participants estimate a lower time needed to control sound in Experiment 1 than in Experiment 2" showed that there was no statistically significant difference in estimating times needed to control sound between experiments 1 and 2 ($Z = -.66, p = .51 > .05$), $T = 17, p > .05, r = -.13$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-.66}{\sqrt{13 \times 2}} = -.13$ (see appendix E.12.16), although the estimated time needed to control the sound in experiment 1 (median of 21 (12 to 26.5), standing for "20 s") was a little bit higher than in experiment 2 (median of 17 (12.5 to 25.5), standing for "10 s").

4.4.2 Group With Musical Knowledge

4.4.2.1 Experiments 1, 2, and 3 Data Results

After splitting the data into two groups by musical knowledge, we performed non-parametric Friedman's ANOVA tests on hypotheses H1 to H9, presented in section 3.6.4.1, because we wanted to examine differences between the two groups in the three experimental conditions (Field, 2009, pp. 573, 581). Thus, we had one categorical independent variable (*experimental condition*) with three levels and one ordinal dependent variable (*answer on a 5-point scale or time estimates*).

4.4.2.1.1 Hypothesis Test H1

The Friedman's ANOVA test on hypothesis H1 showed that (see appendix E.12.12): 1) the median levels were of 5 (5 to 5) in experiments 1 and 2, and of 5 (4 to 5) in experiment 3; and 2) the significance value was equal to .10, or .14 for the exact significance, which was above .05.

Therefore, we can conclude that there were no statistically significant differences in how quickly participants with musical knowledge would understand the three experiments, $\chi^2(2) = 4.57, p > .05$. The descriptive

statistic median of 5 in all experiments, standing for "I totally agree", supports the idea that participants considered that they understood the three experiments quickly (see appendix E.12.5). No *post hoc* (cf. 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.2.1.2 Hypothesis Test H2

The Friedman's ANOVA test on hypothesis H2 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 22 (17 to 28), 21 (14 to 25.5), and 17 (14 to 26.5), respectively; and 2) the significance value was equal to .009, or .007 for the exact significance, which was below .05. Therefore, we can conclude that there was a statistically significant difference in the estimation of time it took for participants with musical knowledge to understand the three experiments, $\chi^2(2) = 9.49, p < .05$.

However, because the Friedman's ANOVA test was significant (Field, 2009, p. 577) and because it does not pinpoint which experiments in particular differ from each other, *post hoc* procedures were performed. Thus, we used three separate Wilcoxon signed-rank tests based on the different combinations of our experimental conditions, in which experiment 1 was compared with experiment 2, experiment 1 with experiment 3, and experiment 2 with experiment 3 (see appendix E.12.7). In order to ensure that Type I errors, which occur "... when we believe that there is a genuine effect in our population, when in fact there isn't" (Field, 2009, p. 56), would not build up to more than .05 because we were making multiple comparisons, we used a Bonferroni correction. This means that we used a critical value for significance of .05 divided by the number of tests we carried out, that is, .05 / 3 = .017, instead of .05, in each test (Field, 2009, p. 577).

The results showed that there were no significant differences between experiment 1 and experiment 2, although there was an overall reduction of the estimated time in experiment 2 (median of 21, standing for "20 s") relatively to experiment 1 (median of 22, standing for "30 s") ($Z = -1.95, p = .052 > .017$), $T = 41, p > .017, r = -.26$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.95}{\sqrt{29 \times 2}} = -.26$ (see figure 4.29).

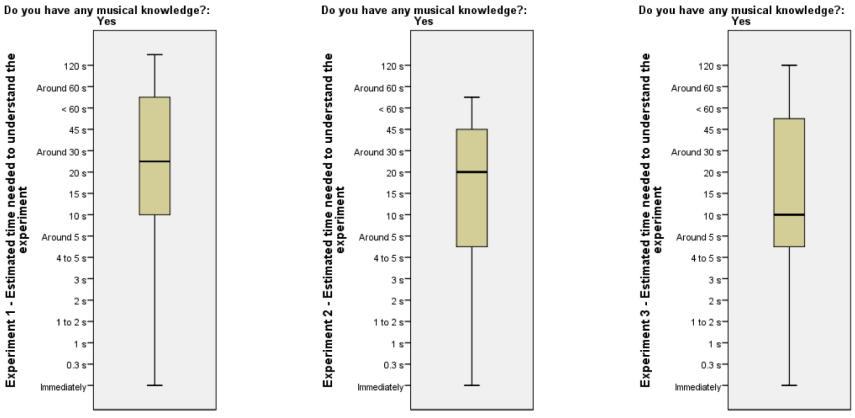


Figure 4.29: Boxplots of estimated times needed to understand the experiments by musical knowledge (hypothesis tests H2).

In addition, there were also no significant differences between experiment 1 and experiment 3, although there was a general reduction of the estimated time in experiment 3 (median of 17, standing for "10 s") relatively to experiment 1 (median of 22, standing for "30 s") ($Z = -2.12, p = .034 > .017$), $T = 42.5, p > .017, r = -.28$,

where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.12}{\sqrt{29 \times 2}} = -.28$.

Finally, there were no significant differences between experiment 2 and experiment 3 as well, although there was an overall reduction of the estimated time in experiment 3 (median of 17, standing for "10 s") relatively to experiment 2 (median of 21, standing for "20 s") ($Z = -.13, p = .90 > .017$), $T = 50.5, p > .017, r = -.017$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-.13}{\sqrt{29 \times 2}} = -.017$. Thus, the estimated times were not less than the 2 seconds we hypothesized.

4.4.2.1.3 Hypothesis Test H3

The Friedman's ANOVA test on hypothesis H3 showed that (see appendix E.12.12): 1) the median levels were of 4 (4 to 5) in experiments 1 and 2, and of 4 (3 to 5) in experiment 3; and 2) the significance value was equal to .31, or .32 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in how easily participants with musical knowledge interacted with the installation in the three experiments, $\chi^2(2) = 2.34, p > .05$. The descriptive statistic median of 4 in the three experiments, standing for "I agree", supports this idea and confirms that participants rated their ease of interaction highly (cf. appendix E.12.5). No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.2.1.4 Hypothesis Test H4

The Friedman's ANOVA test on hypothesis H4 showed that (see appendix E.12.12): 1) the median levels were of 4 (4 to 5) in experimental condition 1, of 4 (4 to 4) in experiment 2, and of 3 (3 to 4) in experiment 3; and 2) the significance value was equal to .011, or .010 for the exact significance, which was below .05. Therefore, we can conclude that participants with musical knowledge admitted that they easily defined the origin of sound in significantly different ways, $\chi^2(2) = 9.00, p < .05$.

Post hoc procedures were performed because the Friedman's ANOVA test was significant (Field, 2009, p. 577). Thus, we used three Wilcoxon signed-rank tests, in which experiment 1 was compared with experiment 2, experiment 1 with experiment 3, and experiment 2 with experiment 3. We applied a Bonferroni correction, which means that we used a critical value for significance of .017, instead of .05, in each test (cf. section 4.3.1.1).

The outcomes show that there were no significant differences between experiments 1 and 2 ($Z = -1.06, p = .29 > .017$), $T = 36, p > .017, r = -.14$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.06}{\sqrt{29 \times 2}} = -.14$, although it was a little bit easier to define the origin of sound in experiment 1 than in experiment 2 (the higher sum of ranks associated to the negative ranks indicates that).

However, there were statistically significant differences between experiments 1 and 3 ($Z = -3.07, p = .002 < .017$), $T = 22.5, p < .017, r = -.40$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-3.07}{\sqrt{29 \times 2}} = -.40$ means that there was a medium to large effect or a moderate to big difference in easily defining the origin of sound between experiments 1 and 3 (Field, 2009, pp. 57, 170). Thus, it was easier to define the origin of sound in experiment 1 than in experiment 3.

There were no significant differences between experiments 2 and 3 ($Z = -1.55, p = .12 > .017$), $T = 64.5, p > .017, r = -.20$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.55}{\sqrt{29 \times 2}} = -.20$, although it was easier to define the origin of sound in experiment 2 than in experiment 3, according to the higher sum of ranks associated to the negative ranks.

4.4.2.1.5 Hypothesis Test H5

The Friedman's ANOVA test on hypothesis H5 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 4 (4 to 5); and 2) the significance value was equal to .85, or 1.00 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in how highly participants with musical knowledge rated the suggested gesture as being adequate to any of the three experiments, $\chi^2(2) = .33, p > .05$. The descriptive statistic median of 4 in all experiments, standing for "I agree", supports this idea and confirms that participants rated its adequacy highly (cf. appendix E.12.5). No *post hoc* (see section 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.2.1.6 Hypothesis Test H6

The Friedman's ANOVA test on hypothesis H6 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 4 (3 to 4), 4 (4 to 4), and 4 (3 to 4), respectively; and 2) the significance value was equal to .036, or .037 for the exact significance, which was below .05. Therefore, we can conclude that there were statistically significant differences in the degree of appreciation whether the system's response to gesture was immediate, $\chi^2(2) = 6.63, p < .05$.

Post hoc procedures were performed because the Friedman's ANOVA test was significant (Field, 2009, p. 577). Thus, we used three separate Wilcoxon signed-rank tests, in which experiment 1 was compared with experiment 2, experiment 1 with experiment 3, and experiment 2 with experiment 3. We applied a Bonferroni correction, which means that we used a critical value for significance of .017, instead of .05, in each test (see section 4.3.1.1).

The outcomes show that there were no significant differences between experiments 1 and 2 ($Z = -2.24, p = .025 > .017$), $T = 19.5, p > .017, r = -.29$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.24}{\sqrt{29 \times 2}} = -.29$, although the higher sum of ranks associated to the positive ranks indicates that the rating tendency was slightly higher in experiment 2 than in experiment 1.

There were also no significant differences between experiments 1 and 3 ($Z = -.28, p = .78 > .017$), $T = 25, p > .017, r = -.037$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-.28}{\sqrt{29 \times 2}} = -.037$, but the higher sum of ranks associated to the negative ranks suggests that the rating tendency was higher in experiment 1 than in experiment 3.

Finally, there were no significant differences between experiments 2 and 3 as well ($Z = -2.30, p = .022 > .017$), $T = 8, p > .017, r = -.30$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.30}{\sqrt{29 \times 2}} = -.30$. The higher sum of ranks associated to the negative ranks indicates that the rating tendency was a little bit higher in experiment 2 than in experiment 3. The descriptive statistic median of 4 in all experiments, standing for "I agree" (cf. appendix E.12.5), confirms

that participants with musical knowledge most likely considered that the system's response to gesture was immediate.

We also conducted a Friedman's ANOVA test on the inverted questions "The system's response to the gesture was fast", as already explained in section 4.3.1.6, which showed that: 1) the median levels for the three experimental conditions were of 4 (3 to 4.5), 4 (4 to 5), and 4 (3 to 4.5), respectively; and 2) the significance value was equal to .11, or .12 for the exact significance, which was above .05.

Therefore, we can conclude that there were no statistically significant differences in the degree of appreciation whether the system's response to gesture was fast in any of the three experiments, $\chi^2(2) = 4.44, p > .05$. The descriptive statistic median of 4 in the three experiments, standing for "I agree", supports this idea and confirms that volunteers with musical knowledge more firmly considered that the system's response to gesture was fast (cf. appendix E.12.5). Once again, no *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.2.1.7 Hypothesis Test H7

The Friedman's ANOVA test on hypothesis H6 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 12.5 (9 to 21.5), 12 (9 to 19.25), and 12 (9.75 to 21), respectively; and 2) the significance or the exact significance values were both equal to .21, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in the estimation of the shorter times of the system's response to gesture in the three experiments, $\chi^2(2) = 3.17, p > .05$. The descriptive statistic median of 12 in all experiments, standing for "1 s", supports this idea and confirms that participants with musical knowledge estimated a short time of the system's response to gesture (cf. appendix E.12.5). No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.2.1.8 Hypothesis Test H8

The Friedman's ANOVA test on hypothesis H8 showed that (see appendix E.12.12): 1) the median levels were of 5 (4 to 5) in all experimental conditions; and 2) the significance value was equal to .24, or .26 for the exact significance, which was well above .05.

Therefore, we can conclude that there were no statistically significant differences in how strongly volunteers with musical knowledge felt surrounded by sound in the installation, $\chi^2(2) = 2.87, p > .05$. The descriptive statistic median of 5 in the three experiments, standing for "I totally agree", corroborates this idea and confirms that participants with musical knowledge considered that they felt surrounded by sound in the installation (cf. appendix E.12.5). No *post hoc* (see 4.3.1.1) procedures were performed, because the Friedman's ANOVA test was not significant (Field, 2009, p. 577).

4.4.2.1.9 Hypothesis Test H9

The Friedman's ANOVA test on hypothesis H9 showed that (see appendix E.12.12): 1) the median levels for the three experimental conditions were of 4 (4 to 5), 4 (1.5 to 4), and 4 (3 to 4), respectively; and 2) the significance

value was equal to .033, or .031 for the exact significance, which was below .05. Therefore, we can conclude that there were statistically significant differences in how participants considered that their gesture coincided with the origin of sound, $\chi^2(2) = 6.84, p < .05$.

Post hoc procedures were performed because the Friedman's ANOVA test was significant (Field, 2009, p. 577). Thus, we used three separate Wilcoxon signed-rank tests, in which experiment 1 was compared with experiment 2, experiment 1 with experiment 3, and experiment 2 with experiment 3. We applied a Bonferroni correction, which means that we used a critical value for significance of .017, instead of .05, in each test (cf. section 4.3.1.1).

The outcomes show that there was a statistically significant reduction tendency (as can be read from the highest sum of ranks associated to the negative ranks) in considering that the gesture coincided with the origin of sound in experiment 2 relatively to experiment 1 ($Z = -2.72, p = .007 < .017$), $T = 38.5, p < .017, r = -.36$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.72}{\sqrt{29 \times 2}} = -.36$, which means that there was a medium effect or a moderate difference in this rating (Field, 2009, pp. 57, 170).

There was also a statistically significant reduction tendency in admitting that the gesture coincided with the origin of sound in experiment 3 relatively to experiment 1 ($Z = -2.84, p = .005 < .017$), $T = 27, p < .017, r = -.37$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-2.84}{\sqrt{29 \times 2}} = -.37$, which corresponds to a medium to large effect or a moderate to big difference in this rating.

Nevertheless, there was no significant difference in assuming that the gesture coincided with the origin of sound in experiments 2 and 3 ($Z = -1.73, p = .083 > .017$), $T = 52.5, p > .017, r = -.23$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.73}{\sqrt{29 \times 2}} = -.23$, although a small effect or difference could be still observed: the rating tendency was slightly lower in experiment 2 than in 3. Finally, the descriptive statistic median of 4 in all experiments, standing for "I agree" (cf. appendix E.12.5 as well), confirms that participants with musical knowledge more firmly considered that the gesture coincided with the origin of sound.

4.4.2.2 Experiment 1 Data Results

Parametric assumptions, such as the interval level assumption, were violated, as already mentioned in section 4.4. Therefore, we conducted non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H10 and H11, presented in section 3.6.4.2, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.13).

4.4.2.2.1 Hypothesis Test H10

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H10 that "the estimated time needed to understand Experiment 1 is less than 2 seconds if participants consider that they quickly understand it" showed that there was no significant relationship between admitting that they quickly understood experiment 1 and the estimated time needed to understand it, $r_s = .088, p > .05$ and $\tau = .074, p > .05$, both one-tailed (see appendix E.12.13).

In addition, the descriptive statistic median of 5 in the first variable, standing for "I totally agree", and the median of 22 in the second variable, standing for "30 s", support the idea that participants with musical

knowledge more likely admitted that they quickly understood Experiment 1, but that the estimated time of 30 s is still a long time when compared with the meaning of "quickly understood the experiment" (cf. appendix E.12.5).

4.4.2.2.2 Hypothesis Test H11

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H11 that "participants need a time of less than 1 second to control sound in Experiment 1 if they more likely consider that they feel immediate control over sound" showed that there was a statistically significant relationship between admitting that they felt immediate control over sound and the short estimated time to control sound in experiment 1, $r_s = -.38, p < .05$ and $\tau = -.31, p < .05$, both one-tailed (see appendix E.12.13). The descriptive statistic median of 4 in the first variable, standing for "I agree" (a high rating), and the median of 17 in the second variable, standing for "10 s", confirm this outcome (cf. appendix E.12.5), although the estimated time was not less than 1 second as we hypothesized.

4.4.2.3 Experiment 2 Data Results

Non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H12, H13, and H14, presented in section 3.6.4.3, were conducted as well, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.14).

4.4.2.3.1 Hypothesis Test H12

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H12 that "the estimated time needed to understand Experiment 2 is less than 2 seconds if participants consider that they quickly understand it" showed that there was no significant relationship between assuming that participants with musical knowledge quickly understood experiment 2 and the estimated time needed to understand it, $r_s = -.16, p > .05$ and $\tau = -.14, p > .05$, both one-tailed (see appendix E.12.14). The descriptive statistic median of 5 in the first variable, standing for "I totally agree" (a high rating), and the median of 21 in the second variable, standing for "20 s", which is still a long time when compared with the meaning of "quickly understood the experiment", support this outcome (cf. appendix E.12.5).

4.4.2.3.2 Hypothesis Test H13

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H13 that "participants most likely reckon that their gesture does not coincide with the origin of sound in Experiment 2 and that the proposed experiment does not confuse them if they are convinced that it is easy to define the origin of sound" showed that there was no significant relationship between considering that it was easy to define the origin of sound in experiment 2 and that their gesture did not coincide with the origin of sound, $r_s = -.16, p > .05$ and $\tau = -.17, p > .05$, both one-tailed (see appendix E.12.14). In addition, there was also no significant relationship between admitting that their gesture did not coincide with the origin of sound and that the proposed

experiment did not confuse them, $r_s = .019, p > .01$ and $\tau = .007, p > .01$, both one-tailed. Nevertheless, there was a statistically significant relationship between assuming that it was easy to define the origin of sound in experiment 2 and that the proposed experiment did not confuse them, $r_s = .54, p < .01$ and $\tau = .50, p < .01$, both one-tailed.

4.4.2.3.3 Hypothesis Test H14

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H14 that "participants estimate a time of less than 1 second to control sound in Experiment 2 if they find that they feel immediate control over sound" showed that there was a statistically significant relationship between admitting that participants with musical knowledge felt immediate control over sound in experiment 2 and the estimated time needed to control it, $r_s = -.45, p < .05$ and $\tau = -.36, p < .05$, both one-tailed (see appendix E.12.14). The descriptive statistic median of 4 in the first variable, standing for "I agree" (a high rating), and the median of 13 in the second variable, standing for "5 s", support this outcome (cf. appendix E.12.5), although the estimated time was not less than 1 second as we hypothesized.

4.4.2.4 Experiment 3 Data Results

Non-parametric Kendall's and Spearman's correlation coefficient tests on hypotheses H15, H16, and H17, presented in section 3.6.4.4, were conducted as well, in order to examine relationships or correlations between ordinal variables (Field, 2009, p. 180) (see also appendix E.12.15).

4.4.2.4.1 Hypothesis Test H15

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H15 that "the estimated time needed to understand Experiment 3 is less than 2 seconds if participants consider that they quickly understand it" showed that there was no significant relationship between considering that participants quickly understood experiment 3 and the estimated time needed to understand it, $r_s = -.30, p > .05$ and $\tau = -.25, p > .05$, both one-tailed (see appendix E.12.15). The descriptive statistic median of 5 in the first variable, standing for "I totally agree" (a high rating), and the median of 17 in the second variable, standing for "10 s", support this outcome (cf. appendix E.12.5).

4.4.2.4.2 Hypothesis Test H16

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H16 that "participants estimate a time of less than 3 seconds to locate sound in Experiment 3 if they admit that they quickly locate it" showed that there was a statistically significant relationship between assuming that participants quickly located sound in experiment 3 and the estimated time needed to locate it, $r_s = -.56, p < .01$ and $\tau = -.45, p < .01$, both one-tailed (see appendix E.12.15). The descriptive statistic median of 3 in the first variable, standing for "Not always / Sometimes", and the median of 12 in the second variable, standing for "5 s", confirm this result (cf. appendix E.12.5), although the estimated time was not less than 3 seconds as we hypothesized.

4.4.2.4.3 Hypothesis Test H17

The Kendall's and Spearman's correlation coefficient tests on the hypothesis H17 that "participants are most likely to admit that their gesture coincides with the origin of sound in Experiment 3 if they mean that it is easy to define the origin of sound and assume that they manage to locate sound" showed that there was a statistically significant positive relationship between admitting that it was easy to define the origin of sound in experiment 3 and that participants managed to locate sound, $r_s = .37, p < .05$ and $\tau = .35, p < .05$, both one-tailed (see appendix E.12.15). In addition, there was also a statistically significant positive relationship between considering that it was easy to define the origin of sound and that the gesture coincided with the origin of sound, $r_s = .47, p < .05$ and $\tau = .44, p < .05$, both one-tailed. Finally, there was once again a statistically significant positive relationship between assuming that participants managed to locate sound and that the gesture coincided with the origin of sound, $r_s = .42, p < .05$ and $\tau = .39, p < .05$, both one-tailed.

4.4.2.5 Experiments 1 and 2 Data Results

We performed non-parametric Wilcoxon signed-rank tests on hypotheses H18 and H19, presented in section 3.6.4.5, in which differences between experiments 1 and 2, where the same participants had been involved, were examined (Field, 2009, p. 552) (see appendix E.12.16).

4.4.2.5.1 Hypothesis Test H18

The Wilcoxon signed-rank test on the hypothesis H18 that "participants feel more immediate control over sound in Experiment 1 than in Experiment 2" showed that there was no statistically significant difference of the immediate control over sound between experiments 1 and 2 ($Z = -1.13, p = .26 > .05$), $T = 30.5, p > .05, r = -.16$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.13}{\sqrt{26 \times 2}} = -.16$ (see appendix E.12.16), although there was a slightly increase of immediate control in experiment 2 (median of 4 (3.75 to 4), standing for "I agree") relatively to experiment 1 (median of 4 (3 to 4), standing for "I agree" as well).

4.4.2.5.2 Hypothesis Test H19

The Wilcoxon signed-rank test on the hypothesis H19 that "participants estimate a lower time needed to control sound in Experiment 1 than in Experiment 2" showed that there was no statistically significant difference in estimating times needed to control sound between experiments 1 and 2 ($Z = -1.38, p = .17 > .05$), $T = 47.5, p > .05, r = -.19$, where $r = \frac{Z}{\sqrt{N \times 2}} = \frac{-1.38}{\sqrt{26 \times 2}} = -.19$ (see appendix E.12.16), although the estimated time needed to control the sound in experiment 1 (median of 17 (7.25 to 23), standing for "10 s") was a little bit higher than in experiment 2 (median of 13 (5 to 23), standing for "5 s").

4.5 Installation's Usefulness in Practical Life

Regarding the last question of the Inquiry Mode Questionnaire (InQ) (see appendix A), the valid results show that only 4.8% of all volunteers answered that our interactive installation would not be useful at all, having

these negative responses been given by people belonging to the group without any musical knowledge, and that 95.2% of all participants (see appendix E.12.2) found that our installation would be useful in practical life, particularly: in hearing tests with the aim of developing awareness of the origin of sound; in tests for auditory and psychoacoustic diagnosis; in detecting hearing problems; in localization exercises; in multi-sensory training; in multidimensional auditory tests; in music therapy; in hearing exercises or educational games; in locating loudspeakers according to our listening needs with a high accuracy; in the development of attention through the perception of the origin of sound sources; in the development of auditory acuity; in the improvement of the quality of life of people with cognitive and motor disabilities; in assisting the guidance of the visually impaired; in training of deaf people when using cochlear implants; in working with people with special educational needs; in the development of auralization systems for various applications; in teaching, as a possible didactic material so that anyone who would not be able to manipulate any musical instrument could compose or create music; in music training classes; in music studies; in the identification of repertoires, timbres, and rhythms; and in learning choreographies, in particular classical dance.

It could be also used: in dance for greater relation with the public; as a gesture-music interaction means in an orchestra, in an opera; to improve and increase the potential of the creative teams involved in Performing Arts; in entertainment business; in video games, in multimedia games, in interactive video; in changing the spatialization of sound and objects in space; in artistic installations, for instance sound installations; in the creation of artistic objects; in musical, theatrical, and other performances; in the interactivity between sound and gesture produced on stage and sound generated in the audience; in science centres, commercial spaces, parties; in passenger compartments, air-plane cockpits; as an excellent means of guidance, for example in public spaces; as a good solution to get sound localized without the use of headphones in immersive multimedia systems; and to easily define the sound surrounding us in a cinema or in a home cinema by choosing what we would like to hear and where.

Furthermore, it could be used as well: to turn on or off different devices in a room; to command or control sound consoles; to control devices effortlessly; to control sound spatialization of a performance with the performer's body; in everyday life, for sound activation from body gestures; and as a means to bring the human being closer to technology in a simple way.

4.6 Discussion and Evaluation

4.6.1 All Participants

Whereas hypothesis H1 was verified as most participants, taking all volunteers into account, considered that they quickly understood the three experiments, hypothesis H2 was not verified because there were significant differences in the estimation of time participants needed to understand each of the three experimental conditions. Here we reinforce the idea that the estimated times would help us gaining insight into each participant's time perception in each experiment and, at the same time, detecting inconsistencies in responses, as we already mentioned in section 4.1. Those estimated times of 30 s, 20 s, and 10 s were relatively long when compared with the meaning of "quickly understood the experiment" and with "less than 2 seconds" we had

hypothesized. In addition, it seems that the decreasing estimated time needed to understand the experiments from the first to the last experiment was due to familiarization with the interactive installation, which we were aware of and had even stimulated, as we explained in section 3.6. In this sense and after analysing the videos, whereas participants triggered 27 sounds on average in experiment 1, they only triggered 19 sounds on average in experiment 2. Hypotheses H10 and H12, which were not verified as well, confirmed these results. Hypothesis H15 was partially verified because, while most participants admitted that they quickly understood the last experiment, the estimated time needed to understand it was short. Nevertheless, it was not less than 2 seconds as we had hypothesized.

Hypothesis H3 was verified and so it confirms that most participants interacted easily with our installation. Furthermore, hypothesis H5 was also verified as most volunteers agreed that the suggested gesture was adequate to any of the three experiments. Hypothesis H6 was firstly partially verified as most participants considered that the system's response to gesture was immediate in any of the three experiments, but the rating tendency in experiment 2 was higher than in experiment 3. However, afterwards we considered it as verified because the inverted questions related to hypothesis H6 confirmed that most volunteers would end up regarding the system's response to gesture equally fast in the three experiments. In addition, hypothesis H7 was verified as well and reinforced this idea, since the estimated time of the system's response to gesture in the three experiments was equal to around 1 s, which is consistent with the system's latency measurement of approximately 470 milliseconds (cf. section 3.7.1). Hypothesis H8 was verified as most participants felt surrounded by sound in our interactive installation using Ambisonics Equivalent Panning (cf. sections 2.3.4.7 and 3.3).

Hypothesis H4 was partially verified as most participants considered that they would define the origin of sound more easily in the first experiment and then in the last one, as we had hypothesized. Since the direction of a deictic gesture (see section 2.6) coincided with the origin of a sound in experiment 1, as we had decided, and most directions of sounds were determined randomly by the software to be laterally incident on a volunteer when he or she faced the camera in experiment 3 (on average, 6 to the left, 7 to the right, and only 2 in the central zone) (see appendix E.8), this partial result points towards being easier to define the origin of sound when its incidence is frontal in the horizontal plane, in addition to auditory perception being more accurate in this position, where the Minimum Audible Angle (MAA) is best, just as sound source localization theory predicts (see sections 2.2.4, 2.5, and 2.5.1). However, most participants also considered that the origin of sound in experiment 2 would be almost as easily to define as in experiment 1, which we had not hypothesized. Looking just at the answers of the volunteers who correctly realized that their gesture did not coincide with the origin of sound in experiment 2 (cf. section 3.6), as we had determined, corresponding to 30,3% of all participants (= 23,3% who answered "I totally disagree" + 7,0% who answered "I disagree" to question 4 related to experiment 2 - see section 4.1), and examining their answers to the question about the easiness of the definition of the origin of sound at the same time (cf. appendix E.12.2), we concluded that only 25.6% of all participants found the definition of the origin of sound in experiment 2 easy as well, although slightly less than in experiment 1. Thus, at least 55.9% of all participants, who answered "I agree" or "I totally agree" to question 4 related to experiment 2, did not perceive that their gesture could not have coincided with the origin of sound in experiment 2. This result, together with the analysis of the videos, seems to demonstrate that when

sounds, such as those we had chosen for the experiments, are produced frontally or from the back on the axis of intersection of the horizontal plane with the median plane and a person has the task of controlling their motion with a deictic gesture at the same time, his or her ability to identify the source of sound is in general further reduced, in addition to the already well-known diminished ability to identify the origin of sound when it is in the median plane if the head is not rotated, as explained in sections 2.5 and 2.5.2, and in addition to the "... relative inefficiency of the human auditory system in processing spatial information..." (Marentakis et al., 2008, p. 1), as already mentioned in section 2.3. Although participants were allowed to freely move their body, legs, arms, and the head while controlling the motion of sound, in the videos it appears that most of them did not use the head rotation or other cues to undo the front-back ambiguity produced by almost identical inter-aural signal differences during this task, because the hand movement was mostly accompanied by a natural congruent head movement as well.

Furthermore, hypothesis H9 was partially verified since most participants considered that their gesture coincided with the origin of sound in the first and then in the third experimental conditions. However, most volunteers did not notice that in experiment 2 their gesture could not have been coincident with the origin of sound, as already mentioned in the previous paragraph. Thus, hypothesis H13 was not verified either because most participants assumed that their gesture coincided with the origin of sound in experiment 2, even considering that it would be easy to define the origin of sound and that the proposed experiment would not confuse them.

Hypothesis H11 was partially verified as most participants admitted that they felt immediate control over sound and at the same time they estimated a relatively short time to control it in experiment 1. Nevertheless, this time of 10 s was longer than 1 second we had hypothesized. According to Pedroso de Lima (2012, p. 287), an auditory stimulus takes around 8 to 10 milliseconds to reach the brain (ASB) and simple reaction times (SRT) for auditory information are of approximately 140 to 160 milliseconds long. Thus, the total time for a participant to hear a sound and react to it in our interactive installation would be of around 647 ms = 470 ms (latency of our whole system) + 7 ms (time that a sound needs to travel from a loudspeaker to the central listening point) + 10 ms (ASB) + 160 ms (SRT), which is why we had hypothesized a time of less than 1 second. The outcome was almost the same in experiment 2, so that the hypothesis H14 was partially verified as well. So, most volunteers meant that they felt immediate control over sound and at the same time they estimated a relatively short time to control it in experiment 2. Nevertheless, this time of 5 s was longer than 1 second we had hypothesized. The fact that most participants felt almost the same immediate control over sound in experiments 1 and 2 caused hypothesis H18 to be not verified. Hypothesis H19 was also not verified since most volunteers did not estimate a lower time needed to control sound in experiment 1 than in experiment 2.

Hypothesis H16 was partially verified because most participants were highly convinced that they had quickly located sound and at the same time they estimated a short time to locate it in experiment 3. However, this time of 5 s was not less than the 3 seconds we had hypothesized experimentally. The analysis of the videos showed that they actually needed 9.7 s on average to localize sounds. Hypothesis 17 was verified as in experiment 3 most volunteers considered that it would be easy to define the origin of sound, that they would manage to locate sound, and that their gesture would coincide with the origin of sound.

Whereas 86.1% of all volunteers agreed or totally agreed that the gesture coincided with the origin of sound

Table 4.18: Percentage of Perceptual Sound Source Localizations - All Participants

Experiment	"Your gesture coincided with the origin of sound"
1	86.1%
3	58.1%

in experiment 1, only 58.1% considered the same in experiment 3 (see table 4.18). The analysis of the videos also showed that most volunteers made 2 attempts on average to localize sounds with their deictic gesture in experiment 3 and that a sound only had to be heard 1 time on average to be localized. These 2 attempts correspond to 50% of correct perceptual sound source localizations, a percentage almost equal to the 58.1% by volunteers about the gesture's coincidence with the origin of sound in this experiment. In addition, the deviations made by most volunteers to the left and to the right of the actual sound directions, when they tried to identify the system-predetermined localizations of perceptual sound sources with their deictic gestures, which are usually performed in a peripheral area of the body (cf. section 2.6), were of 7.97º and -7.19º on average, respectively. Thus, the absolute average deviation was of 7.76º. Comparing these deviations with those made by participants using their left hand (deviations of 6.86º to the left and -6.35º to the right of the actual sound directions) and with those using their right hand (deviations of 8.46º to the left and -7.38º to the right of the actual sound directions), we concluded that the results were quite similar to each other (cf. appendix E.8).

According to Odowichuk (2012, p. 59), "the addition of gestural control to the creation of 3D spatial audio has the potential to improve..." the creation of immersive auditory scenes. In our case, since most participants interacted easily with our installation using Ambisonics Equivalent Panning and they felt surrounded by sound, while controlling sound with their deictic gesture, it appears that sound spatialization could improve their performance in an interactive installation, although the correlation between gesture and localization of sound sources in space is not as perfect as it could possibly be.

4.6.2 By Musical Knowledge

The conclusions drawn for hypotheses H1, H3, H5, H7, H8, H12, H13, H16, H17, H18, and H19, when dividing all participants by musical knowledge into two groups, one with and the other without musical knowledge, were similar to those obtained when taking all volunteers into account (cf. section 4.6.1).

However, whereas hypothesis H2 was not confirmed when taking all participants into account, it was partially confirmed in each group of volunteers, respectively with and without musical knowledge, since there were no statistically significant differences in the estimation of time it took for participants in each group to understand the three experimental conditions, and the estimated time was not less than 2 s, as we had hypothesized. Hypothesis H10 was not confirmed in the group with musical knowledge, but it was partially confirmed in the group of participants without any musical knowledge due to a high rating in quickly understanding experiment 1, although the estimated time was not less than 2 s, as we had hypothesized. Furthermore, hypothesis H15 was not confirmed in each group of participants, for the reason that there was no significant relationship between considering that participants quickly understood experiment 3 and the estimated time needed to understand it.

Hypothesis H6 was confirmed in both groups of volunteers, as most participants considered that the sys-

tem's response to gesture was immediate in any of the three experiments.

Whereas hypothesis H4 was partially confirmed in the group with musical knowledge and also when taking all participants into account, it was not confirmed at all in the group without any musical knowledge, because there were no statistically significant differences in how easily participants without musical knowledge defined the origin of sound in the three experiments.

In the group of participants with musical knowledge, hypothesis H9 was partially confirmed for the same reasons when taking all participants into account (see section 4.6.1). However, in the group without any musical knowledge, hypothesis H9 was not confirmed, since there were no statistically significant differences in how volunteers considered that their gesture coincided with the origin of sound.

Hypothesis H11 was partially confirmed in the group with musical knowledge, just like when taking all volunteers into account. In the group without any musical knowledge, this hypothesis was not confirmed, as there was no statistically significant relationship between admitting that participants felt immediate control over sound in experiment 1 and the respective short estimated time to control sound.

Concerning hypothesis H14, it was partially confirmed in the group with musical knowledge and when taking all volunteers into account as well, but it was not confirmed at all in the group without any musical knowledge, because there was no statistically significant relationship between assuming that participants felt immediate control over sound in experiment 2 and the estimated time needed to control it.

In conclusion, we can state that between the two groups, the differences are only found in the results for hypotheses H4, H9, H10, H11 and H14. When individually compared with the situation in which all participants were involved, the group of participants with musical knowledge differs in only 3 cases (hypotheses H2, H6, and H15), while the group of volunteers without any musical knowledge differs in 8 cases (hypotheses H2, H4, H6, H9, H10, H11, H14, and H15).

This means that there were some hypothetical deviations in the results of the analysed data based on musical knowledge. Whereas participants in the group without any musical knowledge did not show significant differences in how easily they defined the origin of sound in the three experiments, volunteers with musical knowledge did somehow define the origin of sound more easily in the first an then in the third experimental condition. In addition, most participants in the group without any musical knowledge did not demonstrate any significant differences in how they considered that their gesture coincided with the origin of sound, although most of the volunteers in both groups failed to detect that in experiment 2 the gesture did not coincide with the origin of sound. In section 4.2.1, we mentioned that we had hypothesized that participants with musical knowledge, although they might not be, even so, specialized listeners, would be more likely to be focused on listening when compared to those without such knowledge. However, after we had analysed the data (cf. section E.12.1), we found out that this was not exactly the case. Only 31.0% of the participants with musical knowledge were able to perceive that their gesture did not coincide with the origin of sound in experiment 2. 28.6% of the volunteers without musical knowledge were able to do so, too.

Whereas 82.8% of the volunteers with musical knowledge agreed or totally agreed that the gesture coincided with the origin of sound in experiment 1, 92.9% of the participants without any musical knowledge considered the same (see table 4.19). Relatively to experiment 3, 65.5% of the volunteers with musical knowledge agreed or totally agreed that the gesture coincided with the origin of sound and only 42.8% of the participants

Table 4.19: Percentage of Perceptual Sound Source Localizations - By Musical Knowledge

Musical Knowledge	Experiment	"Your gesture coincided with the origin of sound"
Yes	1	82.8%
No	1	92.9%
Yes	3	65.5%
No	3	42.8%

without any musical knowledge considered the same. The analysis of the videos showed that most volunteers either with or without musical knowledge made 2 attempts on average to localize sounds with their deictic gesture in experiment 3 and that a sound only had to be heard 1 time on average to be localized, such as when taking all participants into account. These 2 attempts correspond to 50% of correct perceptual sound source localizations, a percentage quite different from the 65.5% by volunteers with musical knowledge or from the 42.8% by participants without any musical knowledge. Furthermore, the deviations made by participants with musical knowledge to the left and to the right of the actual sound directions were of 8.41º and -7.43º on average, respectively. Thus, the absolute average deviation was of 8.08º. The time they needed to localize sounds with their deictic gestures was of 10 s on average. In turn, the deviations made by participants without any musical knowledge to the left and to the right of the actual sound directions were of 7.06º and -6.70º on average, respectively. Thus, their absolute average deviation was of 7.08º. The time they needed to localize sounds with their deictic gestures was of 9.2 s on average. These outcomes revealed that the overall performance of the volunteers without any musical knowledge was surprisingly a little bit better than that of those with musical knowledge. However, these deviations were very close to each other and to those when all participants were taken into account, which allowed us to state that there were no significant differences between participants with or without musical knowledge in the localization of perceptual sound sources with their deictic gestures (cf. section 4.2.1).

4.7 Chapter Conclusions

In this chapter, we firstly presented our method of analysis of the collected data based on our Inquiry Mode Questionnaire (InQ) (cf. appendix A), followed by the data results and the respective actual analysis. We started with the descriptive statistics concerning demographic information of all 43 participants in our study and then of two groups of volunteers from the sample containing all participants, one with and the other without any musical knowledge, in order to test hypothetical deviations in the results of the analysed data based on musical knowledge. Our sample of participants resulted randomly from a population of volunteers, which was considered to be a self-selection sampling, made up of friends, relatives, students from different academic subjects studying, and colleagues working, at the School of Arts at the Portuguese Catholic University (EA-UCP), at the School of Music and Performing Arts at the Polytechnic Institute of Porto (ESMAE-IPP) and at the Music Academy of Espinho (AME), with or without any musical knowledge, so that outcomes obtained from that sample could be generalisable to this entire population.

We concluded that the decrease of the auditory sensibility with age was not a problem to take into account

in the data analysis, since the majority of the female participants was younger than 45 (under 35 in the group without any musical knowledge and under 55 in the group with musical knowledge) and the vast majority of men was aged between 25 and 34 years or less. A standard audiometric analysis of each participant could have supported this assumption more consistently and could have revealed the hearing handicaps of the volunteers, but we did not have the opportunity to carry it out.

Descriptive and inferential statistics of collected data with regard to the three experiments undertaken in our investigation were presented afterwards. This means that our main hypotheses, found in section 3.6.4, were all tested, firstly taking all volunteers into account and then the two groups of participants. The confidence degree we have chosen, that is, the degree of certainty that the characteristics of our collected data represent the characteristics of the entire population, is of 95%. The confidence interval or margin error, that is, the accuracy required for any estimates made from our sample, has therefore been chosen to be of 5%.

We found out that most participants took much more than the 2 seconds we had experimentally hypothesized to understand each of the three experiments and that this estimated time decreased from the first to the last experiment, apparently due to familiarization with our interactive system, although they considered that they had done it quickly. Furthermore, most volunteers interacted easily with our installation and they agreed that the suggested gesture was adequate to any of the three experiments. We also found out that participants estimated the system's response to gesture as being immediate in the three experiments, that is, of around 1 second, which is consistent with the system's latency measurement of about 470 milliseconds. In addition, most volunteers felt surrounded by sound in our interactive installation using Ambisonics Equivalent Panning.

The outcomes point towards being easier to define the origin of sound when its incidence is frontal in the horizontal plane, in addition to auditory perception being more accurate in this position, just as sound source localization theory predicts (see section 2.5.1). In this sense, we discovered that most volunteers (at least 55.9% of all participants - cf. section 4.6.1) did not perceive that their gesture could not have coincided with the origin of sound in experiment 2, which, together with the analysis of the videos, seems to demonstrate that when sounds, such as those we had chosen for the experiments, are produced frontally or from the back on the axis of intersection of the horizontal plane with the median plane and a person has the task of controlling their motion with a deictic gesture at the same time, his or her ability to identify the source of sound is in general further diminished, in addition to the already well-known reduced ability to identify the origin of sound when it is in the median plane if the head is not rotated. Thus, this experiment did not confuse them. It appears that most of the participants did not use the head rotation or other cues to undo the front-back ambiguity produced, because the hand movement was mostly accompanied by a natural congruent head movement.

Most participants also felt immediate control over sound in experiments 1 and 2, but the time they estimated to control it was much longer than the time of approximately 650 milliseconds needed for a human being to hear a sound and react to it in our interactive installation, as explained in section 4.6.1. We also discovered that the actual average time needed to localize sounds with the use of a deictic gesture in experiment 3 was of about 10 seconds, a time which was much longer than the 3 seconds we had experimentally hypothesized, and that volunteers made 2 attempts on average to localize sounds with their deictic gesture. Furthermore, a sound had to be heard 1 time only on average to be localized. However, when dividing all participants by musical knowledge, we found out that the time needed for participants without any musical knowledge to

localize sounds with their deictic gestures was surprisingly slightly lower than that of the volunteers with musical knowledge, that is, about 9 seconds against 10 seconds on average, respectively.

Whereas 86.1% of all volunteers agreed or totally agreed that the gesture coincided with the origin of sound in experiment 1, only 58.1% considered the same in experiment 3 (see table 4.18 in section 4.6.1). The analysis of the videos also showed that most volunteers made 2 attempts on average to localize sounds with their deictic gesture in experiment 3 and that a sound only had to be heard 1 time on average to be localized. These 2 attempts correspond to 50% of correct perceptual sound source localizations, a percentage almost equal to the 58.1% rated by volunteers about the gesture's coincidence with the origin of sound in this experiment.

When taking two groups of volunteers from the sample containing all participants into account, one with and the other without any musical knowledge, whereas 82.8% of the volunteers with musical knowledge agreed or totally agreed that the gesture coincided with the origin of sound in experiment 1, 92.9% of the participants without any musical knowledge considered the same (cf. table 4.19 in section 4.6.2). Relatively to experiment 3, 65.5% of the volunteers with musical knowledge agreed or totally agreed that the gesture coincided with the origin of sound and only 42.8% of the participants without any musical knowledge considered the same. The analysis of the videos showed that most volunteers either with or without musical knowledge made 2 attempts on average to localize sounds with their deictic gesture in experiment 3 and that a sound only had to be heard 1 time on average to be localized, such as when taking all participants into account. These 2 attempts correspond to 50% of correct perceptual sound source localizations, a percentage quite different from the 65.5% estimated by volunteers with musical knowledge or from the 42.8% estimated by the participants without any musical knowledge.

We discovered, as well, that the deviations made by most volunteers to the left and to the right of the actual sound directions, when they tried to identify the system-predetermined localizations of perceptual sound sources with their deictic gestures, which are usually performed in a peripheral area of the body (cf. section 2.6), were of 7.97º and -7.19º on average, respectively. Thus, the absolute average deviation was of 7.76º. Comparing these deviations with those made by participants using their left hand (deviations of 6.86º to the left and -6.35º to the right of the actual sound directions) and with those of test subjects using their right hand (deviations of 8.46º to the left and -7.38º to the right of the actual sound directions), we concluded that the results were quite similar to each other (cf. appendix E.8). Furthermore, the deviations made by participants with musical knowledge to the left and to the right of the actual sound directions were of 8.41º and -7.43º on average, respectively. Thus, the absolute average deviation was of 8.08º. In turn, the deviations made by participants without any musical knowledge to the left and to the right of the actual sound directions were of 7.06º and -6.70º on average, respectively. Thus, their absolute average deviation was of 7.08º. All values are consequently quite similar to each other.

Therefore, we concluded that, using an interactive installation like ours with a target audience similar to the one we had, there is a relatively high correlation between gesture and localization of sound sources in space, but that it is not as perfect as it could possibly be due to our hearing system's limitations and seemingly to our natural head's movement dependence on gesture. So, it appears that sound spatialization can improve performance in an interactive installation, but in a moderate way.

Finally, the aspect of the usefulness of our interactive installation in practical life was then addressed, giving us some clues for future work.

Chapter 5

Conclusions and Future Work

A review of the literature on sound spatialization / localization and gesture as the main subjects was carried out by us over the last years with the aim of studying their relationship in the field of human-computer interaction (HCI), more specifically in the context of interactive installations. At the time when we began our investigation, we found out that there were many studies approaching these subjects: 1) separately from each other and/or in other contexts, as in Blauert (1997), Pulkki (1999), Pulkki & Karjalainen (2001), Pulkki (2001a), Bates et al. (2007), Hammershøi (2009), McNeill (1992), Coutaz & Crowley (1995), Choi (2000), Cadoz & Wanderley (2000), Nehaniv (2005), Campbell (2005), or Godøy & Leman (2010); 2) from a more technical point of view, as in Harada et al. (1992), Jensenius et al. (2006), Marshall et al. (2006), Schacher (2007), Neukom & Schacher (2008), Zelli (2009), Marshall et al. (2009), Bhuiyan & Picking (2009), or Schumacher & Bresson (2010); 3) from a more artistic point of view, as in Bencina et al. (2008) or Grigoriou & Floros (2010); 4) but very few trying to involve or address both in order to study their relationship from a more perceptual point of view as in Gröhn (2002), de Götzen (2004), or Marentakis et al. (2008).

Therefore, our study addresses a very specific aspect of the field of HCI, which is the correlation between gesture and localization of sound sources in space. Based on hearing and gesture theories, research questions were raised and respective hypotheses were formulated (see section 3.6.4) and tested (cf. chapter 4), in order to complement and support the main hypothesis that there is a significantly high relationship between a deictic gesture (cf. section 2.6) and localization of perceptual sound sources in space (see section 2.5).

5.1 Summary of Contributions

The main outcomes of our study point towards being easier to define the origin of sound when its incidence is frontal in the horizontal plane (cf. section 2.3.2), in addition to auditory perception being more accurate in this position, just as sound source localization theory predicts (see section 2.5.1). In this sense, we found out that 30.3% of all volunteers realized that their gesture could not have coincided with the origin of sound in our experiment 2 (cf. section 3.6) and that at least 55.9% of all participants (cf. section 4.6.1) did not perceive this fact, which, together with the analysis of the videos (see appendix E.8), seems to demonstrate that when sounds, such as those we had chosen for the experiments (cf. section 3.4), are produced frontally

or from the back on the axis of intersection of the horizontal plane with the median plane and a person has the task of controlling their motion with a non tactile deictic gesture (see section 2.6) at the same time, his or her ability to identify the source of sound is in general further diminished, in addition to the already well-known reduced ability to identify the origin of sound when it is in the median plane if the head is not rotated. Thus, this experiment did not confuse them. It appears that most of the participants did not use the head rotation or other cues to undo the front-back ambiguity produced, because the hand movement was mostly accompanied by a natural congruent head movement, a fact that we would like to study further in the future (see section 5.2). In addition, 86.1% of all volunteers considered that their gesture coincided with the origin of sound in experiment 1 and only 58.1% admitted the same in experiment 3 (cf. table 4.18 in section 4.6.1).

Most participants also felt immediate control over sound in our experiments 1 and 2 (see section 3.6), but the times they estimated to control it were much longer than the time of approximately 650 milliseconds needed for a human being to hear a sound and react to it in our interactive installation, as explained in section 4.6.1. When considering all participants, the estimated median times were of 10 seconds in experiment 1 and of 5 seconds in experiment 2. When splitting all volunteers into two groups, one with and the other without any musical knowledge, these times were respectively of 20 and 10 seconds in the group without any musical knowledge, and of 10 and 5 seconds in the group with musical knowledge.

We also discovered that the actual average time needed to localize sounds with the use of a deictic gesture in our experiment 3 was of about 10 seconds, a time which was much longer than the 3 seconds we had experimentally hypothesized, and that volunteers made 2 attempts on average to localize sounds with their deictic gesture. Furthermore, a sound had to be heard 1 time only on average to be localized.

However, when dividing all participants by musical knowledge, we found out that the time needed for participants without any musical knowledge to localize sounds with their deictic gestures was surprisingly slightly lower than that of the volunteers with musical knowledge, that is, about 9 seconds against 10 seconds on average, respectively.

Therefore, participants had more or less the same type of control over sound when it was in front or behind them, but when sound was not aligned with the front to back axis of the head, then they had more difficulties in determining its origin and its localization.

We discovered, as well, that the deviations made by most volunteers to the left and to the right of the actual sound directions, when they tried to identify the system-predetermined localizations of perceptual sound sources with their deictic gestures, which are usually performed in a peripheral area of the body (cf. section 2.6), were of 7.97º and -7.19º on average, respectively. Thus, the absolute average deviation was of 7.76º. Comparing these deviations with those made by participants using their left hand (deviations of 6.86º to the left and -6.35º to the right of the actual sound directions) and with those using their right hand (deviations of 8.46º to the left and -7.38º to the right of the actual sound directions), we concluded that the results were quite similar to each other (cf. appendix E.8). Furthermore, the deviations made by participants with musical knowledge to the left and to the right of the actual sound directions were of 8.41º and -7.43º on average, respectively. Thus, the absolute average deviation was of 8.08º. In turn, the deviations made by participants without any musical knowledge to the left and to the right of the actual sound directions were of 7.06º and -6.70º on average, respectively. Thus, their absolute average deviation was of 7.08º. All values are consequently quite similar to

each other.

We found out that much more than the 2 seconds we had experimentally hypothesized were taken up by most participants to understand each of the three experiments and that this estimated time decreased from the first to the last experiment, apparently due to familiarization with our interactive system, although they considered that they had done it quickly.

Furthermore, most volunteers interacted easily with our installation and they agreed that the suggested gesture was adequate to any of the three experiments.

We also found out that participants estimated the system's response to gesture as being immediate in our three experiments, that is, of around 1 second, which is consistent with the system's latency measurement of about 470 milliseconds (see section 3.7.1).

In addition, most volunteers felt surrounded by sound in our interactive installation using Ambisonics Equivalent Panning (cf. section 2.3.4.7).

Therefore, we have concluded that, using an interactive installation like ours with a target audience similar to the one we had, there is a relatively high correlation between gesture and localization of sound sources in space, but that it is not as perfect as it could possibly be due to our hearing system's limitations and apparently to our natural head's movement dependence on gesture. So, it seems that sound spatialization can improve performance in an interactive installation, but in a moderate way.

5.2 Future Work

The interactive installation we developed has nevertheless some limitations for various reasons. Therefore, we would like to improve it in the future by: 1) using more than one depth camera with better tracking capabilities and resolution than the Microsoft Xbox Kinect we used (and whose cost we had ourselves to bear), in order to obtain more accurate results in a 360º view without any restrictions and to also track the positions and directions of the head and eyes; 2) using faster programming languages like C++[1] or open source toolkits like openFrameworks[2] to reduce the overall latency time due to data processing; 3) using more than one computer with the aim of essentially reducing the overall latency time as well, each one performing a specific function: for instance, one for tracking fingers, hands, head, and eyes, the other for audio processing, and a third one for recording all data with the highest resolution as possible; and 4) using more loudspeakers in positions other than only in the horizontal plane, in order to explore the correlation between gesture and localization of sound sources beyond the horizontal plane. The tracking of the head and eyes would allow us to verify to what extent the head movement is dependent on the deictic gesture and if the eyes control somehow, or have any influence in, the head movement when a deictic gesture is used.

In addition, we would also like to carry out a standard audiometric analysis for each participant, in order to support the assumption that the data of part 2 of our Inquiry Mode Questionnaire (InQ) (see appendix A) are reliable for analysis in a more consistently way (cf. section 4.2).

The interactive system we developed could also be used in other contexts. Based on our own professional

[1] Retrieved 05/11/2017, from `https://en.wikipedia.org/wiki/C%2B%2B`
[2] Retrieved 05/11/2017, from `https://en.wikipedia.org/wiki/OpenFrameworks`

experience and perception of things, it could be realistically used: 1) in dance or theatre, so that a dancer or an actor could manipulate the direction of sound during a performance, according to the gestures performed – the directivity of sound could be used both for the dancer or actor on a stage as for the audience, in a manner similar to what Pierre Schaeffer and Pierre Henry did with the *potentiomètre d'espace* (cf. section 2.3.1 in chapter 2), but in a much more evolved way; 2) in multimedia games, in which the gesture would control the directivity of sound associated with any action; 3) to improve the hearing ability of dancers, musicians, and sound engineers, among other people.

In addition to the researchers' own proposals on this issue, the majority of the participants contributed as well for it in the last question of the InQ. Thus, our installation would be useful in practical life (see section 4.5), particularly: in hearing tests with the aim of developing awareness of the origin of sound; in tests for auditory and psychoacoustic diagnosis; in detecting hearing problems; in localization exercises; in multi-sensory training; in multidimensional auditory tests; in music therapy; in hearing exercises or educational games; in locating loudspeakers according to our listening needs with high accuracy; in the development of attention through the perception of the origin of sound sources; in the development of auditory acuity; in the improvement of the quality of life of people with cognitive and motor disabilities; in assisting the guidance of the visually impaired; in training of deaf people when using cochlear implants; in working with people with special educational needs; in the development of auralization systems for various applications; in teaching, as a possible didactic material so that anyone who would not be able to manipulate any musical instrument could compose or create music; in music training classes; in music studies; in the identification of repertoires, timbres, and rhythms; and in learning choreographies, in particular classical dance.

It could also be used: in dance for greater relation with the public; as a gesture-music interaction means in an orchestra, in an opera; to improve and increase the potential of the creative teams involved in Performing Arts; in entertainment business; in video games, in multimedia games, in interactive video; in changing the spatialization of sound and objects in space; in artistic installations, for instance sound installations; in the creation of artistic objects; in musical, theatrical, and other performances; in interactivity between sound and gesture produced on stage and sound generated in the audience; in science centres, commercial spaces, parties; in passenger compartments, air-plane cockpits; as an excellent means of guidance, for example in public spaces; as a good solution to get sound localized without the use of headphones in immersive multimedia systems; and to easily define the sound surrounding us in a cinema or in a home cinema by choosing what we would like to hear and where.

Furthermore, it could be used as well: to turn on or off different devices in a room (e.g. radio, heater, light); to command or control sound consoles; to control devices effortlessly; to control sound spatialization of a performance with the performer's body; in everyday life, for sound activation from body gestures; and as a means to bring the human being closer to technology in a simple way. Although some of these suggestions have in the meantime been explored or even implemented by engineers (for example control of roller shutters and blinds or home lighting systems), we permit ourselves to mention in passing that before beginning our investigation work we already had intuitions and precise ideas about the relevance of interactive systems. Moreover, these lines presented just at the end of our work had originally been conceived as a guide and desideratum.

Bibliography

Aaronson, N. L., & Hartmann, W. M. (2014). Testing, correcting, and extending the woodworth model for interaural time difference. *Journal of the Acoustical Society of America*, *135*(2), pp. 817–823. Retrieved 11/06/2016, from https://www.pa.msu.edu/acoustics/woodworth.pdf

Almeida, G. A. N. (2013). *Influência da acústica do palco na performance musical: um novo paradigma na simulação de eventos acústicos para auralização interactiva com músicos* (PhD thesis, Escola das Artes - Universidade Católica Portuguesa). Retrieved 22/09/2015, from http://repositorio.ucp.pt/handle/10400.14/15316

Alvarsson, J. (2013). *Perspectives on wanted and unwanted sounds in outdoor environments: Studies of masking, stress recovery, and speech intelligibility* (PhD thesis, Stockholm University). Retrieved 05/11/2015, from http://su.diva-portal.org/smash/get/diva2:661988/FULLTEXT02.pdf

Arau, H. (1999). *Abc de la acustica arquitectonica*. Barcelona, ES: Grupo Editorial Ceac, S.A.

ASA. (1951, July). *American standard acoustical terminology.* Book. American Standards Association. Retrieved 14/01/2016, from https://ia600304.us.archive.org/20/items/ameri00amer/ameri00amer.pdf (revision of Z24.1-1942)

Attenborough, K. (2014). Sound propagation in the atmosphere. In T. D. Rossing (Ed.), *Springer handbook of acoustics* (2nd ed., pp. 117–155). Heidelberg, DE: Springer-Verlag Berlin Heidelberg.

Audio Products Division of National Semiconductor Analog Products Group. (n.d.). *3d sound - what is it?* (Vol. 4). Audio Education Series eBook. Retrieved 23/07/2011, from http://www.national.com/assets/en/other/3DSound.pdf

Austin, J. L. (1962). *How to do things with words*. London, UK: Oxford University Press.

Azevedo, M. (2004). *Teses, relatórios e trabalhos escolares: Sugestões para estruturação da escrita* (4.ª ed.). Lisboa, PT: Universidade Católica Editora.

Ballou, G. M. (1987). *Handbook for sound engineers: The new audio cyclopedia*. Indianapolis, IN, USA: Howard W. Sams & Company.

Bates, E., Kearney, G., Boland, F., & Furlong, D. (2007). Localization accuracy of advanced spatialization techniques in small concert halls. *Journal of the Acoustical Society of America*, *121*(5), pp. 3069–3070. Retrieved 05/06/2015, from http://www.endabates.net/ASA07EndaBates.pdf

Bauer, B. B. (1961). Phasor analysis of some stereophonic phenomena. *Journal of the Acoustical Society of America*, *33*(11), pp. 1536–1539.

Bencina, R., Wilde, D., & Langley, S. (2008). Gesture ≈ sound experiments: Process and mappings. In *Proceedings of the 2008 conference on new interfaces for musical expression (nime08)* (pp. 197–202). NIME. Retrieved 25/02/2015, from http://www.criticalsenses.com/resources/papers/bencinawildelangley-gesturesound-nime2008.pdf

Bennett, J. C., Barker, K., & Edeko, F. O. (1985). A new approach to the assessment of stereophonic sound system performance. *Journal of the Audio Engineering Society*, *33*(5), pp. 314–321. Retrieved 15/04/2016, from http://decoy.iki.fi/dsound/ambisonic/motherlode/source/A%20new%20approach%20to%20the%20assessment%20of%20stereophonic_Bennett%20et%20al_1985.pdf

Berkhout, A. J., de Vries, D., & Vogel, P. (1993). Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, *93*(5), pp. 2764–2778. Retrieved 25/05/2016, from https://www.researchgate.net/publication/229099797_Vogel_D_Acoustic_control_by_wave_field_synthesis

Bhuiyan, M., & Picking, R. (2009, November). Gesture-controlled user interfaces, what have we done and what's next? In *Proceedings of the fifth collaborative research symposium on security, e-learning, internet and networking (sein 2009).* Retrieved 19/05/2014, from http://www.glyndwr.ac.uk/computing/research/pubs/sein_bp.pdf

BIPM. (2006). *The international system of units (si)* (8th ed.). SI Brochure. Organisation Intergouvernementale de la Convention du Mètre. Retrieved 17/11/2015, from http://www.bipm.org/utils/common/pdf/si_brochure_8_en.pdf (updated 2014 by supplement)

Birkner, C. (2004). *Practical recording 5: Surround sound*. London, UK: SMT.

Bisig, D., Neukom, M., & Flury, J. (2007). Interactive swarm orchestra. In *Proceedings of the generative art conference.* Retrieved 16/01/2016, from http://www.generativeart.com/on/cic/papersGA2007/ISO_GA_2007.pdf

Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization*. Cambridge, Massachusetts, USA: The MIT Press.

Blumlein, A. D., Electric, & Musical Industries, L. (1931, December). *U.k. patent 394325: Improvements in and relating to sound-transmission, sound-recording and sound-reproduction systems.* Milestone-Proposal. Retrieved 08/03/2016, from http://ieeemilestones.ethw.org/images/5/5f/GB394325A%28OCR%29.pdf

Boone, Jr., H. N., & Boone, D. A. (2012). Analysing likert data. *Journal of Extension*, *50*(2). Retrieved 20/12/2012, from http://www.joe.org/joe/2012april/tt2p.shtml

Borenstein, G. (2012). *Making things see: 3d vision with kinect, processing, arduino, and makerbot*. Sebastopol, CA, USA: O'Reilly Media, Inc.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, Massachusetts, USA: The MIT Press.

Brüel & Kjaer. (2015). *7. audiology and psychoacoustics.* Sound, Vibration, Education. Retrieved 29/11/2015, from http://www.bksv.com/doc/sve/7.%20Audiology%20(0210).pdf

Brown, A. D., Stecker, G. C., & Tollin, D. J. (2015, February). The precedence effect in sound localization. *Journal of the Association for Research in Otolaryngology*, *16*(1), pp. 1–28. Retrieved 10/07/2016, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310855/pdf/10162_2014_Article_496.pdf

Brown, C. H., & May, B. J. (2010). Comparative mammalian sound localization. In A. N. Popper & R. R. Fay (Eds.), *Sound source localization* (pp. 124–178). New York, USA: Springer Science + Business Media, Inc.

Brown, P. (2008). Fundamentals of audio and acoustics. In G. M. Ballou (Ed.), *Handbook for sound engineers* (4th ed., pp. 21–39). Oxford, UK: Elsevier Inc.

Brown, R. (2010). *Sound: A reader in theatre practice*. Basingstoke, Hampshire, UK: Palgrave Macmillan.

Brunner, S., Maempel, H.-J., & Weinzierl, S. (2007). On the audibility of comb-filter distortions. *AES E-Library - 122nd Convention of the Audio Engineering Society*, pp. 1–7. Retrieved 28/03/2016, from https://www2.ak.tu-berlin.de/~akgroup/ak_pub/2007/Brunner%20Maempel%20Weinzierl%202007_On%20the%20audibility%20of%20comb%20filter%20distortions%20AES.pdf

Brutti, A., Omologo, M., Svaizer, P., & Irst, F. B. K. (2008). Comparison between different sound source localization techniques. In *Hands-free speech communication and microphone arrays* (pp. 69–72). IEEE. Retrieved 25/03/2013, from http://dicit.fbk.eu/publications/brutti_omologo_svaizer_hscma08.pdf

Cadoz, C., & Wanderley, M. M. (2000). Gesture - music. In M. M. Wanderley & M. Battier (Eds.), *Trends in gestural control of music* (pp. 71–94). Paris, FR: IRCAM - Centre Georges Pompidou. Retrieved 26/09/2011, from http://www.idmil.org/_media/wiki/cadoz_wanderley_trends.pdf

Calmes, L. (2013). *Binaural sound source localization - basics.* Retrieved 25/03/2013, from http://www.laurentcalmes.lu/soundloc_basics.html

Campbell, L. (2005, July). *The observation of movement.* Retrieved 26/03/2013, from www.idmil.org/_media/publications/mumt_609_louise_essay93cf3.pdf?id=publications&cache=cache (Unpublished report)

Castellanos, J. (2006). *Design of a framework for adaptive spatial audio rendering* (Master's thesis, University of California). Retrieved 28/07/2011, from http://stomach.v2.nl/Docs/TechPubs/Audio/CastellanosSpatial.pdf

Chamness, M. (1994). *Objective analysis of loudspeaker polar response* (Technical Information White Paper). Eastern Acoustic Works, Inc. Retrieved 23/01/2014, from http://eaw.com/docs/6_Technical_Information/White_Papers/PolarResponseAnalysis.pdf

Cheng, C. I., & Wakefield, G. H. (2001). Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *Journal of the Audio Engineering Society*, *49*(4), pp. 231–249. Retrieved 22/02/2013, from http://505606.pbworks.com/f/HRTF.pdf

Choi, I. (2000). Gestural primitives and the context for computational processing in an interactive performance system. In M. M. Wanderley & M. Battier (Eds.), *Trends in gestural control of music* (pp. 139–172). Paris, FR: IRCAM - Centre Georges Pompidou. Retrieved 26/09/2011, from http://www.music.mcgill.ca/~mwanderley/Trends/Trends_in_Gestural_Control_of_Music/DOS/P.CHOgp.pdf

Choi, I. (2003). A component model of gestural primitive throughput. In *Proceedings of the 2003 conference on new interfaces for musical expression (nime-03)* (pp. NIME03-201–NIME03-204). NIME. Retrieved 26/09/2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.2.1997&rep=rep1&type=pdf

Chowning, J. (1971). The simulation of moving sound sources. *Journal of the Audio Engineering Society*, *19*(1), pp. 2–6. Retrieved 04/08/2015, from https://ccrma.stanford.edu/courses/220a-fall-2001/chowning.pdf

Collins Cobuild English Language Dictionary. (1993). *English language*. London, UK: William Collins Sons & Co Ltd.

Coutaz, J., & Crowley, J. L. (1995). *Interpreting human gesture with computer vision.* Retrieved 08/08/2016, from http://iihm.imag.fr/publication/CC95b/ (Unpublished paper: CHI'95 (Conference on Human Factors in Computing Systems) Workshop on Gesture at the User Interface, Denver, Colorado, USA)

Dahl, S., Bevilacqua, F., Bresin, R., Clayton, M., Leante, L., Poggi, I., & Rasamimanana, N. (2010). Gestures in performance. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 36–68). New York, NY, USA: Routledge.

Daniel, J. (2001). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia* (PhD thesis, Université Paris 6). Retrieved 23/03/2016, from http://gyronymo.free.fr/audio3D/download_Thesis_PwPt.html#Thesis_download

Davis, D., & Davis, C. (1997). *Sound system engineering* (2nd ed.). Oxford, UK: Focal Press.

Davis, G., & Jones, R. (1989). *Sound reinforcement handbook* (2nd ed.). Milwaukee, WI, USA: Hal Leonard Corporation.

Davis, M. F. (2014). Audio and electroacoustics. In T. D. Rossing (Ed.), *Springer handbook of acoustics* (2nd ed., pp. 779–817). Heidelberg, DE: Springer-Verlag Berlin Heidelberg.

de Götzen, A. (2004). The sounding gesture: An overview. In *Proceedings of the 7th international conference on digital audio effects (dafx'04)* (pp. 5–10).

de Vries, D., & Boone, M. M. (1999). Wave field synthesis and analysis using array technology. In *Proceedings 1999 ieee workshop on applications of signal processing to audio and acoustics* (pp. 15–18). IEEE. Retrieved 26/05/2016, from http://www.umiacs.umd.edu/~ramani/cmsc828d_audio/deVries_Boone_WFS_WASPAA_99.pdf

Duda, R. O., Avendano, C., & Algazi, V. R. (1999). An adaptable ellipsoidal head model for the interaural time difference. In *Proceedings of the 1999 ieee international conference on acoustics, speech, and signal processing* (Vol. 2, pp. 965–968). IEEE. Retrieved 11/06/2016, from http://interface.cipic.ucdavis.edu/pubs/Icassp99.pdf

Eargle, J. (2001). *The microphone book*. Boston, Massachusetts, USA: Focal Press.

Eargle, J., & Foreman, C. (2002). *Jbl audio engineering for sound reinforcement*. Milwaukee, WI, USA: Hal Leonard Corporation.

Everest, F. A. (2001). *Master handbook of acoustics* (4th ed.). New York, San Francisco, USA: McGraw-Hill.

Fellgett, P. B. (1974). Ambisonic reproduction of directionality in surround-sound systems. *Nature*, *252*, pp. 534–538. Retrieved 27/03/2016, from http://decoy.iki.fi/dsound/ambisonic/motherlode/source/Ambisonic%20Reproduction%20Of%20Directionality%20in%20Surround_Sound%20Systems%20Fellgett%201974.pdf

Field, A. (2009). *Discovering statistics using spss - third edition* (3rd ed.). London, UK: SAGE Publications Ltd.

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.), *The psychology of music* (pp. 149–180). New York, NY, USA: Academic Press, Inc.

Frank, M. (2013). *Phantom sources using multiple loudspeakers in the horizontal plane* (PhD thesis, University of Music and Performing Arts Graz). Retrieved 23/03/2016, from http://iem.kug.ac.at/fileadmin/media/iem/projects/2013/frank_matthias_diss.pdf

Frank, M., Zotter, F., & Sontacchi, A. (2008). Localization experiments using different 2d ambisonics decoders. In *25th tonmeistertagung - vdt international convention.* Retrieved 19/01/2014, from http://iem.kug.ac.at/fileadmin/media/iem/projects/2008/tmt08.pdf

Françoise, J. (2013). Gesture-sound mapping by demonstration in interactive music systems. In *Proceedings of the 21st acm international conference on multimedia (mm'13)* (pp. 1051–1054). ACM. Retrieved 04/09/2016, from https://hal.archives-ouvertes.fr/hal-01061221/document

Freixo, M. J. V. (2012). *Metodologia científica: Fundamentos, métodos e técnicas* (4th ed.). Lisboa, PT: Instituto Piaget.

Gambetta, C. L. (2005). *Conducting outside the box: Creating a fresh approach to conducting gesture through the principles of laban movement analysis* (PhD thesis, Faculty of The Graduate School at The University of North Carolina at Greensboro). Retrieved 27/07/2016, from http://www.uio.no/studier/emner/hf/imv/MUS2006/v14/litteratur/Gambetta_2005.pdf

Gerzon, M. A. (1970, August). The principles of quadraphonic recording - part one: Are four channels really necessary? *Studio Sound*, *12*, pp. 338–342. Retrieved 13/10/2011, from http://www.michaelgerzonphotos.org.uk/articles/Principles%201.pdf

Gibbs, T. (2007). *The fundamentals of sonic art & sound design*. Lausanne, CH: AVA Publishing SA.

Gibson, D. (1997). *The art of mixing*. Emeryville, CA, USA: MixBooks.

Godøy, R. I. (2010). Gestural affordances of musical sound. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 103–125). New York, NY, USA: Routledge.

Godøy, R. I., & Leman, M. (2010). *Musical gestures: Sound, movement, and meaning*. New York, NY, USA: Routledge.

Gough, C. (2014). Musical acoustics. In T. D. Rossing (Ed.), *Springer handbook of acoustics* (2nd ed., pp. 567–701). Heidelberg, DE: Springer-Verlag Berlin Heidelberg.

Grigoriou, N., & Floros, A. (2010). ambistar: A virtual space representation using 3d human interaction. In *Proceedings of the 3rd international conference on human system interaction (hsi 2010)* (pp. 430–434). Institute of Electrical and Electronics Engineers. Retrieved 18/08/2014, from https://www.academia.edu/1029867/ambiStar_A_virtual_space_representation_using_3D_human_interaction

Gröhn, M. (2002). Localization of a moving virtual sound source in a virtual room, the effect of a distracting auditory stimulus. In *Proceedings of the 2002 international conference on auditory display* (pp. ICAD02-1–ICAD02-9). International Conference on Auditory Display. Retrieved 05/06/2015, from http://www.icad.org/websitev2.0/Conferences/ICAD2002/proceedings/43_MattiGrohn.pdf

Guedes, C. (2005). *Mapping movement to musical rhythm: A study in interactive dance* (PhD thesis). The Steinhardt School of Education, New York University.

Guevara, M. A., & Corsi-Cabrera, M. (1996). Eeg coherence or eeg correlation? *International Journal of Psychophysiology*, pp. 145–153. Retrieved 02/03/2016, from https://www.researchgate.net/publication/14259218_EEG_coherence_or_EEG_correlation

Halmrast, T., Guettler, K., Bader, R., & Godøy, R. I. (2010). Gesture and timbre. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 183–211). New York, NY, USA: Routledge.

Hammershøi, D. (2009). Localization capacity of human listeners. In Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, & H. Kato (Eds.), *Principles and applications of spatial hearing* (pp. 3–13). World Scientific Publishing Company. Retrieved 22/02/2013, from http://www.worldscientific.com/doi/suppl/10.1142/7674/suppl_file/7674_chap01.pdf

Harada, T., Sato, A., Hashimoto, S., & Ohteru, S. (1992). Real time control of 3d sound space by gesture. In *Proceedings of the 1992 international computer music conference* (pp. 85–88). The International Computer Music Association. Retrieved 11/09/2011, from http://quod.lib.umich.edu/cgi/p/pod/dod-idx?c=icmc;idno=bbp2372.1992.023

Harpe, S. E. (2015). How to analyze likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*(6), pp. 836–850. Retrieved 01/05/2017, from https://www.google.pt/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=

0ahUKEwjS9MiykM_TAhWJWRoKHS2nCLsQFggvMAA&url=https%3A%2F%2Fwww.researchgate.net%2Ffile.PostFileLoader.html%3Fid%3D5728c83c93553b59b90ea14f%26assetKey%3DAS%253A357686404239366%2540462290492242&usg=AFQjCNEIawU573SqcHXhyJkl-n4uiVIPbg

Henrique, L. L. (2007). *Acústica musical* (2nd ed.). Lisboa, PT: Fundação Calouste Gulbenkian.

Hines, W. W., Montgomery, D. C., Goldsman, D. M., & Borror, C. M. (2003). *Probability and statistics in engineering* (4th ed.). New York, NY, USA: John Wiley.

Hollerweger, F. (2006). *Periphonic sound spatialization in multi-user virtual environments* (Master's thesis, Institute of Electronic Music and Acoustics (IEM), Graz University of Music and Dramatic Arts; Center for Research in Electronic Art Technology (CREATE), University of California). Retrieved 13/09/2011, from http://flo.mur.at/writings/thesis-hollerweger.pdf

Holman, T. (2000). *5.1 surround sound: Up and running*. Boston, USA: Focal Press.

Howard, D. M., & Angus, J. A. S. (2001). *Acoustics and psychoacoustics* (2nd ed.). Oxford, Auckland, Boston, Johannesburg, Melbourne, New Delhi: Focal Press.

International Telecommunication Union. (2003, December). *General methods for the subjective assessment* (Recommendation ITU-R BS.1284-1). International Telecommunication Union. Retrieved 25/08/2016, from https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1284-1-200312-I!!PDF-E.pdf

International Telecommunication Union. (2012, August). *Multichannel stereophonic sound system with and without accompanying picture* (Recommendation ITU-R BS.775-3). International Telecommunication Union. Retrieved 28/04/2016, from https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.775-3-201208-I!!PDF-E.pdf

International Telecommunication Union. (2015, February). *Methods for the subjective assessment of small impairments in audio systems* (Recommendation ITU-R BS.1116-3). International Telecommunication Union. Retrieved 25/08/2016, from http://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1116-3-201502-I!!PDF-E.pdf

Jensenius, A. R., Kvifte, T., & Godøy, R. I. (2006). Towards a gesture description interchange format. In *Proceedings of the 2006 international conference on new interfaces for musical expression (nime06)* (pp. 176–179). New Interfaces for Musical Expression. Retrieved 26/03/2013, from www.nime.org/proceedings/2006/nime2006_176.pdf

Jensenius, A. R., Wanderley, M. M., Godøy, R. I., & Leman, M. (2010). Musical gestures: Concepts and methods in research. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 12–35). New York, NY, USA: Routledge.

Kane, J. W., & Sternheim, M. M. (1988). *Physics* (3rd ed.). New York, NY, USA: John Wiley & Sons, Inc.

Kendall, G. S. (1995). A 3-d sound primer: Directional hearing and stereo reproduction. *Computer Music Journal*, *19*(4), pp. 23–46. Retrieved 10/06/2016, from http://music.columbia.edu/cmc/courses/g6631/fall2012/page4/files/A%203D%20Sound%20Primer.pdf

Klapuri, A., & Davy, M. (2006). *Signal processing methods for music transcription*. New York, NY, USA: Springer Science+Business Media LLC.

Laureano, R. M. S. (2013). *Testes de hipóteses com o spss - o meu manual de consulta rápida* (2nd ed.). Lisboa, PT: Edições Sílabo, Lda.

Leman, M., & Godøy, R. I. (2010). Why study musical gestures? In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 3–11). New York, NY, USA: Routledge.

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), pp. 1–55. Retrieved 07/01/2017, from http://www.voteview.com/pdf/Likert_1932.pdf

Linkwitz, S. (2015, December). *Sound reproduction.* Retrieved 25/02/2016, from http://www.linkwitzlab.com/reproduction.htm

Litovsky, R. Y., Colburn, H. S., Yost, W. A., & Guzman, S. J. (1999, October). The precedence effect. *Journal of the Acoustical Society of America*, *106*(4), pp. 1633–1654. Retrieved 11/07/2016, from https://www.waisman.wisc.edu/bhl/about_publications/1997LitovskyJAcoustSocAm.pdf

Longstaff, J. S. (2005). Rudolf laban's notation workbook, an historical survey of dance script methods from choreographie (1926). In M. Bastien & R. A. Ploch (Eds.), *Proceedings of the 24th biennial conference of the international council of kinetography laban (ickl)* (pp. 203–238). International Council of Kinetography Laban. Retrieved 03/08/2016, from http://www.laban-analyses.org/jeffrey/2005_laban-choreographie-notation-workbook/Longstaff_2005_proceedings_paper.doc

Lourtie, I. M. G. (2007). *Sinais e sistemas* (2nd ed.). Lisboa, PT: Escolar Editora.

Magee, D. J. (2008). *Orthopedic physical assessment* (5th ed.). St. Louis, Missouri, USA: Saunders, Elsevier Inc.

Malham, D. G. (1998). Approaches to spatialisation. *Organised Sound*, *3*(2), pp. 167–177. Retrieved 29/01/2012, from http://atlas.dxarts.washington.edu/courses/567/current/malham.pdf

Marentakis, G., Malloch, J., Peters, N., Marshall, M., Wanderley, M., & McAdams, S. (2008). Influence of performance gestures on the identification of spatial sound trajectories in a concert hall. In P. Susini & O. Warusfel (Eds.), *Proceedings of the 14th international conference on auditory display (icad2008)* (pp. ICAD08-1–ICAD08-8). IRCAM (Institut de Recherche et Coordination Acoustique/Musique). Retrieved 25/03/2013, from http://marktmarshall.com/publications

Marentakis, G., & McAdams, S. (2013). Perceptual impact of gesture control of spatialization. In *Proceedings of the acm transactions on applied perception* (Vol. 10, pp. 22:1–22:21). Association for Computing Machinery. Retrieved 25/02/2013, from http://www.mcgill.ca/mpcl/files/mpcl/marentakis_2013_acmtap.pdf

Marshall, M. T., Malloch, J., & Wanderley, M. M. (2009). Gesture control of sound spatialization for live musical performance. In M. S. Dias, S. Gibet, M. M. Wanderley, & R. Bastos (Eds.), *Gesture-based human-computer interaction and simulation: 7th international gesture workshop, gw 2007, lisbon, portugal, may 23-25, 2007, revised selected papers* (pp. 227–238). Berlin, DE: Springer-Verlag. Retrieved 12/03/2012, from http://marktmarshall.com/_media/writings/gesturecontrolspatialization.pdf

Marshall, M. T., Peters, N., Jensenius, A. R., Boissinot, J., Wanderley, M. M., & Braasch, J. (2006). On the development of a system for gesture control of spatialization. In *Proceedings of the 2006 international computer music conference* (pp. 360–366). The International Computer Music Association. Retrieved 11/09/2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.186.8627&rep=rep1&type=pdf

McCarthy, B. (2007). *Sound systems: Design and optimization - modern techniques and tools for sound system design and alignment*. Oxford, UK: Elsevier Ltd.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL, USA; London, UK: The University of Chicago Press.

McNeill, D. (2000). *Language and gesture*. Cambridge, UK: Cambridge University Press.

McNeill, D. (2011). Gesture. In P. C. Hogan (Ed.), *The cambridge encyclopedia of the language sciences* (pp. 344–346). New York, NY, USA: Cambridge University Press. Retrieved 12/08/2016, from https://pt.scribd.com/doc/243148394/Cambridge-Encyclopedia-of-Language-Sciences-pdf

Meares, D. J. (1973, May). *Systems of quadraphony: A first assessment* (BBC Engineering: A record of BBC technical experience and developments in radio and television broadcasting No. 94). British Broadcasting Corporation. Retrieved 24/11/2011, from http://downloads.bbc.co.uk/rd/pubs/archive/pdffiles/engineering/bbc_engineering_94.pdf

Miranda, E. R., & Wanderley, M. M. (2006). *New digital musical instruments: Control and interaction beyond the keyboard*. Middleton, WI, USA: A-R Editions, Inc.

Monro, G. (2000). In-phase corrections for ambisonics. In *Proceedings of the 2000 international computer music conference* (Vol. 2000, pp. 1–4). The International Computer Music Association. Retrieved 17/05/2016, from http://quod.lib.umich.edu/cache//b/b/p/bbp2372.2000.194/bbp2372.2000.194.pdf#page=1;zoom=75

Moore, B. C. J. (2013). *An introduction to the psychology of hearing* (6th ed.). Leiden, NL: Koninklijke Brill NV.

Mulder, A. (1996, February). *Hand gestures for hci* (Hand Centered Studies of Human Movement Project No. 96-1). School of Kinesiology, Simon Fraser University. Retrieved 08/08/2016, from http://www.xspasm.com/x/sfu/vmi/HCIgestures.htm

Mustard, J. (2006). *The integrated sound, space and movement environment: The uses of analogue and digital technologies to correlate topographical and gestural movement with*

sound (Master's thesis, Faculty of Communications and Creative Industries, Edith Cowan University). Retrieved 27/07/2011, from http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1084&context=theses&sei-redir=1#search=%22integrated%20sound%2C%20space%22

Nachbar, C., Zotter, F., Deleflie, E., & Sontacchi, A. (2011). Ambix - a suggested ambisonics format. In *Proceedings of the ambisonics symposium 2011* (pp. 1–11). IEM Ambisonics Group. Retrieved 11/05/2016, from http://iem.kug.ac.at/fileadmin/media/iem/projects/2011/ambisonics11_nachbar_zotter_sontacchi_deleflie.pdf

Nehaniv, C. L. (2005). Classifying types of gesture and inferring intent. In *Proceedings of the 2005 artificial intelligence and simulation of behaviour symposium on robot companions: Hard problems and open challenges in robot-human interaction.* The Society for the Study of Artificial Intelligence and Simulation of Behaviour. Retrieved 25/09/2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.9295&rep=rep1&type=pdf

Nehaniv, C. L., Dautenhahn, K., Kubacki, J., Haegele, M., Parlitz, C., & Alami, R. (2005). A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication*, *2005*, pp. 371–377. Retrieved 10/08/2016, from https://uhra.herts.ac.uk/bitstream/handle/2299/606/101146.pdf?sequence=1

Neukom, M., & Schacher, J. C. (2008). Ambisonics equivalent panning. In *Proceedings of the 2008 international computer music conference.* The International Computer Music Association. Retrieved 07/02/2012, from http://www.jasch.ch/pub/ICMC08_AEP_paper.pdf

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Springer Science+Business Media B. V.*. Retrieved 27/01/2017, from http://blogs.helsinki.fi/kvantiblogi/files/2013/09/SampleSizeNorman2010.pdf

Odowichuk, G. (2012). *Free-space gesture mappings for music and sound* (Master's thesis, University of Victoria). Retrieved 08/12/2013, from http://dspace.library.uvic.ca:8080/bitstream/handle/1828/4288/Odowichuk_Gabrielle_MASc_2012.pdf?Sequence=4

Ostergaard, P. B. (2003). Physics of sound and vibration. In E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, & M. Layne (Eds.), *The noise manual* (5th ed., pp. 19–39). Fairfax, VA, USA: American Industrial Hygiene Association.

Pareyon, G. (2011). *On musical self-similarity: Intersemiosis as synecdoche and analogy*. Helsinki, FI: The International Semiotics Institute at Imatra (ISI).

Pauk, S. (2006). *Use of long-term average spectrum for automatic speaker recognition* (Master's thesis, University of Joensuu). Retrieved 14/12/2015, from ftp://ftp.cs.joensuu.fi/pub/Theses/2006_MSc_Pauk_Sergey.pdf

Pedersen, T. H., & Zacharov, N. (2008). How many psycho-acoustic attributes are needed? In *Proceedings acoustics'08 paris* (pp. 1215–1220). Société Française d'Acoustique. Retrieved 16/11/2015, from http://webistem.com/acoustics2008/acoustics2008/cd1/data/articles/002939.pdf

Pedroso de Lima, J. J. (2012). *Ouvido, ondas e vibrações: Aspectos físicos e biofísicos*. Coimbra, PT: Imprensa da Universidade de Coimbra.

Penha, R., & Oliveira, J. P. (2013). Spatium, tools for sound spatialization. In *Proceedings of the sound and music computing conference 2013 (smc 2013)* (pp. 660–667). Sound and Music Computing). Retrieved 16/01/2014, from http://smcnetwork.org/system/files/SPATIUM%2C%20TOOLS%20FOR%20SOUND%20SPATIALIZATION.pdf

Pestana, M. H., & Gageiro, J. N. (2014). *Análise de dados para ciências sociais - a complementaridade do spss* (6th ed.). Lisboa, PT: Edições Sílabo, Lda.

Pierce, A. D. (2014). Basic linear acoustics. In T. D. Rossing (Ed.), *Springer handbook of acoustics* (2nd ed., pp. 29–115). Heidelberg, DE: Springer-Verlag Berlin Heidelberg.

Pinto de Lima, J. (1983). *Linguagem e acção - da filosofia analítica à linguística pragmática*. Lisboa, PT: Apáginastantas - Cooperativa de Serviços Culturais, crl.

Plack, C. J., & Oxenham, A. J. (2005). Overview: The present and future of pitch. In C. J. Plack, A. J. Oxenham, R. R. Fay, & A. N. Popper (Eds.), *Pitch: Neural coding and perception* (pp. 1–6). New York, NY, USA: Springer Science+Business Media, Inc.

Power, P., Dunn, C., Davies, B., & Hirst, J. (2013). *Localisation of elevated sources in higher-order ambisonics* (Research & Development White Paper No. WHP 261). British Broadcasting Corporation. Retrieved 19/01/2014, from http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP261.pdf

Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, *45*(6), pp. 456–466. Retrieved 22/05/2016, from http://lib.tkk.fi/Diss/2001/isbn9512255324/article1.pdf

Pulkki, V. (1999). Uniform spreading of amplitude panned virtual sources. In *Proceedings 1999 ieee workshop on applications of signal processing to audio and acoustics* (pp. W99-1–W99-4). IEEE. Retrieved 29/01/2012, from http://www.acoustics.hut.fi/~ville/papers/waspap.pdf

Pulkki, V. (2001a). Localization of amplitude-panned virtual sources ii: Two- and three-dimensional panning. *Journal of Audio Engineering Society*, *49*(9), pp. 753–767. Retrieved 29/01/2012, from ftp://ftp.ircam.fr/private/salles/markus/bt2010/references/periphonic/2001_Pulkki_VBAPLocalization2_JAES.pdf

Pulkki, V. (2001b). *Spatial sound generation and perception by amplitude panning techniques* (PhD thesis, Helsinki University of Technology). Retrieved 29/01/2012, from http://lib.tkk.fi/Diss/2001/isbn9512255324/isbn9512255324.pdf

Pulkki, V., & Karjalainen, M. (2001). Localization of amplitude-panned virtual sources i: Stereophonic panning. *Journal of Audio Engineering Society*, *49*(9), pp. 739–752. Retrieved 29/01/2012, from http://lib.tkk.fi/Diss/2001/isbn9512255324/article6.pdf

Reas, C., & Fry, B. (2007). *Processing: a programming handbook for visual designers and artists*. Cambridge, Massachusetts, USA: The MIT Press.

Reas, C., & Fry, B. (2010). *Getting started with processing*. Sebastopol, CA, USA: O'Reilly Media, Inc.

Reis, E. (2012). *Estatística descritiva* (7th ed.). Lisboa, PT: Edições Sílabo, Lda.

Roads, C., Strawn, J., Abbott, C., Gordon, J., & Greenspun, P. (1996). *The computer music tutorial*. Cambridge, Massachusetts, USA: Massachusetts Institute of Technology.

Rodrigues, I. M. G. (2007). *O corpo e a fala - comunicação verbal e não-verbal na interacção face a face*. Lisboa, PT: Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia.

Rossing, T. D. (1990). *The science of sound* (2nd ed.). Reading, Massachusetts, USA: Addison-Wesley Publishing Company.

Rovan, J., & Hayward, V. (2000). Typology of tactile sounds and their synthesis in gesture-driven computer music performance. In M. M. Wanderley & M. Battier (Eds.), *Trends in gestural control of music* (pp. 355–368). Paris, FR: IRCAM - Centre Georges Pompidou. Retrieved 09/05/2014, from http://www.music.mcgill.ca/~mwanderley/Trends/Trends_in_Gestural_Control_of_Music/DOS/RovHay.pdf

Rumsey, F. (2001). *Spatial audio*. Oxford, UK: Focal Press.

Rumsey, F. (2008). Signal processing for 3-d audio. *Journal of Audio Engineering Society*, *56*(7/8), pp. 640–645. Retrieved 18/03/2010, from http://www.aes.org/tmpFiles/tutorials/20100317/aesTutorial93.pdf

Rumsey, F., & McCormick, T. (1997). *Sound and recording - an introduction* (3rd ed.). Oxford, UK: Focal Press.

Saberi, K., Dostal, L., Sadralodabai, T., & Perrott, D. R. (1991). Minimum audible angles for horizontal, vertical, and oblique orientations: Lateral and dorsal planes. *Acustica*, *75*, pp. 57–61. Retrieved 05/06/2015, from http://www.auditorylab.com/saberietal1991acustica.pdf

Salema, C., & Ferreira, A. (2009). Som, luz e cor. In F. Pereira (Ed.), *Comunicações audiovisuais: Tecnologias, normas e aplicações* (pp. 29–74). Lisboa, PT: Instituto Superior Técnico.

Schacher, J. C. (2007). Gesture control of sounds in 3d space. In *Proceedings of the 7th international conference on new interfaces for musical expression (nime07)* (pp. 358–362). ACM. Retrieved 04/04/2013, from http://www.jasch.ch/pub/NIME07_Gesture_Jasch.pdf

Schneider, A. (2010). Music and gestures: A historical introduction and survey of earlier research. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 69–100). New York, NY, USA: Routledge.

Schumacher, M., & Bresson, J. (2010). Compositional control of periphonic sound spatialization. In *Proceedings of the 2nd international symposium on ambisonics and spherical acoustics.* IRCAM (Institut de Recherche et Coordination Acoustique/Musique). Retrieved 22/02/2013, from http://mcgill.academia.edu/MarlonSchumacher/Papers

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review*, *89*(4), pp. 305–333.

Smith, S. W. (1997). *The scientist and engineer's guide to digital signal processing*. San Diego, California, USA: California Technical Publishing.

Solmer, A. (1999). *Manual de teatro*. Lisboa, PT: Cadernos ContraCena.

Sonnenschein, D. (2001). *Sound design: The expressive power of music, voice, and sound effects in cinema*. Studio City, CA, USA: Michael Wiese Productions.

Spiegel, M. R. (2000). *Estatística*. Lisboa, PT: McGraw-Hill.

Stark, S. H. (2002). *Live sound reinforcement: A comprehensive guide to p.a. and music reinforcement systems and technology*. Vallejo, CA, USA: Artistpro.com, LLC.

Stern, R. M., Brown, G. J., & Wang, D. (2005). Binaural sound localization. In D. Wang & G. J. Brown (Eds.), *Computational auditory scene analysis* (pp. 1–34). John Wiley & Sons, Inc. Retrieved 22/02/2013, from http://www.cs.cmu.edu/~robust/Papers/SternWangBrownChapter.pdf

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), pp. 677–680. Retrieved 03/05/2017, from http://marces.org/EDMS623/Stevens%20SS%20(1946)%20On%20the%20Theory%20of%20Scales%20of%20Measurement.pdf

Stitt, P., Bertet, S., & van Walstijn, M. (2013). Perceptual investigation of image placement with ambisonics for non-centred listeners. In *Proceedings of the 16th international conference on digital audio effects (dafx'13)* (pp. DAFx-1–DAFx-7). Digital Audio Effects. Retrieved 19/01/2014, from http://dafx13.nuim.ie/papers/61.dafx2013_submission_32.pdf

Streicher, R., & Everest, F. A. (1998). *The new stereo soundbook* (2nd ed.). Pasadena, CA, USA: Audio Engineering Associates.

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, *5*(4), pp. 541–542. Retrieved 27/01/2017, from http://www.jgme.org/doi/pdf/10.4300/JGME-5-4-18

Sundberg, J. (1991). *The science of musical sounds*. San Diego, CA, USA: Academic Press, Inc.

Sutil, N. S. (2013). Rudolf laban and topological movement: A videographic analysis. *Space and Culture: International Journal of Social Spaces*, *16*(2), pp. 173–193. Retrieved 03/08/2016, from http://thirdworld.nl/order/fcaa9a1d994140d3ecc278d5fb682cf33fc785c0

Tanaka, K., Parker, J., Baradoy, G., Sheehan, D., Holash, J. R., & Katz, L. (2012). A comparison of exergaming interfaces for use in rehabilitation programs and research. *Loading... The Journal of the Canadian Game Studies Association*, *6*(9), pp. 69–81. Retrieved 28/12/2012, from http://journals.sfu.ca/loading/index.php/loading/article/download/107/118

Theile, G., & Wittek, H. (2004). Wave field synthesis: A promising spatial audio rendering concept. *Journal of the Acoustical Society of Japan*, *25*(6), pp. 393–399. Retrieved 26/05/2016, from http://www.umiacs.umd.edu/~ramani/cmsc828d_audio/WFS_Review_2004_Theile_Wittek_ASJ.pdf

Thigpen, B. (2009). Spatialization without panning. *The Electronic Journal of Electroacoustics*. Retrieved 30/09/2011, from http://cec.sonus.ca/econtact/11_4/thigpen_spatialization.html

Tipler, P. A. (1999). *Physics for scientists and engineers* (4th ed.). New York, NY, USA: W. H. Freeman and Company/Worth Publishers, Inc.

Toole, F. E. (2008). *Sound reproduction: Loudspeakers and rooms*. Oxford, UK: Focal Press.

Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. In *Proceedings of the national academy of sciences of the united states of america* (pp. 1–6). National Academy of Sciences of the United States of America. Retrieved 02/09/2013, from http://www.pnas.org/content/early/2013/08/16/1221454110.full.pdf+html

University of Salford. (2015, December). *Defining sound quality.* Retrieved 27/12/2015, from http://www.salford.ac.uk/computing-science-engineering/research/acoustics/psychoacoustics/sound-quality-making-products-sound-better/sound-quality-testing/defining-sound-quality

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 ieee computer society conference on computer vision and pattern recognition* (pp. I-511–I-518). The Institute of Electrical and Electronics Engineers, Inc. Retrieved 25/04/2012, from http://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf

Wanderley, M. M., & Battier, M. (2000). Resources. In M. M. Wanderley & M. Battier (Eds.), *Trends in gestural control of music* (pp. 736–763). Paris, FR: IRCAM - Centre Georges Pompidou. Retrieved 26/09/2011, from http://www.music.mcgill.ca/~mwanderley/Trends/Trends_in_Gestural_Control_of_Music/

Wanderley, M. M., & Depalle, P. (2004). Gestural control of sound synthesis. In *Proceedings of the ieee* (Vol. 92, pp. 632–644). IEEE. Retrieved 28/08/2016, from http://www.music.mcgill.ca/~mwanderley/MUMT-615/Papers/Class02/P.WanDep.pdf

Warren, J. D., Uppenkamp, S., Patterson, R. D., & Griffiths, T. D. (2003). Separating pitch chroma and pitch height in the human brain. In M. E. Raichle (Ed.), *Proceedings of the national academy of sciences of the united states of america* (Vol. 100, pp. 10038–10042). National Academy of Sciences. Retrieved 12/01/2016, from http://www.pnas.org/content/100/17/10038.full.pdf

Wikipedia. (2014, May). *Sound localization.* Retrieved 09/05/2014, from http://en.wikipedia.org/wiki/Sound_localization

Wikipedia. (2015a, November). *Colors of noise.* Retrieved 14/11/2015, from https://en.wikipedia.org/wiki/Colors_of_noise

Wikipedia. (2015b, November). *Inverse-square law.* Retrieved 30/11/2015, from https://en.wikipedia.org/wiki/Inverse-square_law

Wikipedia. (2015c, December). *Psychophysics.* Retrieved 20/12/2015, from https://en.wikipedia.org/wiki/Psychophysics

Wikipedia. (2015d, November). *Speed of sound.* Retrieved 18/11/2015, from https://en.wikipedia.org/wiki/Speed_of_sound

Wikipedia. (2015e, October). *Standard anatomical position.* Retrieved 18/11/2015, from https://en.wikipedia.org/wiki/Standard_anatomical_position

Wikipedia. (2016a, January). *Christian doppler.* Retrieved 25/01/2016, from https://en.wikipedia.org/wiki/Christian_Doppler

Wikipedia. (2016b, March). *Headphones.* Retrieved 01/03/2016, from https://en.wikipedia.org/wiki/Headphones

Wikipedia. (2016c, December). *Quadrant (plane geometry).* Retrieved 03/12/2016, from https://en.wikipedia.org/wiki/Quadrant_(plane_geometry)

Wikipedia. (2016d, October). *Variable (computer science).* Retrieved 31/10/2016, from https://en.wikipedia.org/wiki/Variable_(computer_science)

Wikipedia. (2017a, March). *Parametric statistics.* Retrieved 02/03/2017, from https://en.wikipedia.org/wiki/Parametric_statistics

Wikipedia. (2017b, January). *Spss.* Retrieved 29/01/2017, from https://en.wikipedia.org/wiki/SPSS

William A. Kuperman, P. R. (2014). Underwater acoustics. In T. D. Rossing (Ed.), *Springer handbook of acoustics* (2nd ed., pp. 157–212). Heidelberg, DE: Springer-Verlag Berlin Heidelberg.

Winters, R. M., & Wanderley, M. M. (2012). New directions for sonification of expressive movement in music. In J. F. Michael A. Nees Bruce N. Walker (Ed.), *Proceedings of the 18th international conference on auditory display (icad2012), atlanta, ga, usa* (pp. 227–228). The International Community for Auditory Display. Retrieved 26/03/2013, from http://www.idmil.org/_media/publications/winters_icad2012.pdf?id=publications&cache=cache

Xiong, X. (2013, March). *How does the brain locate sound sources?* Retrieved 25/03/2015, from http://knowingneurons.com/2013/03/15/how-does-the-brain-locate-sound-sources

Yoshioka, K. (2005). *Linguistic and gestural introduction and tracking of referents in l1 and l2 discourse* (PhD thesis, University of Groningen). Retrieved 08/08/2016, from http://www.rug.nl/research/portal/files/2931162/thesis.pdf

Zelli, B. (2009). Space and computer music: A survey of methods, systems and musical implications. *The Electronic Journal of Electroacoustics*. Retrieved 30/09/2011, from http://cec.sonus.ca/econtact/11_4/zelli_space.html

Zvonar, R. (2005a). An extremely brief history of spatial music in the 20th century. *The Electronic Journal of Electroacoustics*. Retrieved 22/05/2014, from http://cec.sonus.ca/econtact/7_4/zvonar_spatialmusic-short.html

Zvonar, R. (2005b). A history of spatial music: Historical antecedents from renaissance antiphony to strings in the wings. *The Electronic Journal of Electroacoustics*. Retrieved 22/05/2014, from http://cec.sonus.ca/econtact/7_4/zvonar_spatialmusic.html

Appendix A

Inquiry Mode Questionnaire (InQ)

UNIVERSIDADE CATÓLICA PORTUGUESA PORTO CAMPUS FOZ · CITAR · programa de doutoramento em CIÊNCIA e TECNOLOGIA das ARTES · ARTE INTERACTIVA

INQUÉRITO POR QUESTIONÁRIO

Este inquérito tem como objectivo recolher dados para a realização de um trabalho de investigação, no domínio de Ciência e Tecnologia das Artes - Arte Interactiva, na Escola das Artes da Universidade Católica Portuguesa, no Porto. Pretende-se, assim, estudar a correlação entre gesto e a localização de fontes sonoras no espaço.

Constituem a população-alvo pessoas com conhecimentos musicais e pessoas com poucos ou nenhuns conhecimentos musicais.

Os dados recolhidos são anónimos e serão objecto de tratamento estatístico.

PARTE 1 – Informação Demográfica

Assinale com um **X**, sempre que possível.

1. **Idade**
   De 15 a 24 ☐ De 25 a 34 ☐ De 35 a 44 ☐
   De 45 a 54 ☐ De 55 a 64 ☐ 65 ou mais ☐

2. **Sexo**
   Feminino ☐ Masculino ☐

3. **Com que mão escreve?**
   Esquerda ☐ Direita ☐

4. **Habilitações literárias**
   12º Ano ☐ Bacharelato ☐ Licenciatura ☐ Pós-Graduação ☐
   Mestrado ☐ Doutoramento ☐ Outras ☐ _____

5. **Tem conhecimentos musicais?**
   Sim ☐ Se sim, quais? _____
   Não ☐

6. **Tem dificuldades auditivas?**
   Sim ☐ Se sim, de que tipo? _____
   Não ☐

PARTE 2 – Dados do Questionário

Dê a sua opinião, marcando apenas um **X** em cada linha.

1 – Discordo completamente **2** – Discordo **3** – Nem sempre / Por vezes **4** – Concordo **5** – Concordo totalmente

EXERCÍCIO 1 – Definir a origem do som com o gesto

	1	2	3	4	5
Entendeu rapidamente o exercício					
Foi fácil interagir com a instalação					
Foi fácil definir a origem do som					
O gesto sugerido é adequado ao exercício					
Sentiu controlo imediato sobre o som					
A resposta do sistema ao gesto foi imediata					
A resposta do sistema ao gesto foi lenta					
Sentiu-se envolvida(o) pelo som na instalação					
O seu gesto coincidiu com a origem do som					

Faça uma estimativa do tempo:

- que precisou para entender o exercício: _____
- que precisou para controlar o som: _____
- de resposta do sistema ao gesto: _____

Figure A.1: Inquiry Mode Questionnaire (InQ) in Portuguese - Page 1.

1 – Discordo completamente **2** – Discordo **3** – Nem sempre / Por vezes **4** – Concordo **5** – Concordo totalmente

EXERCÍCIO 2 – Definir a origem do som com o gesto, mas o som provém da direcção oposta

	1	2	3	4	5
Entendeu rapidamente o exercício					
Foi fácil interagir com a instalação					
O exercício proposto confundiu-me					
Foi fácil definir a origem do som					
O gesto sugerido é adequado ao exercício					
Sentiu controlo imediato sobre o som					
A resposta do sistema ao gesto foi imediata					
A resposta do sistema ao gesto foi lenta					
Sentiu-se envolvida(o) pelo som na instalação					
O seu gesto coincidiu com a origem do som					

Faça uma estimativa do tempo:

- que precisou para entender o exercício: _____
- que precisou para controlar o som: _____
- de resposta do sistema ao gesto: _____

EXERCÍCIO 3 – Software define a origem do som e utilizador tenta identificá-la através do gesto

	1	2	3	4	5
Entendeu rapidamente o exercício					
Foi fácil interagir com a instalação					
Foi fácil definir a origem do som					
O gesto sugerido é adequado ao exercício					
O seu gesto coincidiu com a origem do som					
A resposta do sistema ao gesto foi imediata					
A resposta do sistema ao gesto foi lenta					
Sentiu-se envolvida(o) pelo som na instalação					
Conseguiu localizar o som					
Localizou rapidamente o som					

Faça uma estimativa do tempo:

- que precisou para entender o exercício: _____
- de resposta do sistema ao gesto: _____
- que precisou para localizar o som: _____

Vê alguma utilidade desta instalação na vida prática?

Sim ☐ Se sim, qual ou quais? _____

_____

Não ☐

Terminou o preenchimento deste inquérito.

Obrigado pela sua colaboração!

Figure A.2: Inquiry Mode Questionnaire in Portuguese - Page 2.

INQUIRY MODE QUESTIONNAIRE (InQ)

This survey aims to collect data for the accomplishment of a research work in the field of Science and Technology of the Arts - Interactive Art, in the School of Arts of the Portuguese Catholic University, in Porto. The aim is to study the correlation between gesture and localization of sound sources in space.

The target population is people with musical knowledge and people with little or no musical knowledge.

The collected data are anonymous and will be subject to statistical treatment.

PART 1 – Demographic Information

Mark with an X, whenever possible.

1. Age

15 to 24 ☐ 25 to 34 ☐ 35 to 44 ☐
45 to 54 ☐ 55 to 64 ☐ 65 or over ☐

2. Gender

Female ☐ Male ☐

3. Which hand do you use for writing?

Left ☐ Right ☐

4. Educational qualifications

12th grade ☐ Bachelor's degree ☐ Graduation ☐ Postgraduate studies ☐
Master's degree ☐ PhD ☐ Other ☐ _____
_____

5. Do you have any musical knowledge?

Yes ☐ If "Yes", which ones? _____
No ☐

6. Do you have hearing handicaps?

Yes ☐ If "Yes", what type? _____
No ☐

PART 2 – Questionnaire Data

Give your opinion by scoring only one X on each line.

1 – I totally disagree **2** – I disagree **3** – Not always / Sometimes **4** – I agree **5** – I totally agree

EXPERIMENT 1 – Define the origin of sound with gesture

	1	2	3	4	5
You quickly understood the experiment					
It was easy to interact with the installation					
It was easy to define the origin of sound					
The suggested gesture is adequate to the experiment					
You felt immediate control over sound					
The system's response to gesture was immediate					
The system's response to gesture was slow					
You felt surrounded by sound in the installation					
Your gesture coincided with the origin of sound					

Estimate the time:

- needed to understand the experiment: _____
- needed to control sound: _____
- of the system's response to gesture: _____

Figure A.3: Inquiry Mode Questionnaire (InQ) translated into English - Page 1.

1 – I totally disagree **2** – I disagree **3** – Not always / Sometimes **4** – I agree **5** – I totally agree

EXPERIMENT 2 – Define the origin of sound with gesture, but the sound comes from the opposite direction

	1	2	3	4	5
You quickly understood the experiment					
It was easy to interact with the installation					
The proposed experiment confused me					
It was easy to define the origin of sound					
The suggested gesture is adequate to the experiment					
You felt immediate control over sound					
The system's response to gesture was immediate					
The system's response to gesture was slow					
You felt surrounded by sound in the installation					
Your gesture coincided with the origin of sound					

Estimate the time:

- needed to understand the experiment: _____
- needed to control sound: _____
- of the system's response to gesture: _____

EXPERIMENT 3 – Software defines the origin of the sound and user attempts to identify it by gesture

	1	2	3	4	5
You quickly understood the experiment					
It was easy to interact with the installation					
It was easy to define the origin of sound					
The suggested gesture is adequate to the experiment					
Your gesture coincided with the origin of sound					
The system's response to gesture was immediate					
The system's response to gesture was slow					
You felt surrounded by sound in the installation					
You managed to locate sound					
You quickly located sound					

Estimate the time:

- needed to understand the experiment: _____
- of the system's response to gesture: _____
- needed to locate sound: _____

Do you see any usefulness of this installation in practical life?

Yes ☐ If "Yes", which one or which ones? _____

_____

No ☐

You have completed this survey.

Thank you for your collaboration!

Figure A.4: Inquiry Mode Questionnaire translated into English - Page 2.

Appendix B

Email Text Sent to Invite as Many People as Possible to Participate in the Practical Experiments

In Portuguese:

”Boa noite, cara(o) colega.

Como é do teu conhecimento, encontro-me a fazer o doutoramento em Ciência e Tecnologia das Artes - Arte Interactiva, na Escola das Artes da Universidade Católica Portuguesa, no Porto.

O objectivo principal do trabalho de investigação proposto é estudar a correlação entre gesto e localização de fontes sonoras no espaço, de maneira a clarificar “o papel que a espacialização sonora desempenha na melhoria da performance numa instalação interactiva” (título da tese).

Assim, convido-te a colaborar num exercício prático com a duração aproximada de 10 minutos, a realizar nas instalações da Escola das Artes (Rua de Diogo Botelho, 1327, na Foz), a partir da tarde do dia 24 de Setembro de 2012 até à manhã do dia 28 de Setembro de 2012.

O exercício prático será seguido do preenchimento de um inquérito por questionário (a duração do preenchimento é de aproximadamente 1 minuto). Os dados são anónimos e serão posteriormente objecto de tratamento estatístico.

Se estiveres disposta(o) a participar neste estudo, envia-me a tua disponibilidade (dias e horas possíveis) o mais brevemente possível, por favor, para que possa efectuar a calendarização da melhor forma.

Para este estudo necessito de um mínimo de 40 voluntários.

Obrigado e até breve,
Diogo Leichsenring Franco”

Translation into English:

”Good evening, dear fellow.

As you know, I am attending the PhD course in Science and Technology of the Arts - Interactive Art, at the School of Arts of the Portuguese Catholic University, in Porto.

The main objective of the proposed research is to study the correlation between gesture and localization of sound sources in space, in order to clarify ”the role that sound spatialization plays in improving performance in an interactive installation” (thesis title).

Therefore, I invite you to collaborate in a practical exercise with a duration of approximately 10 minutes, to be held at the premises of the School of Arts (Rua de Diogo Botelho, 1327, in Foz), from the afternoon of September 24, 2012 until the morning of September 28, 2012.

The practical exercise will be followed by the completion of an inquiry mode questionnaire (the duration of completion is approximately 1 minute long). The data are anonymous and will subsequently be subject to statistical treatment.

If you are willing to participate in this study, please send me your availability (possible days and times) as soon as possible, so that I can make the schedule in the best way.

For this study I need a minimum of 40 volunteers.

Thank you and see you soon,
Diogo Leichsenring Franco”

Appendix C

Timetable of Participants in the Practical Experiments

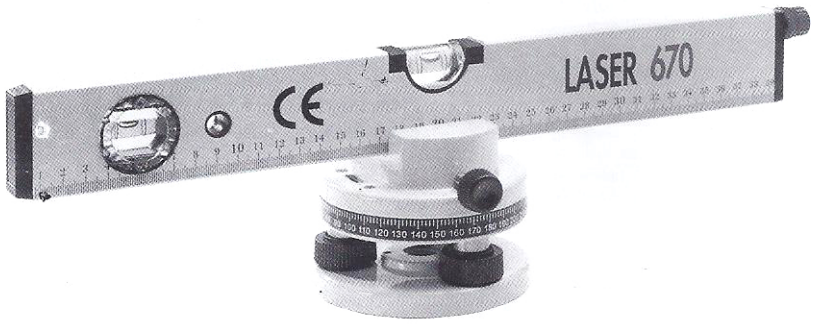
Horas	2ª-feira (24-09-2012)	3ª-feira (25-09-2012)	4ª-feira (26-09-2012)	5ª-feira (27-09-2012)	6ª-feira (28-09-2012)	Sábado (29-09-2012)
09h00-09h15						
09h15-09h30						
09h30-09h45						
09h45-10h00						
10h00-10h15			15	22		
10h15-10h30						
10h30-10h45						
10h45-11h00						
11h00-11h15			16			
11h15-11h30						
11h30-11h45						40
11h45-12h00						
12h00-12h15					32	41
12h15-12h30					33	
12h30-12h45				23		42
12h45-13h00						
13h00-13h15						
13h15-13h30						
13h30-13h45				24	34	
13h45-14h00						
14h00-14h15				25		
14h15-14h30		6	17			
14h30-14h45		7	18			
14h45-15h00	1	8	19			
15h00-15h15	2			26		
15h15-15h30	3			27		
15h30-15h45	4				35	
15h45-16h00				28	36	
16h00-16h15						
16h15-16h30						
16h30-16h45						
16h45-17h00	5	9				
17h00-17h15		10			37	
17h15-17h30						
17h30-17h45		11		29		
17h45-18h00		12				
18h00-18h15			20			
18h15-18h30		13		30		
18h30-18h45				31		
18h45-19h00						
19h00-19h15						
19h15-19h30						
19h30-19h45			21			
19h45-20h00		14				43
20h00-20h15					38	
20h15-20h30					39	
20h30-20h45						
20h45-21h00						
21h00-21h15						
21h15-21h30						
21h30-21h45						
21h45-22h00						
22h00-22h15						
22h15-22h30						
22h30-22h45						
22h45-23h00						
23h00-23h15						
23h15-23h30						
23h30-23h45						
23h45-00h00						

Figure C.1: Timetable of participants in the practical experiments.

Appendix D

Equipment Specifications

PARKSIDE



Laser-type Spirit Level

Assembly and adjusting instructions

Scope of supply	Technical data	
Laser-type spirit level with 3 functions	Laser class 2	DIN EN 60825-1:1997
- pointed beam	P£	1 mW
- laser line function	length of level	40 cm
- 90° angle option	measuring accuracy	± 0.5 mm/m
Elevator tripod	measuring range	≈ 50 m
Levelling base	height of tripod	47.5 - 120 cm
2 batteries	laser	≤1 mW λ = 630-680nm
	wavelength	670 nm
	power supply	2 x 1.5 Micro (AAA)

Figure D.1: Parkside Laser-type Spirit Level 670 specifications.

Appendix E

DVD-ROM

E.1 Ambisonics Equivalent Panning in EXCEL

E.2 Selected Sounds for the Research

E.3 Processing Experiments Code

E.4 MAX MSP Experiments Code

E.5 Processing Data Readers

E.6 Processing and MAX MSP Complete Code

E.7 Full Screen Recorded Information

E.8 Video Analysis

E.9 Audio Latency Measurement - Loudspeaker to Central Hearing Point

E.10 Whole System Latency Measurement

E.11 Filled Inquiry Mode Questionnaires (InQ)

E.12 Calculations in SPSS

E.12.1 Raw Data in SPSS

E.12.2 Raw Frequency Tables and Bar Charts - All Variables - All Participants

E.12.3 Raw Statistics - All Variables - All Participants

E.12.4 Raw Frequency Tables and Bar Charts - All Variables - Participants by Musical Knowledge

E.12.5 Raw Statistics - All Variables - Participants by Musical Knowledge

E.12.6 SPSS Normality Tests - All Participants

E.12.7 Friedman's ANOVA - Experiments 1, 2, and 3 - Hypotheses H1 to H9 - All Participants

E.12.8 Kendall's and Spearman's Correlation Coefficient Tests - Experiment 1 - Hypotheses H10 and H11 - All Participants

E.12.9 Kendall's and Spearman's Correlation Coefficient Tests - Experiment 2 - Hypotheses H12 to H14 - All Participants

E.12.10 Kendall's and Spearman's Correlation Coefficient Tests - Experiment 3 - Hypotheses H15 to H17 - All Participants

E.12.11 Wilcoxon Signed-Rank Tests - Experiments 1 and 2 - Hypotheses H18 and H19 - All Participants

E.12.12 Friedman's ANOVA - Experiments 1, 2, and 3 - Hypotheses H1 to H9 - By Musical Knowledge

E.12.13 Kendall's and Spearman's Correlation Coefficient Tests - Experiment 1 - Hypotheses H10 and H11 - By Musical Knowledge

E.12.14 Kendall's and Spearman's Correlation Coefficient Tests - Experiment 2 - Hypotheses H12 to H14 - By Musical Knowledge

E.12.15 Kendall's and Spearman's Correlation Coefficient Tests - Experiment 3 - Hypotheses H15 to H17 - By Musical Knowledge

E.12.16 Wilcoxon Signed-Rank Tests - Experiments 1 and 2 - Hypotheses H18 and H19 - By Musical Knowledge