3D audio technologies: applications to sound capture, post-production and listener perception

Giulio Cengarle

Tesi Doctoral UPF / 2012

Directors de la tesi
Prof Vicente López
Dr Toni Mateos
Department of Information and Communication Technologies

Dipòsit Legal:
ISBN:

Acknowledgments

This section might very well be the longest of the thesis if full justice had to be done to all the people I mention here, so I will just try to condensate my big gratitude in small sentences.

The first big thanks goes to my co-supervisor Toni Mateos, for his nice attitude, extended support, fresh ideas and high efficiency well beyond the Ph.D. supervision! I am equally grateful to my co-supervisor Prof. Vicente López for the opportunity to carry out my research at Barcelona Media and to continue the adventure in *imm sound*!

My research has been supported by a FIE grant from the government of Catalonia. Part of the work has been carried out within the European projects 20203D-Media (FP7-ICT-2007-1) and iMP (FP7-ICT-2007-3).

I am proud of having shared the last four years with the ex colleagues at the audio group of Barcelona Media, in particular Jordi Arques, Daniel Arteaga, Pau Arumí, Monica Caballero, Jaume Durany, David Garcia, Adan Garriga, Dirk Griffioen, Natanael Olaiz, Toni Mateos, Tim Schmele, Carlos Spa and John Usher; I am grateful for their support and for creating a smooth and highly pleasant working environment. A big thanks as well goes to all the staff at Barcelona Media.

My doctoral research has benefited from the help of the following people: Domenico Stanzial, Davide bonsi and Diego Gonzalez from the acoustics lab of Fondazione Scuola di San Giorgio, Venice, for introducing me to audio research and providing a solid background in the field; Oriol Guasch and Pere Artis from La Salle University for the availability of the anechoic and reverberation chambers for measurements; Davide Bonsi again, for the Microflown probe and the anechoic chamber; Francesc Jarque and the people at Mediapro for the logistic support in the football recording project; Alex Pereda and Miguel Barreda from the Cognition and Perception group at Barcelona Media, without whom I wouldn't have figured out how to do user tests, and especially for their availability as victims of the tests (subjects, in their jargon); all the people who participated in the psychoacoustic tests; finally, all the researchers whose contributions have made 3D audio possible.

A special thanks goes to all the colleagues and the management staff at *imm sound* for making a really nice working environment and for giving me the possibility and support to conclude my Ph.D. research.

I would like to thank the secretariat staff at Universitat Pompeu Fabra for their kindness and efficiency.

As always, I am deeply grateful to my family for their presence and support; since I don't find many opportunities to publicly acknowledge them, I shall not miss this one!

Finally, to Ping, with whom I happily shared these years!

Abstract

The recent advent of 3D audio to the market is dictating changes in several stages of the audio production work-flow, from recording systems and microphone configurations, to post-production methodologies and workstations, to playback loudspeaker configurations. This thesis tackles aspects related to 3D audio arising in the various stages of production. In the recording part, we present a study on the accuracy of tetrahedral microphones from the point of view of three-dimensional sound intensity, and a solution for obtaining second-order Ambisonics responses from first-order transducers using a small number of sensors; in the production stage, we introduce an application for automated assisted mixing of sport events, to reduce the complexity of managing multiple audio channels in real time; a clipping detector is proposed for the rendering of layout-independent audio content to generic playback systems, where the signal levels sent to the speakers are unknown until the decoding stage; finally, psychoacoustic experiments are presented for the validation of perceptual and aesthetic aspects related to 3D audio.

Resumen

La reciente llegada del sonido envolvente 3D en el mercado del audio está imponiendo cambios en varias etapas del flujo de trabajo, desde los sistemas de captación y las técnicas microfónicas, hasta las metodologías de postproducción y las Workstation, y por supuesto las configuraciones de altavoces. Esta tesis trata aspectos relacionados con el audio 3D en las varias fases de la producción: en la parte de captación, presentamos un estudio sobre las características de los micrófonos tetraédricos desde el punto de vista de la intensidad sonora, y una solución para obtener las componentes Ambisonics del segundo orden usando un pequeño número de transductores del primer orden; con respecto a la parte de producción, se presenta una aplicación para la mezcla asistida e automatizada de eventos deportivos, para reducir la complexidad de gestión del multicanal en tiempo real; para la restitución de contenido audio mezclado de manera independiente del sistema de altavoces, en el que los niveles de salida a los altavoces son una incógnita hasta el momento de la decodificación, se propone un detector de clipping independiente del layout. Finalmente, se presentan test psico-acústicos para validar aspectos estéticos y perceptivos relacionados con el audio 3D.

Preface

Multichannel audio appeared in the second half of the twentieth century and developed initially within the realm of electro-acoustic music. At the time, the limitations of the technology in both processing and playback stages led to a very slow development and a scarce adoption of surround sound techniques until the seventies, when four-channel surround sound landed in the industry: the success of the movie Star Wars launched surround sound in cinema, while the same concept of four channels matrixed into a stereo signal was applied in the quadraphonic playback systems for the music consumer market. The quadraphonic systems did not take off seriously, and for many years the advances in surround sound technology were available to audiences only through the cinema industry, which saw the establishment of formats such as the 5.1 surround and 7.1 surround.

5.1 surround appeared in 1992 for the 35mm movie format and later became the standard for home theater releases on DVD, the first support with multichannel audio to reach a vast audience of consumers. In the last decade, many new formats have been proposed for the reproduction of surround sound in the theatrical and domestic environments, including systems for partial or complete 3D playback such as the 10.2 or 22.2. Despite the standardization intents, these proposals are still bound to the research labs, although the principles upon which they are based have been around for many years, as will be briefly described in Chapter 2. However, in the last few years the world of surround audio has undergone a rapid evolution pushed largely by the necessity of the industry to provide newer technology and experiences, in particular to accompany the establishment of 3D video; curiously, the revolution is once again taking place in the movie industry, where horizontal surround sound first landed and settled down almost forty years ago. This is due to the industry's demand for the latest and most innovative technologies, for keeping the edge over the increasing capabilities of home and mobile entertainment systems. The process of implementing 3D audio technologies is nowadays made possible by the vast amount of computing power available in off-the-shelf processors and DSPs at a relatively low cost, which allows to implement techniques that were so far restricted to research and laboratory environments. Thanks to these conditions, audiences can now enjoy 3D audio in a variety of contexts such as cinema, exhibitions

and clubs.

3D surround audio has been studied for a few decades, and its principles are well established from a physical point of view. Various proposals have been made for the recoding and playback, but their implementation is still an open field of research, where different target applications may require different solutions. In fact, the step from a theoretically complete technology and the market is not a smooth and quick transition, but rather a slow process of adaptation, which often requires solving practical problems previously unforeseen. The main obstacle to the success of 3D audio technologies is that hardware solutions and artistic content are strictly interdependent, a bond that can give rise to a "chicken and egg" problem, where nobody is going to produce content for non-existing installations while at the same time nobody is going to invest in hardware without a wide availability of content. Now these problems seem to have been surmounted within the industry and the first 3D surround systems are emerging in the market; therefore, a vast field has appeared for experimentation and research on topics such as: optimum recording techniques and loudspeaker configurations, post-production tools and user interfaces, impact on the listeners and, last but not least, the definition and application of a new aesthetic language.

Perhaps the most revolutionary aspect that came along with 3D audio is the shift of paradigm from a channel-based work-flow, where the content is produced depending on the foreseen playback configuration, to a channel-free one, which is independent from the loudspeaker layout used for playback; one such example is the object-based work-flow, where audio is produced in terms of levels and spatial position of each sound source, leaving to a decoder the task of distributing the signals to the loudspeakers. In fact, the presence of many proposals for a playback standard, or rather the lack of a *de facto* one, make the channel-based concept itself a dangerous bet for future-proof productions.

In this scenario, this thesis developed from circumstances which found the author working within a group of people dedicated to research and development of a 3D audio work-flow, and being involved in different aspects related to spatial audio, mainly regarding cinema productions and sport events. Over the last few years, this has led to tackling different aspect of the production chain, all of which are related by the common denominator of channel-agnostic 3D audio.

List of publications

This thesis is based on the following papers:

- G. Cengarle, T. Mateos and D. Bonsi: *A Second-Order Ambisonics Device Using Velocity Transducers*, Journal of the Audio Engineering Society, vol. 59, n. 9, pg. 656-668, 2011.
- G. Cengarle, T. Mateos and D. Bonsi: *Experimental comparison between Soundfield B-format microphone and Microflown pressure-velocity sound intensity probe* (in italian), presented at the 36^{th} national convention of the Italian Acoustics Association (AIA), Turin, IT, 2009.
- G. Cengarle, T. Mateos: *Comparison of Anemometric Probe and Tetrahedral Microphones for Sound Intensity Measurements*, presented at the 130^{th} Convention of the Audio Engineering Society, London, UK, 2011.
- G. Cengarle, T. Mateos, N. Olaiz and P. Arumí: *A New Technology for the Assisted Mixing of Sport Events: Application to Live Football Broadcasting*, presented at the 128^{th} Convention of the Audio Engineering Society, London, UK, 2010.
- G. Cengarle, T. Mateos: *A clipping detector for layout-independent multichannel audio production*, peer-reviewed paper presented at the 132^{th} Convention of the Audio Engineering Society, Budapest, HU, 2012.
- G. Cengarle, A. Pereda and T. Mateos: *Perceptual aspects of 3D audio*. Currently in preparation for submission (2012).

Contents

List of Figures

List of Tables

1 Introduction and scope

1.1 Introduction

Our perception of sound is three-dimensional: with two ears, we are able to identify the location of sound sources around us and to distinguish sounds coming from left and right, from the front and the back, from the top and the bottom. In our daily activities, we are constantly exposed to three-dimensional sound fields, whether they are the reverberation of a room, the background noise of the city or the natural sounds of the environment. It seems therefore logical that a faithful sound recording and reproduction system should capture and deliver a three-dimensional experience, for the sake of realism. Life-like accuracy aside, audio recording and playback are mostly employed for entertainment, to capture events in the most aesthetically adequate way, and to recreate compelling experiences for the music listeners or the movie goers; in this respect, 3D audio certainly adds a great potential for both getting closer to a "real life" experience and for providing extended creative possibilities for sound engineers.

Bringing 3D sound to the audience requires a re-thinking of the whole production chain as we know it, from sound recording to playback, to adapt it to a new format. To begin with, capturing sound in 3D requires in general highly directional microphones, to separate the sound of sources located in different directions, in those occasions where recording each source separately (in space or in time) is not practical or possible. After the sound is recorded, there comes the post-production stage, where the signals are processed, mixed and prepared for their final presentation; here, dealing with 3D content requires the adaptation of the tools and methods available to sound engineers. The main concerns use to be the increasing number of channels and the difficulty in managing and automating the position of sources in the classical way; the main breakthrough in this regard is the adoption of a channel-free paradigm, where the focus on output channels is replaced by a layout-independent approach, and each audio source or recording track is treated only in terms of its spatialization characteristics, without being tied to the particular playback format that is used. Channel-free approaches are agnostic of the playback configuration; perhaps the most common example is object-based audio, where individual audio tracks have associated position

metadata (usually azimuth and elevation) which are used to locate the sound in the proper position once the playback layout is defined; other examples of channel-free techniques include first- and higher-order Ambisonics, where the spatial information is encoded independently from the output channels, and generic up-mixing algorithms where ambience channels are extracted from a set of channels and can be treated as spatially independent sound sources. With a channel-free paradigm, sound engineers that want to locate a source on the ceiling do not have to worry about which channel to route the audio to: they just specify the desired angular position, adjust the level to taste, and let the decoder route the audio to the proper channels. For this purpose, user interfaces for the post-production also need to adapt to the new paradigm, to simplify a process that would otherwise be too difficult to manage.

Given its artistic potential and capabilities, 3D audio is really a new language: now that more channels are available and the scene extends to at least the upper hemisphere, one has to answer the questions of how to fill the space and which techniques will provide desired or undesired effects. For this reason, the technological evolution must be accompanied by the research on the aesthetic and perceptual implications and impact of the new format.

This thesis addresses topics belonging to the following parts of the work-flow: recording, post-production and perception. A few aspects have been tackled in each part, as detailed in the following sections.

1.2 Contributions

This thesis presents the research done in the three aforementioned parts of the work-flow. The core motivation was tackling some problems that have arisen with the jump from 2D to 3D audio. In the recording part, the main goal was to employ novel anemometric probes to capture higher-order components of the sound field, since their features in terms of size and polar pattern accuracy are very promising. Our work aimed to proposing a simplified method for second-order Ambisonics recording. The main contribution here is the implementation of an analytic approach, based on the Euler equation, for deriving the second-order Ambisonics terms of the acoustic field from a finite difference of first-order signals directly obtained with velocity transducers. Our proposal exploits the fact that a generic microphone of order n can be expressed by sums and differences of signals of order $n-1$. In particular, second-order signals can be obtained by differences of spaced first-order pressure gradient microphones. The Euler equation of fluid dynamics allows to express the pressure gradient in terms of the acoustic velocity, implying that it should be possible to measure the second-order components from combinations of velocity transducers; the result is the design of a second-order three-dimensional microphone which employs nine velocity transducers [Cengarle et al. (2011)]. In order to assess the feasi-

bility of the aforementioned device before implementing it, it was necessary to characterize and compare the accuracy of some existing transducers and techniques for sound recording: extensive work was done in characterizing anemometric transducers and tetrahedral microphones in terms of objective comparative measurements based on sound intensity indicators [Cengarle and Mateos (2011)].

Work done in the production and post-production relates to two different fields: live broadcasting and cinema post-production. Given the current trend of bandwidth availability and the efforts that are being put in bringing new technologies to the broadcasting market, in the author's opinion it is inevitable to foresee the advent of 3D audio broadcasting in the very near future. In order for this to happen, the first requirement is to simplify the sound capture process to allow sound engineers to focus on the three-dimensional soundscape and the use of 3D microphones. In the case of sport events, for example, the largest efforts are spent in capturing the sound of the action, leaving little or no time to focus on the ambient sound. Therefore, one of the first aids to live productions and broadcasting is automating the capture of the point of interest. Our contribution to broadcasting is an algorithm for the assisted mixing of sound events in situations where multiple microphones are located in a wide area to capture the sound from a moving point of interest. Our research tackles in particular the case of football, where a sound engineer traditionally follows the action by manually raising the faders of the microphones that are close to it and lowering the others. The proposed algorithm controls the levels of the microphones given the position of the area of interest: this allows to change the approach to sports mixing from manually controlling the faders of the console to just moving a point on a touch interface and letting a device control the console [Cengarle et al. (2010)]. Since the proposed method only requires the position of the point of interest to optimize its sound capture, the whole process can be slaved and linked to image tracking and be completely independent from the operator. Besides, once the sound of the action is obtained as a monaural signal, it can be easily spatialized, either manually or automatically, according to the camera field of view.

The contribution to cinema post-production is a method for the inference of output levels and the detection of clipping in soundtracks that are produced independently of the playback configuration [Cengarle and Mateos (2012)]. In the scenario of layout-independent audio production, the same soundtrack can be decoded to different loudspeaker layouts and, according to the position of the speakers, the sources of the soundtrack are summed (decoded) in a different way in each layout. Sound engineers are typically monitoring the mix on a reference layout, but when decoding to a different one the level changes may lead to undesired clipping in some speakers. This problem has never been considered previously, since audio productions were always monitored on the intended playback layout; besides, with channel-based audio, the output channels are not re-rendered if the position of the

speakers deviates from the specified layout: for example, in 5.1 the surround content is not changed according to the angle of the speakers in each particular room. In a channel-free approach however, where the output signals are rendered based on the actual position of the speakers, even when decoding the same soundtrack to two almost similar loudspeaker layouts, with similar densities and channel number and a slightly different positioning, level differences of up to 4 dB have been found in some loudspeaker channels; things get even worse when radically different layouts are considered. The proposed solution to the problem is based on the definition of a "worst case" layout, in such a way that absence of clipping in this layout would imply safety with any other layout that meets some particular criteria. The choice of the worst case layout can vary for different kind of content, but in general is related to the minimum required loudspeaker density for a given application. Once the worst case layout is defined, an algorithm decodes the soundtrack to it and reports clipping issues.

Regarding the playback and exhibition, our research has focused on the perception of 3D sound in comparison with stereo and horizontal surround. The research is motivated by the following questions:

- Can 3D audio enhance the emotional involvement of spectators?
- Does 3D audio bring any clear audible advantage in terms of spatial separation of sound sources?
- Can 3D reverberation be pushed to more realistic levels without affecting so much the intelligibility of the program material as it happens with stereo and horizontal surround?

In order to answer the first question, an experiment was designed which consisted in collecting and analyzing psycho-physiological data such as facial electromyography and electro-dermal activity for users watching video content with 3D audio or 5.1 surround audio, as well as collecting responses to questionnaires about audio quality and immersion. The results indicate an increase in emotional arousal and valence provoked by 3D audio, as shown by the measured data; on the other hand, responses to questionnaires did not indicate any conscious perceptual difference between the two formats. The second question arose from conversations with many sound engineers, who reported that mixing in 3D gives a lot of room for placing sound sources in different locations and making them coexist without competing with each other for audibility. These comments reflect almost exactly the opinions of many sound engineers back in the days of the transition from stereo to 5.1, where many of them remarked the benefits of having a much broader panorama where locating the sounds and spread the reverberation. For this reason, another experiment was carried out to study the effect of masking between sound sources in the same critical bands in relationship to the audio format: starting from the empirical evidence that a three-dimensional distribution of sound sources and reverberation in a mix calls for higher density

of soundtracks and higher levels of reverberation before intelligibility is significantly reduced, a study was done to assess how the spatial distribution of a sound source affects the intelligibility of a second, softer sound source. The experiment consisted in the measurement of the masking thresholds for 3D audio, 5.1 and stereo by means of two-alternative forced choice tests and the fitting of the data with a psychometric function to calculate the threshold. The results indicate that the broader the distribution of the sound masker, the higher the intelligibility of the masked sound, given equal sound intensity conditions. These results apply to both audio professionals and unexperienced users. Regarding the intensity of reverberation, it is often found that in recording and playback a lower reverberation level is preferred comparing to a live listening situation: too much reverberation in playback often causes a loss of clarity and results in an unpleasant presentation of the program material. A possible explanation for this is the fact that in real life reverberation is spread in 3D, while in standard playback it is shrunk to a small frontal angle (in the case of stereo) or at most the horizontal plane. In order to evaluate this effect, a third experiment, based on the method of adjustment, was carried out to study the perceived level of reverberation in the three aforementioned audio formats: users were asked to match a reference level of reverberation by adjusting a fader controlling the reverberation in each of the given formats. The results show a subtle tendency to increase the level of the reverberation with the increasing number of channels, although the difference between formats is in the order of 1 dB. This result indicates that the reverberation is perceived roughly with the same intensity regardless of the audio format. Combined with the previous results, this seems to indicate that 3D audio can be exploited to create soundtracks with higher density of sound sources and higher perceptual levels of reverberation before they negatively affect intelligibility. These results are currently being prepared for publication.

Most of the research problems considered here are previously unreported, and a few of them have actually arisen during the practical application or development of 3D audio technologies for music and cinema production. Being new problems, part of the research was devoted to study how to formalize them, put them into a general context and finding strategies to tackle them in a feasible but comprehensive way.

1.3 Outline of the thesis

This thesis begins with an overview of the state of the art in 3D surround technologies: Chapter 2 presents the most advanced and widespread technologies available for spatial sound, together with an analysis of their pros and cons depending on the application. The work presented in later chapters is strictly related to some of these technologies, so the overview, although far from being exhaustive, serves to puts the reader in the context.

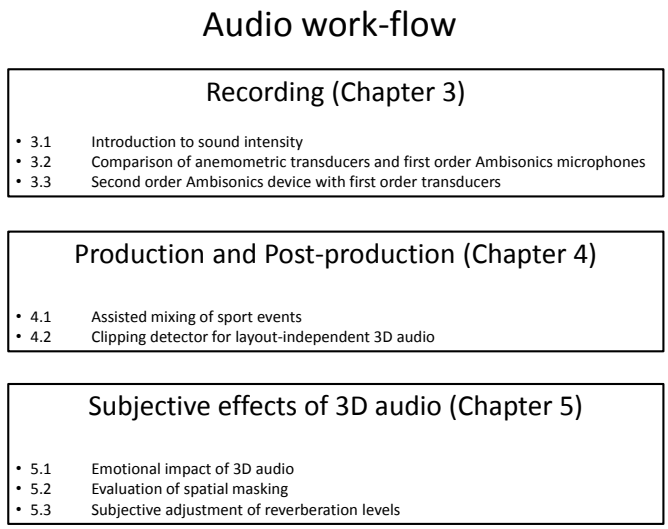

Figure 1.1: Structure of the body of the thesis.

The main body of the thesis consists of three chapters dealing each one with a specific stage of the work-flow, as schematically illustrated in Figure 1.1. Chapter 3 presents the work related to transducers and technology for the capture of 3D surround sound. The chapter begins with an introduction to the concepts of sound intensity that are used in later sections. After the description of the transducers topologies used in this work, we present comparisons between different types of sensors and characterize existing devices in terms of their spatial accuracy. The main section of the chapter presents our study on second-order Ambisonics devices from first-order transducers; here we show how the Euler equation allows to use first-order velocity transducers to derive the second-order spatial properties of acoustic fields: we first derive the equations that relate the second-order derivatives of the pressure field to gradients of the velocity field; then follows a discussion on the frequency response of the higher-order components caused by the differentiation, and the equalization that has to be applied to compensate it; then, these equations are used to obtain a finite-difference approximation for the second-order quantities; finally, we present measurements performed on a prototype device and discuss the performance supporting the experimental data with simulations.

Chapter 4 deals with two aspects related to production and post- production. The first is the algorithm and interface for assisted mixing of sport events where a fixed microphone configuration is used to capture the sound of moving sources. This part includes the details of the algorithm and the

description of the tests that were carried out. The second part of the chapter introduces the problem of unknown signal levels at the playback stage when the production is independent from the loudspeaker layout: the proposal for inferring the occurrence of clipping is based on the definition of a minimum acceptable layout; in this part we describe the algorithm and show examples of its application.

Chapter 5 explores the link between the technical and perceptual aspects of 3D sound. It is divided into three parts, each one presenting an experiment to evaluate a specific perceptual aspect of 3D audio: the first part includes the psychophysical measurements that correlate to the sensation of emotional arousal and indicate an increase in the emotional response to 3D audio compared to standard surround; the second part reports threshold measurements for the discrimination of masking threshold for 3D content compared with 5.1 and stereo. These psychoacoustic tests are based on measurements of the psychometric function, and the results evidence that a 3D distribution of sounds favors intelligibility and reduces masking. In the last part we study how the perceptual level of reverberation is affected by the audio format: a simple experiment, based on the method of adjustment, reveals that the perceived level has a very small variation with the audio format.

Some final considerations and future perspectives are given in Chapter 6.

2 State of the art in 3D audio technologies

In this chapter we present a brief overview of the state of the art in 3D surround sound. The technologies reviewed here span from complete frameworks that account for the whole chain from capture to playback, such as Ambisonics and Wavefield Synthesis, to extensions of existing 2D approaches, like amplitude panning, to a brief mention of hybrid systems and solutions that have been recently introduced to the market.

This chapter is not meant to be a complete and detailed description of the technologies, but just to introduce their most relevant aspects and give the reader a basic knowledge of the subject, providing a context for the topics that are mentioned in the rest of the thesis. References to key research papers and books are provided in each section.

2.1 Binaural audio

Binaural audio is perhaps the most straightforward way of dealing with three-dimensional audio. Since we perceive three-dimensional sound with our two ears, all the relevant information is contained in two signals; indeed, our perception is the result of interpreting the pressure that we receive at the two ear drums, so recording these signals and playing them back at the ears should suffice for recreating life-like aural experiences.

Our perception of the direction of sound is based on specific cues, mostly related to signal differences or similarities between the ears, that our brain interprets and decodes. In the end of the nineteenth century, Lord Rayleigh identified two mechanisms for the localization of sound: time cues (which are also interpreted as phase differences) are used to determine the direction of arrival at frequencies below 700 Hz, while intensity cues (related to signal energy) are dominant above 1.5 kHz [Rayleigh (1896)]. In the low frequency region of the audible spectrum, the wavelength of sound is large compared to the size of the head, therefore sound travels almost unaffected and reaches both ears regardless of the direction of arrival. Besides, unless a sound source is located very close to one ear, the small distance between ears does not cause any significant attenuation of sound pressure due to the

decay with distance. At low frequencies, the only difference between the signals at the ears is therefore a phase difference, related to the difference in time of arrival of sound. According to Gardner (1998), "the question of which ear has the leading phase can be unambiguously determined below 700 Hz". The interaural time delay (IDT) is used by our auditory system to detect the direction of arrival of sound roughly below 1.5 kHz. The other important cue for sound localization is the interaural level difference (ILD), which lateralizes sound towards the ear that receives the signal with the greatest intensity. This cue would in principle work at all frequencies, but the shadowing of the head is not enough to cause a significant level difference at low frequencies, unless the source is very close to one ear. ITD and ILD are the main cues that our hearing system uses for localizing sounds. However, for a given direction of arrival of sound, there exists a whole locus of directions that would give the same ILD and ITD cues: this corresponds to a cone obtained by rotating the line connecting the source to the head around the axis that intersects the two ears, as shown in Figure 2.1. The cones are called cones of confusion, meaning that each point on the surface of a cone produces the same ILD and ITD cues and therefore would make it impossible to discern the position of the source. The most obvious example of such ambiguity of cues is the perception of a sound coming from the front direction versus a sound coming from the back. Fortunately, our hearing system relies on two additional mechanisms to solve this ambiguity and discriminate different directions within the same cone of confusion; firstly, a small rotation of the head can cause a variation in ITD and ILD, which will lead to the discrimination of direction. Besides, the spectral content of sound is modified by the outer ear due to the interaction with the pinna, which introduces peculiar filtering according to the direction of arrival of sound [Blauert (1997)].

The basic concept behind 3D binaural audio is that if one measures the acoustic pressure produced by a sound field in the position of the ears of a listener, and then reproduces exactly the same signal directly at the ears of the listener, the original information will be reconstructed. Binaural audio is perfectly linked with our perception, because it takes implicitly into account the physical mechanisms that take part in our hearing. Binaural recordings are implemented by means of manikin heads with shaped pinnae and ear canals, with two pressure microphones inserted at the end of the ear canal, thus collecting the signals that a human would perceive. Experiments have been done with miniature microphones inserted into the ear canals of a subject, to obtain recordings that are perfectly tailored to a person's shape of the outer ear [Griesinger (1990)]. Binaural playback requires using headphones to deliver each ear the corresponding recorded signal, and the technique delivers good spatial impression. It is worth mentioning that while listening to conventional mono or stereo material through headphones conveys a soundstage located within the head, the use of binaural technique accurately reproduces sounds outside the head, a property which is called

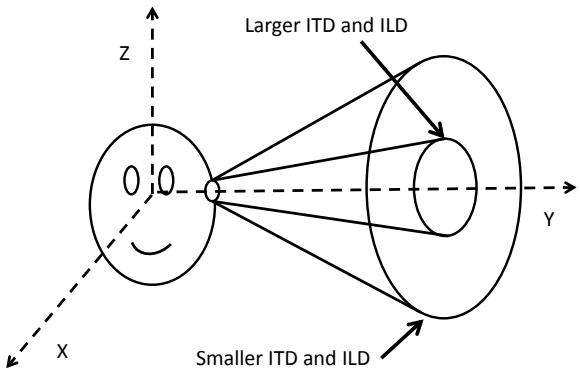

Figure 2.1: Schematic illustration of the concept of cone of confusion. Sound sources located on the surface of the same cone generate the same ITD and ILD cues.

"externalization".

Physically, the signals that reach the ear drums when a sound source emits a sound from a certain position can be expressed as the convolution between the sound emitted by the source and the transfer function between the position of the source and each ear (neglecting effects of the room). The head related transfer functions (HRTF) depend on the position of the source, the distance from the listener and the peculiar shape of the outer ear that is used during recording. Various HRTF databases are available which offer the impulse response recordings done with the source sampling a sphere at a fixed distance (far field approximations are used and distance is usually neglected). With such functions, binaural material can also be synthesized by convolution: once a source and its position are chosen, the left and right binaural signals are obtained by convolving the source with the left and right HRTF corresponding to the position of the source. In this way, realistic virtual reality scenarios can be reproduced. In the real time playback of synthetic sound fields, the adoption of head tracking to detect the orientation of the listener and adapt the sound scene accordingly has been proven invaluable for solving the localization uncertainty related to the circles of confusion or the front-back ambiguity.

Playback of binaural material can also be achieved through loudspeakers, but in this case the left and right signals mix together in the air and reach both ears; since it is essential that the left signal reaches the left ear only and vice versa, cross-talk cancellation filters are required to cancel the left signal reaching the right ear and vice versa. This technique, called transaural playback, is extensively described in Gardner (1998). It provides very good

results as long as the listener sits in the intended position (the position for which the filters have been computed), otherwise the crosstalk cancellation filters would fail in delivering each ear the correct signal.

Binaural playback is certainly a preferred application for mobile audio devices, where most user employ earphones. Binaural 3D audio, consisting of just two channels, can readily be distributed through the conventional media. Among the drawbacks of binaural technology are the need of using headphones for listening (transaural systems are not widespread and too demanding in terms of sweet spot) and the fact that to obtain the best localization effects one would need to use personalized HRTF or recordings done with a replica of one's own ears.

2.2 Ambisonics

Ambisonics is based on the expansion of the sound pressure field at a single point of space into a Fourier–Bessel series; for a monochromatic sound field, after writing the wave equation in spherical coordinates, the pressure at point $\vec{r}$ can be expressed as:

$$p(\vec{r},\omega) = \sum_{m=0}^{\infty} i^m j_m(kr) \sum_{0\leq n\leq m} A^{\sigma}_{mn}(\omega) Y^{\sigma}_{mn}(\theta,\varphi), \tag{2.1}$$

where $Y_{mn}(\theta,\varphi)$ are the spherical harmonics, $j_m(kr)$ are the Bessel functions of the first kind, A^{σ}_{mn} are the coefficients of the expansion, which describe the spatial properties of the field, and $k = \omega/c$ is the wavenumber. In terms of physical magnitudes, the zeroth-order component corresponds to the sound pressure and the first-order ones correspond to the three components of the equalized pressure gradient, which are equivalent to the components of the acoustic velocity vector [Bonsi and Stanzial (2001, 2002); Cotterell (2002)]. Higher-order components are not associated with physical quantities that are directly measurable: they are linear combinations of derivatives of the sound pressure field.

Ambisonics appeared in the 1970s as a way to encode all the relevant information for the recording of the spatial properties of sound fields in a single point of space and their subsequent decoding through suitable configurations of loudspeakers [Gerzon (1973, 1975a)]. After binaural audio, it was the first recording technology to actually take into account the vertical component of sound fields. Its initial formulation was based on the recording of the scalar acoustic pressure and three orthogonal components of the pressure gradient vector in a single point of space, without indications that these quantities were considered as first-order terms in the general expansion of Equation 2.1. The rationale for this approach is that the perception of sound for a listener depends on the pressure and the pressure gradient (in particular, its lateral component, for stereo or horizontal surround) at the

head position. In the encoding (or recording) part, capturing these components was therefore the goal of Ambisonics. Microphones for capturing each component separately exist, and correspond to the omni-directional pressure transducers and the bi-directional pressure-gradient microphones. Since placing four microphones in a coincident configuration, three of which oriented along orthogonal directions, was not feasible without introducing significant acoustic interference by the microphones themselves, an elegant solution was found to obtain the desired signals indirectly, from combinations of mixed pressure and pressure-gradient transducers: the resulting device is a tetrahedral microphone, whose principles and characteristics are described in Section 3.2.2. Beyond first order, the signals cannot be obtained directly from corresponding transducers; one common solution is to employ spherical microphone arrays, a series of pressure transducers arranged on the surface of a sphere, and deduce the higher-order signals by weighting each microphone with the projection of the desired spherical harmonic in its direction [Abhayapala and Ward (2002); Park and Rafaely (2005)].

After the components of the acoustic field are captured, they need to be reproduced through loudspeakers to reconstruct the original acoustic field. An Ambisonics decoder is a device that combines the available Ambisonics signals in a proper way and outputs the signals that have to feed the desired loudspeaker configuration to correctly recreate the recorded information. There are two possible approaches for the decoding of Ambisonics, to which we will refer here as the *physical* approach and the *psychoacoustic* one. In the physical approach, given a certain loudspeaker configuration the goal is to combine the Ambisonics signals to the speakers in such a way so that the reproduced components (the coefficients of the spherical harmonic expansion of the field) in the target sweet spot have the maximum resemblance with the original ones. Ideally, in this case one could record a sound field, decode it, play it back through speakers and record it again with the same microphone used for the original recording to check the resemblance between the original and reconstructed versions. The psychoacoustic approach prescribes building the decoder so that it optimizes the reconstruction of certain physical parameters according to our perception; in order to achieve this, Gerzon (1992) identifies two criteria for optimal playback of Ambisonics: the correct reconstruction of the velocity vector at low frequencies (which is equivalent to the physical decoding) and the energy vector at high frequencies. These conclusions are a consequence of Rayleigh's theories, summarized in Section 2.1, where sound localization at low frequencies is related to the phase difference between the ears, in turn related to the acoustic velocity, while at high frequencies the direction of arrival of the energy is the dominant cue. The physical and the psychoacoustic criteria result in linear decoders, where the signal sent to each loudspeaker is a linear combination of the Ambisonis signals (although in the psychoacoustic decoder proposed by Gerzon the components are split into two frequency bands and a different linear combination is used in each band).

Let us consider a system of N loudspeakers, whose positions are identified by the N unit vectors $\widehat{u}_i$, as seen from a Cartesian reference frame whose origin is located at the listening spot. Each loudspeaker plays back a certain signal with a gain G_i. The velocity vector is defined as

$$\vec{V} = \frac{\sum_{i=1}^{N} G_i \widehat{u}_i}{\sum_{i=1}^{N} G_i} = r_V \widehat{u}_V, \tag{2.2}$$

where $\widehat{u}_V$ is the unit vector in the direction of the velocity vector and r_V its modulus. The energy vector is defined as

$$\vec{E} = \frac{\sum_{i=1}^{N} G_i^2 \widehat{u}_i}{\sum_{i=1}^{N} G_i^2} = r_E \widehat{u}_E, \tag{2.3}$$

where $\widehat{u}_E$ is the unit vector in the direction of the energy vector and r_E its modulus.

According to Gerzon's psychoacoustic criterion, the perceived direction of the sound source is $\widehat{u}_V$ and $\widehat{u}_E$ at low and high frequencies respectively. The goal of an Ambisonics decoder is therefore, given a sound source from a certain direction, to reproduce both velocity and energy vectors aligned with the direction of the source, and to maximize their modulus r_V and r_E. For regular loudspeakers configurations, having equal angular distance between each speaker and its neighbors, the requirements of the optimum decoding can be met, and it has been demonstrated [Daniel (2000)] that there exist two sets of coefficients that maximize r_V and r_E separately, and that in this case the directions of the two vectors coincide. In regular configurations, the reconstructed magnitude of the velocity vector can always reach its maximum value 1, but the magnitude of the energy vector could only reach the unit value if only one loudspeaker participated in the playback. In general, increasing the order of the Ambisonics expansion, in a proper decoder the maximum value of the modulus of the energy vector tends to one. As a consequence of the different criteria applied at low and high frequencies, Ambisonics decoders are often dual band, applying different coefficients above and below 700Hz. Irregular loudspeaker configurations represent a challenge for decoding designers, since the problem of finding the best coefficients is non linear, and in some cases one cannot meet all criteria simultaneously. Various strategies have been proposed for the search of the optimum coefficients, such as those described in Wiggins et al. (2003), Tsang and Cheung (2009), Tsang et al. (2009), and Kirkpatrick et al. (1983).

Other proposals for Ambisonics decoding are based on non-linear processing; among these, we mention the DirAC method [Pulkki (2007)], where a non-linear analysis is performed on the Ambisonics components to extract the directional and diffuse components in frequency bands, using concepts from sound intensity analysis that are detailed in Section 3.1. Once the directional and diffuse parts are separated, the former is played back as a monaural signal spatialized with amplitude panning, while the latter is

decoded using decorrelation techniques to improve the perception of diffuseness.

Increasing the order of the Ambisonics expansion provides more information about the sound field near the measurement point. In the decoding stage, the constraints imposed by matching this additional information result into an extended area where the reconstructed field matches the recorded one, which is interpreted as having a wider accurate listening area, as discussed in Daniel and Moreau (2004), Bertet et al. (2006) and Daniel et al. (2003).

Although strictly tied, the Ambisonics recording and playback parts are separate problems that use to be tackled with separate approaches; to summarize, Ambisonics recording aims to obtaining signals corresponding to the spherical harmonics of a given order by means of suitable arrangements and combinations of microphones, while decoding aims to reconstruct the recorded quantities by matching the physical signals or meeting the psychoacoustic criteria.

The strong point of Ambisonics is the complete encoding of three-dimensional fields in a small set of signals. At first order, the spatial characteristics of the sound field are contained into four channels, while at order N the number of channels is $(N+1)^2$, which is the number of spherical harmonics up to order N. On the other hand, at low orders the perceived localization accuracy of sources turns out to be less than ideal, at least with respect to sharp, pin point sources. This makes first-order Ambisonics more suitable for the playback of diffuse sounds, reverberation or spread sources. When decoding first- or second-order Ambisonics to a given playback configuration, mostly all speakers participate to recreate the field and play back the signal with relatively large intensity. While this gives the correct reconstruction in the sweet spot, as soon as a listener moves outside the sweet spot the sound tends to be perceived as coming from the nearest speaker, therefore correct localization is seriously compromised.

2.3 Wavefield synthesis

Wavefield Synthesis (WFS) was introduced in the 1980s [Berkhout (1988)] as an approach to sound spatialization derived from the Kirchoff-Helmholtz integral theory. WFS is a special case of Holophony, a theory that allows to reconstruct a wave field within a volume given the acoustic variables at the boundary of that volume. Our introduction to WFS follows the treatment given in Nicole (1999), focusing on the essential aspects without going into the mathematical details. Holophony is based on the Huygens' principle, which states that the wave field produced by a source is equivalent to the field generated by a continuous distribution of sources located on the wavefront generated by the original source. Let us consider a region of space Ω divided into a sources' domain Ω_1 and a playback domain Ω_2, so that

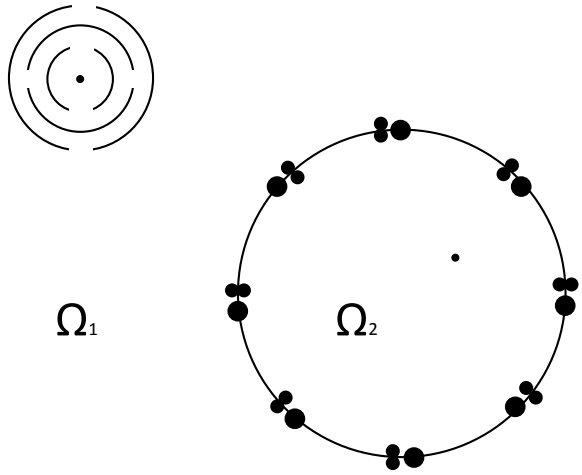

Figure 2.2: WFS recording of pressure and pressure gradient with omnidirectional and bidirectional microphones respectively, in the boundary between the source and playback domains.

the first contains the acoustic sources and the second does not contain any of them. The purpose of Holophony is to express the sound field in Ω_2 as generated by a distribution of sources in the boundary $\partial\Omega$ between the two domains. Mathematically, the problem can be solved via Green functions and the pressure field at time t in the position $\vec{r}$ inside Ω_2 can be expressed as

$$p(\vec{r},t) = \int\int_{\partial\Omega_0} \partial S_0 \widehat{n} \cdot \left\{ \int_{t_1}^{t_2} dt_0 \left[g(\vec{r}-\vec{r}_0, t-t_0)\vec{\nabla}p(\vec{r}_0,t_0) - p(\vec{r}_0,t_0)\vec{\nabla}g(\vec{r}-\vec{r}_0,t-t_0) \right] \right\} \quad (2.4)$$

where $\widehat{n}$ is the vector orthogonal to the surface ∂S, $\vec{r}_0$ is the integration variable, g is the Green function, and t_1 and t_2 define the duration of the sound event. Therefore, the physical reconstruction of the acoustic field would require arrays of pressure and pressure gradient transducers located on a surface enclosing the desired reconstruction domain, as shown in Figure 2.2. This is equivalent to recording the contribution of the secondary sources located at the boundary. The subsequent playback stage would require playing back the pressure and pressure gradient signals by loudspeakers capable of

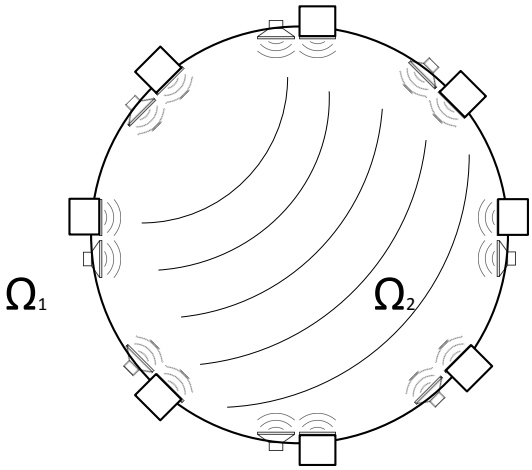

Figure 2.3: Holophonic reconstruction of the field inside Ω_2 with monopole and dipole speakers located on the boundary.

generating equivalent signals located in the same positions as the recording transducers (see Figure 2.3).

Pressure signals are generated by monopole loudspeakers, that is baffled speakers where the drivers are enclosed in a finite volume, while pressure gradients can be generated by loudspeakers where the diaphragm radiates sound on both ends. The whole process, illustrating the ideal link between holophonic recording and playback, is shown in Figure 2.4. The approach described so far requires a continuous distribution of transducers and loudspeakers located on a three-dimensional surface. For a two-dimensional reconstruction of the field, for example in the horizontal plane, it was found that it is sufficient to locate the microphones and the corresponding loudspeakers on the intersection between the enclosing surface and the desired reproduction plane; this result comes from the *stationary phase theorem* and the bi-dimensional simplification is called *stationary phase approximation*. To make the approach feasible, a further approximation allows to make use of a finite number of spaced transducers (microphones and loudspeakers), located in discrete points along a curve on the plane of interest. This approximation is the so called Rayleigh formulation of the Kirchoff-Helmholtz integral, described in Nicole (1999). WFS implements the aforementioned approximation. Although the proposed approach links the recording and the playback stages, WFS is mostly implemented as an exhibition system only, where virtual sound sources are rendered in space from mono tracks

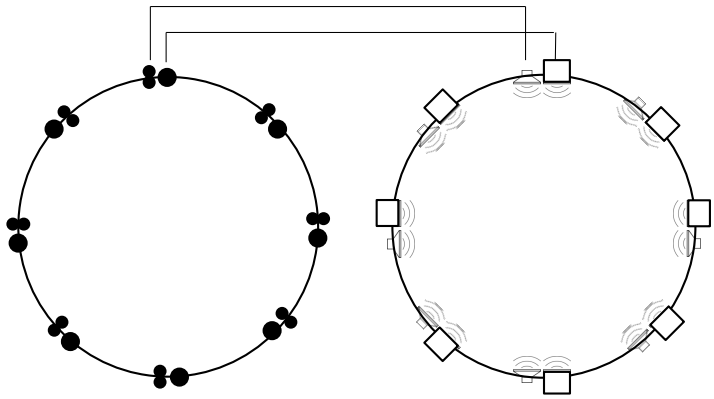

Figure 2.4: Schematic representation of the link between recording and playback with holophonic technology; omnidirectional (pressure) components are reproduced by baffled speakers, while bi-directional (velocity) components are reproduced by dipole speakers.

and spatialization meta-data (angles and distance), from which the signals to play from each loudspeaker are computed. WFS installations feature the ability to reproduce the sound field in a wide listening area, allowing a relatively vast audience to enjoy the "sweet spot". For example, when the technology is used to reproduce a plane wave coming from a certain direction, all the listeners will effectively perceive the sound as coming from the intended direction, no matter their position within the system; on the contrary, with discrete loudspeakers and panning algorithms, the direction of arrival of sound depends on the relative position between the listener and the loudspeaker. The main drawback of WFS is the large number of loudspeakers required; the approximation of the Rayleigh formulation is valid only as long as the separation between the loudspeakers is small, or in the same order, compared to the wavelength of the acoustic field. This implies that the loudspeaker have to be close to each other in order to reproduce correctly the field at high frequencies; for example, obtaining a correct rendering at 5 kHz would require a speaker every circa 8 cm. This condition, together with a required length of several meters to enclose an audience of a few people, implies a number of speakers in the order of a few hundred for rendering in a single horizontal plane! A proposed solution to this inconvenience is to use flat rigid panels with piezoelectric transducers as multi actuators, which

make each panel becoming equivalent to dozens of loudspeakers. While this would not reduce the number of required channels, it would simplify the installations. However, the state of the art in this field is still limited by the frequency response of such actuators when coupled to a rigid plane, which is strongly influenced by the discrete modal distribution of bi-dimensional vibrations in panels [Boone (2004); Corteel et al. (2002)]: this results in a strong filtering of the sound with which equalization cannot cope.

Applications of WFS range from teleconferencing, where multiple participants can be rendered in different positions, thus accompanying the visual localization cues in a conference call, to cinema sound systems, to installations for special events and exhibitions, to sound systems for spatial sound in dancing clubs.

2.4 Amplitude panning

Amplitude panning is a technique in which the same audio signal is reproduced through two or more loudspeakers, with appropriate level differences, so that a phantom virtual source is perceived by the listener in a position between the loudspeakers. The earliest example of amplitude panning dates back to Blumlein (1933); his proposal of a coincident stereo microphone configuration is based on two velocity microphones oriented 45° to the left and right respectively. Since the microphones are located in the same position, the two channels exhibit no difference in time of arrival, but the directivity of the transducers is such that each microphone will output a different signal level depending on the position of the sound source with respect to the array. In the arrangement shown in Figure 2.5, two coincident figure of eight microphones are angled apart θ_m (90° as proposed by Blulmlein); a sound source in the direction θ_s produces a level difference between the microphone signals given by

$$\Delta dB = 20 \log_{10} \frac{\cos(\theta_m/2 - \theta_s)}{\cos(\theta_m/2 + \theta_s)} \tag{2.5}$$

When the left and right signals are reproduced through a stereo pair of loudspeakers, the listener has the impression that the source is reproduced in its original position. Amplitude differences between two loudspeakers generate phantom sources in the line between the speakers; the lower part of Figure 2.5, after Bartlett and Bartlett (1999), shows the perceived position of a sound source as a function of the level difference between channels.

Blumlein's method is related to stereophonic sound recording, but the principle of generating level differences between speakers to recreate phantom sources can be applied to monophonic signals. Two laws have been proposed for stereo panning: the sine law

$$\frac{\sin(\theta_S)}{\sin(\theta_0)} = \frac{g_1 - g_2}{g_1 + g_2} \tag{2.6}$$

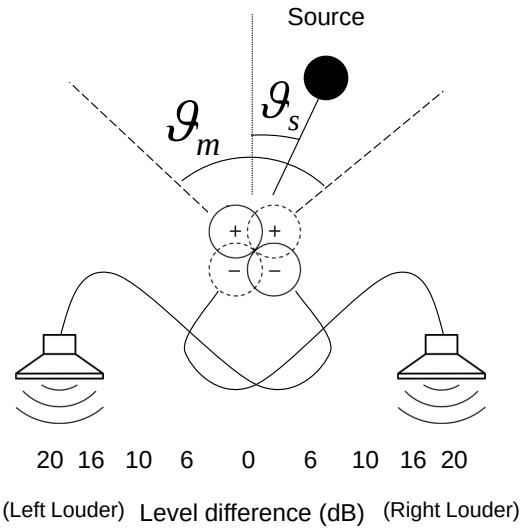

Figure 2.5: The Blumlein recording technique for amplitude stereo panning. The position of the source translates into signal level differences at the output of the two microphones due to their directional pattern. This in turn translates into a different perceived position of the sound source in the stereo panorama according to the level difference between the loudspeakers signals.

and the tangent law

$$\frac{\tan(\theta_S)}{\tan(\theta_0)} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{2.7}$$

where θ_S is the azimuth of the source, θ_0 is the aperture of the speakers and g_1, g_2 are the gains to the speakers. Given g_1 and g_2, these laws allow to predict the perceived source position or vice-versa, given the desired angle one can calculate the required gains. In the case of a stereo system, reproducing the same signal with different amplitude from two speakers generates ILD cues that result in the localization of sound. The panning laws can be extended to 3D loudspeaker layouts, where sound sources are located in space using three loudspeakers that define a triangle inside which the source can be reproduced. The most employed method is Vector Based Amplitude Panning (VBAP) [Pulkki (1997)]: the perceived direction of sound is given by the direction of a vector which is the sum of three vectors aiming to the speakers and having magnitudes proportional to their gains, as seen from the listener's position.

In this formulation, referring to Figure 2.6, the panning vector which gives the position of the phantom source is expressed as

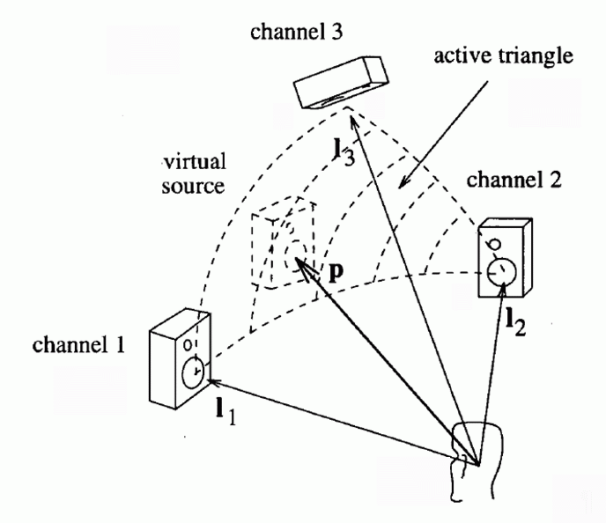

Figure 2.6: Schematic representation and definition of vectors of the VBAP formulation in three dimensions. Image taken from Pulkki (1997).

$$\vec{p} = g_1\vec{l_1} + g_2\vec{l_2} + g_3\vec{l_3}. \tag{2.8}$$

Given the position of the virtual sources, the gains can be calculated via

$$\vec{g} = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix} \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}^{-1}. \tag{2.9}$$

For generic 3D loudspeaker setups, non overlapping triangles are defined using neighbor loudspeakers, and the panning of sources are done by first detecting in which triangle they fall and then calculating the corresponding gains to the speakers of the triangle. The 3D panning laws, as well as the sine and tangent laws, specify the relationship between the gains; however, to spatialize a sound source between speakers and maintain its perceptual level for different positions, a further constraint has to be imposed on the normalization of the gain factors: one has to choose between maintaining a constant energy, which is proportional to the square root of the sum of the square gains and expressed by

$$\sqrt{\sum_{n=1}^{N} g_n^2} = 1, \tag{2.10}$$

or a constant amplitude, which is expressed by

$$\sum_{n=1}^{N} g_n = 1. \tag{2.11}$$

The VBAP has been studied in relationship with the perception, in particular considering the ILD and ITD cues that it generates, and extensive listening tests have been conducted to validate the approach. Pair-wise or triplet-wise amplitude panning is relatively simple to implement and works well in rendering the intended position of the sound source. It does not impose strict requirements on the loudspeaker system, other than the coverage has to encompass the desired playback angles and the loudspeaker density does not have to be too scarce, to prevent "holes" of sound between speakers. Normalizations in between amplitude and energy (linear or square sum) are possible to adapt the algorithm to the listening environment. VBAP aims to reproduce sharp source images. Extensions of the algorithm to give size to sound sources can be implemented introducing decorrelation between speakers [Potard (2006), Potard and Burnett (2004)].

2.5 Hybrid approaches

All the techniques reviewed so far have their strong and weak points. The adoption of one or another depends therefore on the context of application and on the type of content as well; if a full 3D soundscape had to be recreated, WFS would hardly be chosen, because so far it is restricted to a single plane, due to the high complexity and number of channels required for a 3D system. Low-order Ambisonics (first- or second-order) works really well for filling the space and reproducing reverberation or ambient sounds with a relatively low number of loudspeakers and with a low number of encoding channels, but fails in reproducing sharp focused sources that must be perceived consistently within a large listening area. Amplitude panning works really well for reproducing sources in a focused position, but may not be the best choice for reproducing a broad, enveloping sound. Binaural sound is ideal for headphone listening, but is not suitable for playback with loudspeakers for a large audience. Nevertheless, some of these techniques can be integrated and used together, to join their strengths and circumvent the weak points. In a layout-independent scenario, Ambisonics and amplitude panning are the most employed techniques; combining both of them is relatively easy, and allows sound engineers to pick the most suitable one according to the type of sound source, or even to complement a sharp, focused source with spread, enveloping reverberation. In this thesis, a hybrid approach of Ambisonics and amplitude panning has been adopted for the rendering of test material used in Sections 4.1 and 4.2. A layout-independent production format was chosen, where the audio engineers do not use information about the loudspeaker layout. The engineers mix the content by controlling the

level of tracks and assigning them to their positions in space, e.g. specifying azimuth and elevation. The gain and position of a source can vary on a sample-by-sample basis. The information about the loudspeaker position is only needed in the layout-specific decoding stage, when calculating the gains of the amplitude-panned sources and the decoding of the Ambisonics content.

Among other hybrid methods, it is worth mentioning Ambiophonics, a proposal for playback of existing stereo and 5.1 surround recordings that combines transaural playback through a stereo set of loudspeakers and generation of enveloping ambience and reverberation through Ambisonics techniques. A description of Ambiophonics' principles is found in Glasgal (1995), while Farina et al. (2001) present the technical details related to the practical implementation.

2.6 Current systems and playback formats

Despite the advanced state of the art in 3D audio, the expression "surround sound" still means horizontal surround sound to most of the music and cinema consumers and professionals. The most common formats beyond stereo are in fact 5.1 surround and its extension, the 7.1 surround.

5.1 is a format which uses a standard stereo Left-Right pair, a Center channel and two channels for Rear Left Surround and Rear Right Surround, plus a channel for Low Frequency Effects (LFE) via a dedicated subwoofer. The loudspeaker layout, specified by the standard ITU-R BS 775, prescribes positioning the five channels at the same distance from the listener, at azimuth angles of $\pm 30°$ for the L and R channels, $0°$ for the Center channel and $\pm 110°$ for the rear surround channels. This is the recommended array for music, while for movie sound the rear surround channels are distributed to multiple loudspeakers uniformly located along the side and rear walls. The 7.1 for music adds two rear surround channels whose recommended position is between $\pm 135°$ and $\pm 150°$, while the two surround channels from the 5.1 are displaced between $\pm 90°$ and $\pm 110°$. For movie theaters, the 7.1 surround splits the existing left and right surround lines of speakers into four channels, to accommodate the Back Left Surround and Back Right Surround channels. 5.1 surround is a sort of extension of stereo on the horizontal plane oriented to movie content: the presence of the Center channel helps anchoring the dialogs to the screen; the L-R pair is used for front panning and stereo content, and to maintain compatibility with the previous standard, while the two rear channels are expected to provide ambience and enveloping sound, together with occasional sound effects. The position of the speakers, in particular the privileged coverage of the screen area at the expense of a lesser density in the rear part, brings an impairment in the accuracy of panning in the rear positions: the wide angles between the rear speakers cause perceivable "holes" in the sound field when sounds are panned out of

the screen area. The position of the rear speakers themselves corresponds to directions where our auditory system is less accurate in localizing sounds. This is good for the playback of diffuse sounds and effects, but detrimental for accurate rendering of a full surround soundstage. Curiously, this limitation imposed by adapting the system to the needs of the cinema ended up consolidating the very same limitations in the aesthetic language of cinema: since the playback system does not allow accurate panning out of the screen, this concept has been implicitly banned from the practice of sound design, and the surround channels are relegated to ambient sounds, reverberation and some special effect.

Despite its absence in the consumer market, 3D audio has been employed for a long time in electroacoustic music, where space is used by some composers as an artistic tool and is given the same importance as pitch or rhythm; notable works and installations with 3D sound were proposed by composers such as Edgar Varèse, Karlheinz Stockhausen and Leo Kupper. More recently, the potential of 3D sound has been exploited to accompany hemispheric projection in special exhibitions such as world expos, museums and amusement parks. Plenty of tools for 3D production have appeared that implement some of the techniques mentioned in this chapter, but few if any have reached success in the market. This is probably due to the absence of a de facto standard distribution format and playback configuration. So far the mainstream industry has always relied upon fixed configuration systems based on discrete channels; this approach is not feasible for 3D surround, mainly because it would be too optimistic to assume that every consumer would adopt the same complex loudspeaker configuration. Object based approaches, such as those proposed in Hoffmann et al. (2003), Potard (2006) and Fascinate (2010), have appeared to circumvent this problem; they represent an important change of paradigm, where the sound engineer leaves behind the problem of thinking in terms of output channels and only considers each sound source in terms of its level and position in space. An interesting example of a tool for object-based audio production is the Soundscape Renderer [ssr (2012)], a platform for rendering an object based scene to various output formats, including WFS, Ambisonic and binaural.

In recent years, 3D audio has appeared in the cinema industry; currently, four companies offer solutions that include loudspeaker setups in the upper hemisphere, tools for doing the audio post-production in 3D and equipment to playback the resulting soundtracks [immsound (2012); Auro3D (2012); iosono (2012); Dolby (2012)].

3 Recording

Recording is the first step in the traditional as well as in the 3D audio work-flow. Most of times, the goal of recording is to capture and recreate the illusion of reality, or an aesthetically appealing version of it: if so, it seems essential to employ techniques that allow capturing the full spatial information in 3D.

Recording complex audio scenes in 3D implies being able to separate the signals associated to spatially separated sources, and also to separate the diffuse, immersive reverberation from the direct, localized sound, for later being able to reconstruct the acoustic field with the original spatial properties. Adaptations of existing microphone techniques from stereo and surround recording have been studied and are successfully applied, especially for channel-based surround systems [Theile and Wittek (2011); Williams (2012)]. Regarding more generic approaches, the state of the art is advanced in the field of spherical harmonic microphones of high order and beamforming, both from the theoretical point of view and from the number of prototypes and their applications [Gerzon (1973, 1975b); Farina and Ayalon (2003); Farina et al. (2007); Poletti (1996, 2005b); Elko (2000); Eigenmike (2012)]. However, given the complexity of these approaches and the physical size of the related prototypes, we wanted to have a second look at first-order transducers and see what are their actual limitations in terms of accuracy from a physical point of view. Among first-order transducers, apart from state of the art microphones, anemometric sensors are available which feature really accurate polar patterns over the whole frequency range and small size; based on these promising features, our main goal in the research on recording was to use them for building simple higher-order microphones; the core of the research presented here is the study of a second-order Ambisonics microphone built upon few first-order anemometric transducers. Our initial step was therefore the characterization of first-order anemometric probes and the comparison with state of the art tetrahedral microphones. In order to compare the two topologies, we have relied on objective indicators based on quantities borrowed from the field of sound intensity.

As shown in Figure 3.1, the chapter begins with an introduction to the main concepts related to sound intensity, where some quantities that will be used in the following sections are defined. The second step is the description

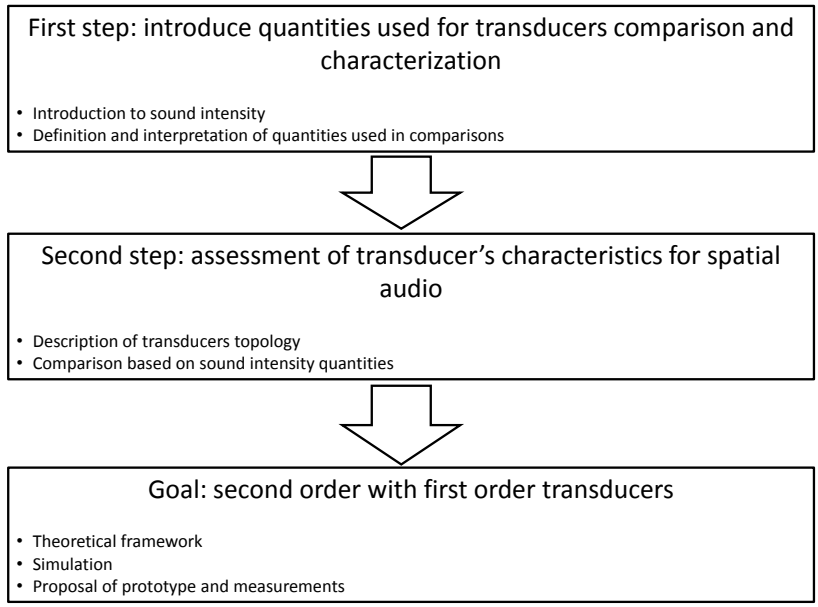

Figure 3.1: Schematic representation of the research steps that constitute the present chapter.

of anemometric transducers and tetrahedral microphones and their comparison in terms of parameters related to sound intensity. Finally, in the main part of the chapter we discuss a proposal for recording the second-order Ambisonics components with first-order transducers, starting with the theoretical framework and then presenting a simulation and measurements on the actual prototype.

3.1 Introduction to sound intensity

The spatial characteristics of sound fields can be quantified within the realm of sound intensity. In this section we firstly review the basic concepts of sound intensity and then introduce some quantities that are related to the description of the spatial qualities of sound fields. The rationale for this is that both anemometric transducers and tetrahedral microphones provide output signals that can be directly used to compute sound intensity and the spatial characteristics of fields, therefore a comparison and analysis of the results in controlled conditions can provide an insight on the transducers accuracy.

Sound waves in the air are perturbations of pressure and density, accompanied by the displacement of the air particles from their rest position

during the motion. The displacement implies a velocity of the air particles, to which kinetic energy is associated. The increase or decrease of pressure implies storage or release of potential energy by the elements of the fluid. Kinetic and potential energy are transported by the sound waves and propagate in the acoustic field.

Sound intensity deals with the measurement of energy transfer in the air by acoustic fields. The energy contained in acoustics fields can be expressed in terms of the acoustic pressure and velocity, which are the signals provided by anemometric probes and tetrahedral microphones. The directional characteristics of sound fields can be interpreted and understood in terms of the direction and magnitude of the energy transfer in the point of interest, which in our case is where the sound engineer would place a microphone and capture what is supposed to be recreated in playback. The acoustic variables depend on space and time, and the dependency is omitted in the following treatment, except when it is convenient to specify it to avoid confusion with other definitions in the frequency domain. The kinetic energy density is the kinetic energy per unit volume associated to the acoustic motion of the air particles and is expressed as

$$K = \frac{1}{2}\rho_0 v^2, \tag{3.1}$$

while the potential energy density is expressed as

$$U = \frac{p^2}{2\rho_0 c^2}. \tag{3.2}$$

c is the speed of sound and ρ_0 is the density of the air at rest in standard conditions. The total energy density of an acoustic field in a point of space is the sum of the potential and kinetic parts, that is

$$w = K + U = \frac{1}{2}\rho_0 \left(\frac{p^2}{z^2} + v^2\right), \tag{3.3}$$

where $z = \rho_0 c$ is the characteristic air impedance.

The instantaneous sound intensity is defined in the time domain as the product of the sound pressure and the particle velocity:

$$\vec{I}(\vec{r}, t) = p(\vec{r}, t)\vec{v}(\vec{r}, t). \tag{3.4}$$

The sound intensity vector gives the direction and magnitude of the energy transfer at any instant of time. In case that the energy is propagating directly from a sound source towards the listener, the direction of the sound intensity vector identifies the direction of the source. The time variation of the total energy density is

$$\frac{\partial w}{\partial t} = \rho_0 \left(\frac{1}{z^2} p \frac{\partial p}{\partial t} + \vec{v} \cdot \frac{\partial \vec{v}}{\partial t}\right) = \rho_0 \left(\frac{c^2}{z^2} p \frac{\partial \rho}{\partial t} + \vec{v} \cdot \frac{\partial \vec{v}}{\partial t}\right). \tag{3.5}$$

Using the mass conservation equation

$$\frac{1}{\rho_0}\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot \vec{v} = 0 \tag{3.6}$$

and the Euler equation

$$\rho_0 \frac{\partial \vec{v}}{\partial t} + \vec{\nabla} p = 0 \tag{3.7}$$

we obtain

$$\frac{\partial w}{\partial t} = -p\vec{\nabla} \cdot \vec{v} - \vec{v} \cdot \vec{\nabla} p = -\vec{\nabla} \cdot (p\vec{v}). \tag{3.8}$$

Substituting Equation 3.4 into Equation 3.8, sound intensity is related to the total energy density by the conservation equation

$$\vec{\nabla} \cdot \vec{I}(\vec{r}, t) = -\frac{\partial w(\vec{r}, t)}{\partial t}. \tag{3.9}$$

According to equation 3.9, the sound intensity vector $\vec{I}$ is associated to the flow of acoustic energy density: for example, integrating the first term of the equation over a surface enclosing a volume, the flow of acoustic intensity through the surface equals the variation of energy density within the volume.

On one hand, measuring the direction and magnitude of the sound intensity vector gives a description of the spatial qualities of a sound field. On the other hand, its measurement with a spatial microphone in controlled conditions can be used to assess the quality of the spatial microphone. Measuring sound intensity requires the direct or indirect measurement of the acoustic velocity. While measuring the sound pressure is a relatively simple task since the invention of the condenser microphone, measuring the acoustic velocity and the sound intensity are more challenging tasks, as indicated by the time lapse between the first patent for a device measuring the sound energy flow, appeared in 1932, to the availability of such equipment on the market, which dates to the 1970s. This delay was mainly due to the technical difficulties in making accurate and reliable transducers that convert the air velocity into electric voltage. The beginning of sound intensity probably dates back to 1931, when Harry Olson filed a patent for a device sensing the energy flow of sound waves, described in Olson (1974). The system was based on a pressure microphone and a velocity microphone (a ribbon transducer), from which the product of pressure and velocity could be deduced from the square of the sum and difference signals, using an equation derived from the square of binomium:

$$(p+v)^2 - (p-v)^2 = 4pv. \tag{3.10}$$

Apparently the device did not find practical use. A decade later, Enns, Firestone and Clapps combined a pressure and pressure gradient microphone for measuring sound intensity, obtaining good results for stationary waves in a tube [Clapp and Firestone (1941)]. Baker (1955) proposed the combination

of a pressure microphone and a hot wire anemometer, together with an electronic multiplier and integrator, for the direct measurement of sound intensity. Although the device was affected by strong thermal noise and was not useful for measurement purposes, the same principle will be re-discovered forty years later and implemented in the Microflown probes, which are used for the measurements presented in this chapter.

Schultz (1955) gave a very important contribution to sound intensity with his patent of a device measuring sound intensity by means of two closely spaced pressure transducers, where the acoustic pressure was obtained as the sum of the signals and the velocity as the difference. This technique has later been improved, while at the same time the researchers gained knowledge on energetic properties of acoustic fields. Since the appearing of the first analogue devices for sound intensity measurements based on pairs of pressure transducers, Schultz's method has been the mostly employed in sound intensity measuring devices until now.

Sound intensity requires therefore the simultaneous and coincident measurement of sound pressure and particle velocity. Traditionally, the particle velocity is measured indirectly via the pressure gradient, employing the linearized Euler equation

$$\frac{\partial \vec{v}}{\partial t} = -\frac{1}{\rho_0}\vec{\nabla} p, \tag{3.11}$$

from which the velocity can be derived by integration:

$$\vec{v} = -\frac{1}{\rho_0}\int_{-\infty}^{t} \vec{\nabla} p dt. \tag{3.12}$$

The pressure gradient is normally measured using a finite difference approximation, as in Schultz's method, employing two pressure transducers in a face to face configuration. The signals of the same pair of transducers are also averaged to provide the acoustic pressure, to ensure the coincidence and phase coherence with the velocity. Sound intensity is therefore calculated as

$$I \approx \frac{p_1 + p_2}{2}\int_{-\infty}^{t} \frac{p_1 - p_2}{\rho_0 \Delta r} dt. \tag{3.13}$$

In the eighties, digital instruments appeared thanks to the development of computers and digital signal processing. With these instruments, it was finally possible to implement digital filters for frequency band analysis. With the availability of measuring devices, sound intensity became a feasible approach to the study of the acoustic fields, and found important applications in source power measurements and noise control.

Recently, sound intensity has been subject to some interpretations where the energetic analysis of sound fields can give insight on its spatial properties. It is worth mentioning, for example, the work of Mann et al. (1987) and Mann and Tichy (1991) on the identification of the energy flux trajectories in acoustic fields, and the work of Schiffrer and Stanzial (1994) and

Stanzial et al. (2003) on the relationship between intensimetric quantities, the magnitude of energy transfer and the complete description of the spatial qualities of sound fields. An application of sound intensity in audio is also found in Pulkki (2007), where directional and diffuse components of the field are distinguished according to the value of the radiation index, which we will introduce later in this section.

In the time domain, the active intensity is defined as the time average of the sound intensity vector:

$$\vec{A}(\vec{r}) \equiv \langle \vec{I}(\vec{r}) \rangle = \langle p(\vec{r})\vec{v}(\vec{r}) \rangle = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} p(\vec{r},t)\vec{v}(\vec{r},t)dt. \tag{3.14}$$

Since the time intensity vector expresses the direction and magnitude of energy transfer at any given time, its average in the time domain expresses the total energy that flows through the measurement point during the time interval considered. Notice that this quantity is in general smaller than the total energy (potential plus kinetic) contained in the sound field during the same interval of time: in fact, the pressure and velocity in the integral of Equation 3.14 are oscillating quantities, therefore the integral of their product depends on their relative phase. In this respect, the term "active" indicates that we are measuring the fraction of the energy that flows through the measurement point, while the remaining part oscillates around it. Although by definition the time interval is extended to infinity, it can be chosen *ad hoc* to focus the analysis on a restricted part of the sound event. For example, one might want to focus only on the first instants of a stationary sound event, which contain the information on the position of the sound source, or one could analyze only the late reverberation caused by a room to evaluate its diffuseness.

The measurement of room impulse responses is the most employed method for the analysis of acoustic fields produced by point sources in enclosures; this method normally only takes the sound pressure into account, but it can be extended to include the three-dimensional acoustic velocity [Bonsi (1998); Stanzial et al. (2000); Farina and Ayalon (2003)]; since the pressure and velocity field generated by a sound source can be expressed as the convolution of the pressure signal emitted by the source with the pressure and velocity impulse responses of the room, the energetic properties of the field can be calculated from such impulse responses: in this case, the components of sound intensity for an impulse are given by the product of the pressure impulse response with the impulse response of each component of the acoustic velocity. For analysis purposes, the active intensity can be calculated in different intervals of the impulse response, such as the early decay or the reverberation tail.

It has been shown that the ratio of the active intensity and the average total energy density, normalized to one by a factor c, represents the fraction

of energy that is radiated by the acoustic field at a given point [Stanzial and Prodi (1997)]. Such fraction is expressed by the radiation index η:

$$\eta = \frac{|\vec{A}|}{c\langle W \rangle}. \tag{3.15}$$

The radiation index is useful to differentiate between directional sound events (or fractions of them) and diffuse ones. Directional sound events are characterized by high values of η, because the sound energy comes from a specific position and travels from the source to the listener; on the contrary, diffuse fields are characterized by energy flowing randomly in all directions, therefore resulting in a low average value of energy flow at the measurement point. For example, for a progressive plane wave the radiation index has a value of 1 in any position, meaning that all the energy is radiated by the field. Conversely, for a stationary wave its value is zero, indicating no energy radiation: in this case, all the acoustic energy oscillates around the measurement point and the time average vanishes. In following sections, we will use the values of radiation index in frequency bands, measured with the transducers under test, to assess their accuracy: the fraction of radiating energy is in fact tightly bound to the phase relationship between pressure and velocity; when pressure and velocity oscillate in phase, the time average of their product (the active intensity) reaches the maximum, while if they are 90° out of phase, the time average vanishes. The radiation index, like the energy transfer and all sound intensity related quantities, depends crucially on the phase relationship between pressure and velocity.

The sound intensity vector gives the direction and magnitude of the energy transfer at any instant of time. Apart from calculating the average during a pre-defined time interval, it is interesting to plot the vector for all instants of time during the interval, to track the energy transfer paths, as will be shown in Sections 3.2.3 and 3.2.4, where we present plots of the intensity vector on certain planes of interest.

3.2 Comparison of anemometric transducers and first-order Ambisonics microphones[1]

There is an obvious link between the variables considered in sound intensity and the first-order Ambisonics signals. They both provide information about the energetic and spatial properties of the sound field that is not available in pressure measurements alone, and both approaches consider a complete set of signals that would ensure the perfect reproduction of a sound field in a single point of space (if one only measures the field in one point, the reconstruction in the surrounding area follows with decreasing accuracy as the distance with the point increases). By complete set we mean here that

[1] This section is based on Cengarle and Mateos (2011)

the local description of the acoustic field requires four variables: the scalar sound pressure and the three components of the acoustic velocity vector.

As seen in Section 3.1, the components of the acoustic velocity are traditionally measured by pairs of closely-spaced pressure transducers, applying a finite difference approximation of the Euler equation. This technique is inherently subject to limitations in the frequency range, caused by the finite spacing of the two transducers, and to inaccuracies caused by a possible phase mismatch between them [Jacobsen (1997)]. Besides, the technique is used normally in one dimension (one pair of transducers only measures the component of the particle velocity along one axis) and the size of the assembly can affect the practicality of its usage in small enclosures or regions of space that are difficult to access. The advent of anemometric transducers for the direct measurement of the velocity has introduced a new player in the field of devices for sound intensity measurements [De Bree et al. (1996)]. Models have been released that include three orthogonal coincident anemometric transducers plus a pressure microphone in a compact package. These devices measure directly the coincident pressure and velocity of the acoustic field, therefore their application to sound intensity measurement is straightforward. Various studies have demonstrated the validity of the approach and the accuracy of the results, comparing them with the pressure gradient method [Jacobsen and De Bree (2005); Tjis et al. (2009)]. Anemometric sound intensity probes can be preferred in applications where compactness is a valuable feature, both for accessing small and busy areas (i.e. near engines and machinery) and for the reduced scattering and diffraction that they introduce in the field.

Sound intensity is not the only application where the acoustic velocity is measured. Microphones responding to the acoustic velocity, based on a thin diaphragm exposed to the acoustic pressure on both sides, have been used for audio applications since the first half of the last century [Olson (1931)]. Initially chosen for their rejection of lateral sounds, they were soon employed for the early stereophonic recording techniques [Blumlein (1933)]. A few decades later, in the seventies, the Ambisonics concept was born, which is based on the encoding of the sound field in terms of the sound pressure and the three coincident orthogonal components of the pressure gradient. Due to practical issues, related to the problem of assembling such transducers in a coincident configuration, the microphone for recording Ambisonics was initially designed as a tetrahedral configuration of mixed pressure-velocity microphones capsules [Gerzon (1975a)], from which the pressure and pressure gradients are derived by linear combinations. This approach was successfully implemented and persists to the present date as the only design of first-order Ambisonics microphone. Due to its ability to provide a good spatial audio rendering, together with a high signal to noise ratio (SNR), the microphone has been adopted in the field of room acoustics too, for measuring the spatial characteristics and impulse responses of spaces and also the common room acoustic parameters, such as reverberation time, clarity,

definition and lateral fraction. It turns out that the signals output by tetrahedral microphones, after proper treatment, correspond to the set of signals required by sound intensity measurement. In fact, what is often referred to as pressure gradient or equalized pressure gradient, is actually the acoustic velocity, as confirmed in various works [Bonsi and Stanzial (2001, 2002); Cotterell (2002)]. In this sense, tetrahedral microphones are another candidate transducer for the measurement of sound intensity and its related quantities.

The author has been using extensively both anemometric transducers and tetrahedral microphones for room acoustics analysis and measurement of impulse responses for spatial audio applications, often calculating intensimetric quantities with both transducer topologies; this led to carrying out their characterization and comparison. The comparison presented in this chapter is based on measurement results in different field conditions, varying from anechoic to reverberant, and comparison, when possible, with the theoretically expected results. In Sections 3.2.1 and 3.2.2 we describe the topology of the transducers used and report on some preliminary measurements and calibrations. In Section 3.2.3 we present the measurements in the anechoic chamber, where the performance in the near-field is assessed, while Section 3.2.4 deals with the behavior in reverberant spaces.

3.2.1 Anemometric transducers

In 1994 a sensor for direct measurement of the acoustic velocity, based on the principle of twin-wire hot-wire anemometry, was introduced [De Bree et al. (1996)]. This sensor exploits the variation in electrical resistivity of an electric conductor as a function of temperature; an air flow causes heat transfer by convection in proximity of a wire heated by an electric current. By setting a pair of parallel wires heated at the same temperature in the presence of an air flow, a temperature difference appears due to the fact that the flow transfers heath along its direction, causing a heat transfer from the upstream wire to the downstream one, as shown in Figure 3.2. The subsequent temperature difference causes a difference in the electric conductivity, which is detected by inserting the wires in a circuit with an electric current. From the variation in resistivity one can deduce the variation in temperature, hence the velocity of the air flow.

The Microflown sensor is made of two parallel platinum wires, 1 mm long and 200 nm thick. The wires are heated by an electric current and reach a steady temperature between $200°C$ and $400°C$. A temperature increase in the wires causes the increase of their resistivity. A single wire undergoes a temperature drop in the presence of an air stream by convection with the coldest air. The wire could therefore measure the modulus of velocity orthogonal to its axis. The Microflown sensor uses twin wires and exploits the temperature difference between the two wires when they are exposed to the air flow to measure the component of the acoustic velocity in the direction orthogonal to the wires on the plane containing them. The Mi-

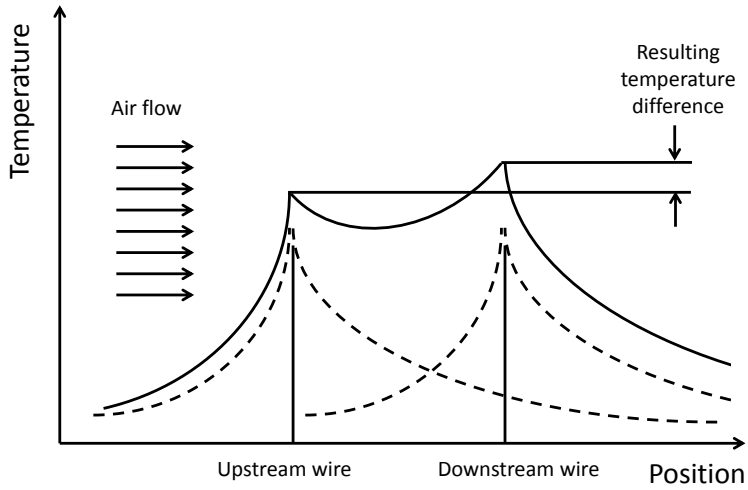

Figure 3.2: Temperature difference between two hot wires due to heat transfer caused by air velocity. Dashed lines are the temperature profiles of the individual sensors in the presence of airflow as indicated in the figure, while the solid line is the resulting combined temperature profile. The downstream wire increases its temperature compared to the upstream one. This temperature difference causes a difference in thermal conductivity which can be measured if the wires are part of an electric circuit.

croflown anemometric transducer has a figure of eight response, described by the equation

$$S(\theta) = A\cos\theta, \tag{3.16}$$

where θ is the angle on the horizontal plane between the direction of the sound source and the front direction of the sensor, as shown in figure 3.3.

The sensor allows measuring air particle velocity in a range approximately between 100 nm/s and 1 m/s in a frequency range between 0 Hz (constant flow) up to nearly 20 kHz. The anemometric sensor has a characteristic response that inherently rolls off high frequencies due to effects of thermal convention and inertia. Neither the amplitude nor phase responses are flat, but their behaviors can be modeled as a series of first-order low pass filters, resulting in a loss of approximately 20 dB at 20 kHz. At high frequencies, the sensitivity decreases due to the time required for the heat to propagate between the two wires and due the thermal capacity of the wires which impedes their instant heating or cooling. The first effect can be approximated by a first-order low pass filter with a corner frequency f_d

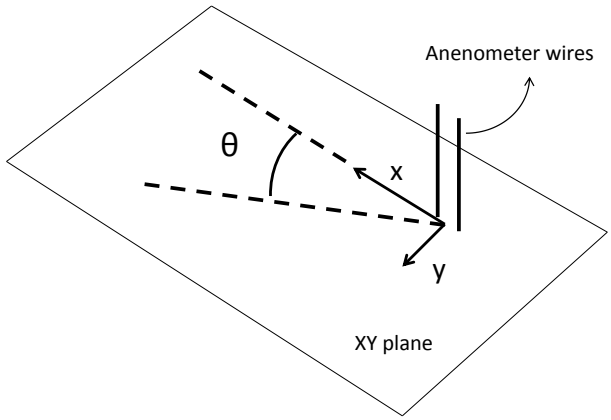

Figure 3.3: Definition of angle of incidence on the plane perpendicular to the wires.

between 500 Hz and 2 kHz; the second effect corresponds to a first-order low pass filter with corner frequency f_h between 2 kHz and 15 kHz. A typical frequency response of the transducer, shown in Figure 3.4, is described by the function

$$S(f) = \frac{LFS}{\sqrt{1 + f_e^2/f^2}\sqrt{1 + f^2/f_d^2}\sqrt{1 + f^2/f_h^2}}, \tag{3.17}$$

where $LFS[\frac{mV}{Pa} \cdot \rho_0 c]$ is the sensitivity at 250 Hz. This equation includes a low frequency rise of 6 dB/oct typical of velocity transducers, associated to a frequency f_e between 30 Hz and 100 Hz. The corresponding phase response is given by

$$P(f) = \arctan\left(\frac{C_1}{f}\right) + \arctan\left(\frac{f}{C_2}\right) + \arctan\left(\frac{f}{C_3}\right). \tag{3.18}$$

These functions have been derived from a physical model describing the behavior of the sensor. The overview presented here is to highlight the most relevant details and specifications of this anemometric sensor. Further details on the operating principles, characteristics and applications can be found in De Bree (2007).

In order to calibrate the anemometric sensors and use them for sound intensity measurement purposes, it is better to measure their actual response and compensate it. The compensation can be done either by fitting the parameters of the model to the measured data, so obtaining the best values of the aforementioned constants, or by inverting their actual response.

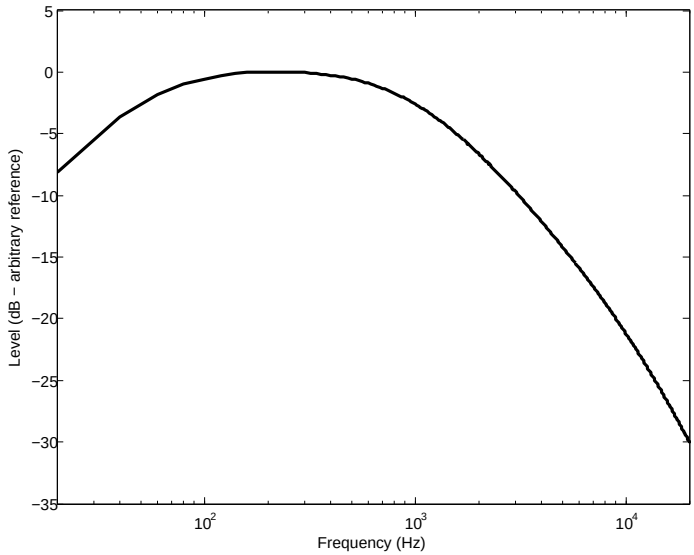

Figure 3.4: Amplitude response as a function of frequency of Microflown anemometric transducer, according to the physical model described by Equation 3.17.

Various methods have been proposed for obtaining the actual response of the velocity transducers and calibrate them. All of them require measuring the acoustic velocity with the Microflown and the acoustic pressure with a reference microphone in a field of known impedance. The actual acoustic velocity can be derived from the acoustic pressure and the impedance, using the following relationship:

$$Z(\omega) = \frac{p(\omega)}{v(\omega)}, \tag{3.19}$$

where $\omega = f \cdot 2\pi$.

The response of the velocity transducer is then given by the transfer function between the actual velocity and the measured one.

Conditions of known impedance are encountered in the following situations:

- Inside a Kundt tube
- In an anechoic chamber (free field)
- Close to a spherical source

The Kundt tube is a cylindrical device with reflective internal surfaces where stationary waves are produced and the amplitude and phase relationship between pressure and velocity at a given frequency is given by an analytic formula. The Kundt's tube was proposed initially for the calibration of Microflowns; the main disadvantage of such device is that the waves are stationary only up to a certain frequency, the cutoff frequency of the tube, which depends on its radius r and is approximated by $\omega \approx 1.8c/r$. Standard Kundt tubes are therefore useful only at low frequencies.

In the anechoic chamber, a small sound source can be used to generate a spherical wave field, where the impedance as a function of distance and frequency is

$$Z(r,\omega) = \frac{p(r,\omega)}{v(r,\omega)} = \rho_0 c \frac{ikr}{1+ikr} \tag{3.20}$$

with $k = \omega/c$. Usually the far field condition is considered, where $kr \gg 1$ and the impedance reduces to $Z = \rho_0 c$. This is equivalent to considering the acoustic field as a plane wave, a condition which holds true only if the distance between the source and the measuring point is roughly at least one order of magnitude greater than the wavelength of sound. Considering typical lengths of anechoic chambers in the order of 10 m, the method is not valid at low frequencies, being therefore a complement rather than an alternative to the Kundt's tube method.

A recent proposal for a full bandwidth calibration uses the field of known impedance generated by a loudspeaker enclosed in a spherical housing. At high frequencies, roughly 100 Hz to 20 kHz, the method is very similar to the free field calibration: the Microflown and a reference pressure transducer are setup in a coincident fashion at distance r from the center of a sphere of radius a containing a loudspeaker with radius b, where the acoustic field generated along the axis of the speaker has an acoustic impedance given by

$$Z(r) = -i\rho c \frac{\sum_{m=0}^{\infty} \left(P_{m-1}(\cos\alpha) - P_{m+1}(\cos\alpha)\right) \frac{h_m(kr)}{h'_m(ka)}}{\sum_{m=0}^{\infty} \left(P_{m-1}(\cos\alpha) - P_{m+1}(\cos\alpha)\right) \frac{h'_m(kr)}{h'_m(ka)}}, \tag{3.21}$$

where P_m is the Legendre function of order m, h_m is the spherical Hankel function of the second kind and order m, and h'_m is its derivative. For large distances from the spherical sources and high frequencies, the above formula is approximated by Equation 3.20. At low frequencies, the same device is used inserting the reference pressure transducer inside the sphere and measuring the acoustic velocity in close proximity of the sphere. Below the internal resonant frequency of the sphere, there is a linear frequency-dependent relationship between the pressure inside the sphere and the velocity at its surface,

$$v = -\frac{i\omega V_0}{\gamma A_0 p_0} p, \tag{3.22}$$

where ω is the angular frequency, V_0 is the volume of the sphere, A_0 the surface area of the loudspeaker piston, p_0 the ambient pressure and γ is the ratio of specific heats (1.4 for air in standard conditions).

The transducer employed for the measurements presented here is the Microflown USP, which incorporates three anemometric sensors, mounted coincidently in orthogonal directions, together with a pressure transducer. This probe allows the direct and simultaneous measurement of the sound pressure and the three components of the acoustic velocity in a single point of space. The probe is 13 cm long with a diameter of 1.25 cm, weighting only 43 grams. The system comes with a signal conditioner unit that powers the probe, providing the current for heating the anemometers, and also provides analogue correction filters tailored to the response of the probe. The analogue filters compensate the response curve, without adjusting for the sensors sensitivity mismatch. A switch allows to bypass the analogue filters and access the raw signal from the anemometers; this choice is useful as it allows to implement the filtering in the digital domain, which brings advantages in the SNR, as will be later explained.

The pressure transducer is an electret microphone with a 1/10 inch diaphragm, a frequency response from 20 Hz to 20 kHz, and a sensitivity of 20 mV/Pa. The velocity transducers work in a frequency range from 1 Hz to 20 kHz with a sensitivity of 20 V/(m/s). Notice that this looks like a huge sensitivity from a physical point of view, but two considerations apply: first of all, this is the signal amplitude after the signal conditioner, which is an active circuit and may therefore provide signal gain; besides, one has to consider that 1 m/s is a very high acoustic velocity. In fact, the value v_0 corresponding to a pressure wave at the threshold of hearing is $5 \cdot 10^{-8}$ m/s, while at the threshold of pain the velocity is in the order of 0.1 m/s. The four sensors are fit within less than 5 mm, a spacing that ensures acoustic coincidence and phase coherence over the whole audio range. The calibration data and parameters have to be measured for each transducer individually. The amplitude of the velocity signal is normalized to the pressure signal by multiplying it by the air characteristic impedance Z_0, so that the pressure and velocity sensors deliver the same output voltage in the presence of a progressive plane wave. Compared to the traditional techniques for indirect measurement of acoustic velocity with a finite difference approximation of the pressure gradient measured by pairs of pressure transducers and the Euler equation, the Microflown has the advantage of avoiding the approximations and inaccuracies inherent in p-p probes. Besides, its small size reduces the diffraction and interference effects on the acoustic field due to the presence of the transducer itself.

The probe was calibrated in an anechoic chamber, using the spherical source provided by the manufacturer, with a diameter of 20.5 cm and a loudspeaker diameter of 6 cm. The high frequency calibration was performed with the transducers placed 31 cm from the surface of the sphere. A Cesva microphone and SPL meter model SC310 was employed as the reference pres-

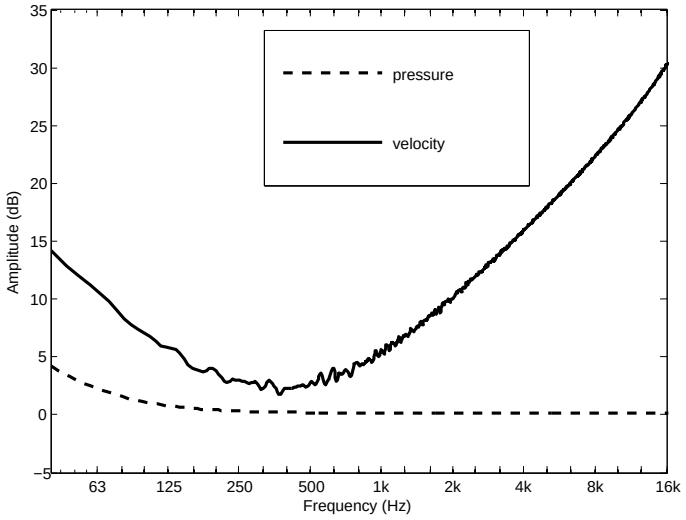

Figure 3.5: Frequency response of the equalization filters for the pressure and the velocity along the x direction. The filters for the velocity transducers along the other axes have similar behaviors.

sure transducer. Once the responses were measured, the parameters of the physical models were calculated using a software provided by the manufacturer. To compensate the measured response, inverse filters were computed by means of the analytic inversion of the minimum-phase zeros-poles representation of the transducers' responses given by the physical model, following a procedure detailed in Bonsi et al. (2005). The resulting equalization filters are shown in Figure 3.5. These filters are obtained in the form of an impulse response, so the correction of a measured signal is performed by convolving it with the IR of the filter in the time domain.

One of the main concerns of the use of anemometric transducers in audio applications is their self-noise. Broadband thermal noise is the dominant noise source in hot-wire anemometers. Besides, the filter boost required to flatten the response increases its audibility significantly, especially at high frequencies. The noise level and spectrum of anemometric transducers have been measured in a quiet anechoic chamber, using a reference sound level meter to assess the absolute values. Firstly, the output of the Microflown has been calibrated in absolute units (mV/Pa) by comparison with the reference microphone in presence of a loud source in the far field, considering a plane-wave acoustic impedance. Subsequently, we measured the output level in absence of sound sources, and derived both the SPL values and the

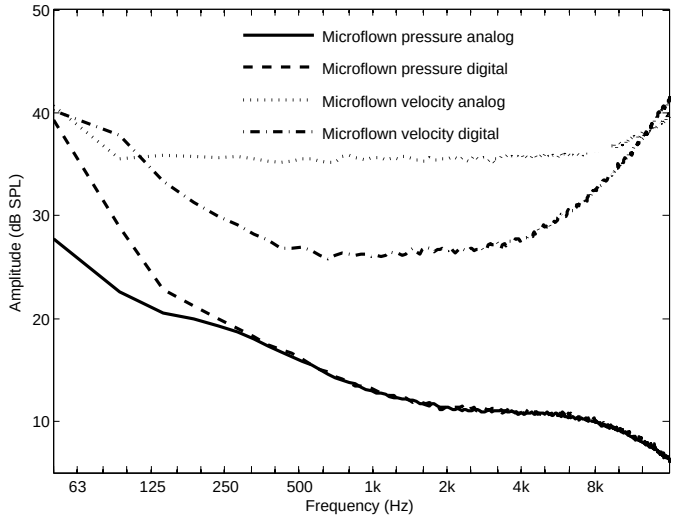

Figure 3.6: Self noise spectra of Microflown pressure and velocity transducers. The benefits of digital filters for velocity transducers lead to a 10 dB improvement in the SNR in the high-mid range.

	Analog	Analog A-wght	Digital	Digital A-wght
p	36	32	46	42
v_x	64	61	61	51
v_y	71	67	60	52
v_z	66	57	57	53

Table 3.1: Self noise of the Microflown transducers in dB SPL, for digital and analog filters.

spectra. The procedure was carried out using both digital and analog filters, evidencing the improvement of the SNR when digital filters are used. The noise spectra are shown in Figure 3.6, while the SPL values, both A-weighted and un-weighted are reported in Table 3.2.1. The improvement in the SNR with the digital filters is about 10 dB for the velocity channels.

3.2.2 Tetrahedral transducers

Due to the physical dimension of a velocity microphone, it is unfeasible to mount three such microphones orthogonally in a coincident position without

causing mutual shadowing. A clever workaround to derive the pressure and three orthogonal components of the velocity in a point of space consists in using a uniform distribution of four microphones located on the surface of a virtual sphere centered on that point. This approach, which can be considered the ancestor of spherical microphone arrays, has been employed since the beginning of the Ambisonics history.

As derived in Beranek (1954), the output of a generic microphone with the diaphragm exposed to the sound waves on one side and facing an acoustic resistance on the other side, for an incoming plane wave

$$p(\vec{r}, t) = \widehat{p} \exp[i(\vec{k} \cdot \vec{r} - \omega t)], \tag{3.23}$$

can be written as

$$V_{out} = A'[p + iB'(\widehat{n} \cdot \vec{\nabla} p)/k], \tag{3.24}$$

where p is the pressure on the diaphragm, $\widehat{n}$ is the unit vector perpendicular to the diaphragm, A' determines the sensitivity of the microphone and B' its polar response. The coefficients A' and B' depend on the acoustic resistance of the rear cavity, the diaphragm's stiffness and its acoustic impedance. Being $\widehat{n} \cdot \vec{\nabla} p$ proportional to the cosine of the angle between the orientation of the microphone and the direction of arrival of the sound wave, the angular response of the microphone can be written in general as

$$V_{out}(\theta) = A + B \cos(\theta), \tag{3.25}$$

with $A, B \in [0, 1]$ and $A + B = 1$. $A = 1$ and $B = 0$ give an omnidirectional, pressure microphone; $A = 0$ and $B = 1$ give a bidirectional, pressure gradient microphone, while intermediate values correspond to various directional patterns, as shown in Figure 3.7.

Since a generic microphone responds to a linear combination of pressure and pressure gradient, on the other way around pressure and pressure gradient can be expressed as linear combinations of suitably arranged generic microphones. This is the principle behind the original implementation of Ambisonics microphones, which use four subcardioid capsules ($A = 0.75$ and $B = 0.25$) arranged as closely as possible and oriented along the faces of a tetrahedron; an early example of such arrangement, implemented with standard microphone bodies, is shown in Figure 3.8

The set of signals produced by the four capsules is referred to as the A-format. Each A-format signal can be expressed as

$$p_{out} = \frac{3}{4} p + \frac{1}{4} \vec{\nabla} p \cdot \widehat{n}, \tag{3.26}$$

where the unit vector $\widehat{n}$ specifies the orientation of the capsule. The orientation of the capsules in a tetrahedral microphone in spherical coordinates (θ, ϕ) are: L_F (left-front): (45°; 54.7°); L_B (left-back): (135°; 125.3°); R_F (right-front): (225°; 54.7°); R_B (right-back): (315°; 125.3°). The pressure

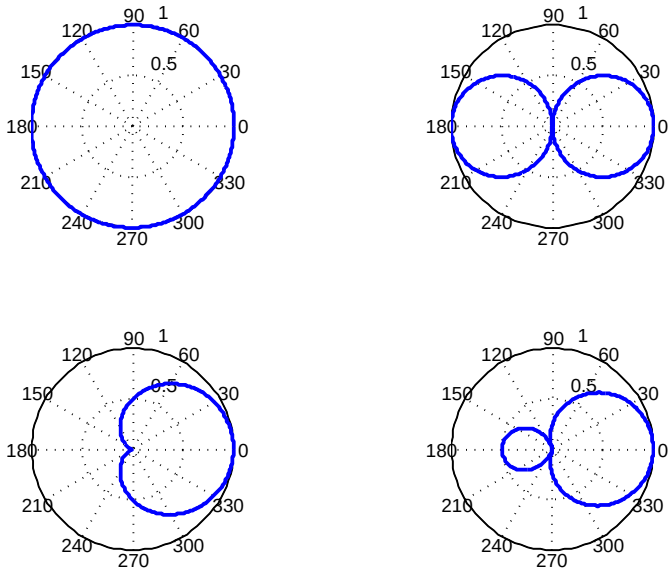

Figure 3.7: Polar patterns of four common directional characteristics of microphones. Top left: omnidirectional; top right: figure of eight; bottom left: cardioid; bottom right: hypercardioid.

and the three velocity components, labeled respectively W, X, Y and Z (a set which is called B-format), are derived as linear combinations of the A-format in the following way

$$\begin{pmatrix} W \\ X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} L_F \\ R_B \\ L_B \\ R_F \end{pmatrix} \tag{3.27}$$

As a convention, the pressure component of the B-format is attenuated 3 dB with respect to the others. In combining spaced microphone capsules, a peculiar filtering occurs, which depends on the spacing of the capsules and the direction of arrival of sound. In the case of tetrahedral microphones, equalization is used to flatten the response for diffuse-field incidence, so that the average response over the spherical angle is rendered flat [Gerzon (1975a)]. Tetrahedral microphones use closely spaced capsules to derive the spherical harmonics of order zero and one of the pressure field, therefore they

Figure 3.8: Four directional microphones in a tetrahedral arrangement for early Ambisonics recordings. Picture from http://www.michaelgerzonphotos.org.uk/tetrahedral-recording-images.html

can be classified as spherical microphone arrays. The main limitation of a spaced approach is the rise of spatial aliasing when the wavelength of sound is comparable to the spacing of the capsules. This effect appears as a deviation of the polar patterns from their ideal shape; in tetrahedral microphones, where the spacing is the in order of a few centimeters, the polar response is affected starting from approximately 8 kHz. Often the characteristics of these microphones are considered in terms of directionality and accuracy of the patterns, discarding the phase relationship between pressure and velocity. In sound intensity applications, the errors related to capsule spacing and spatial aliasing translate into phase inaccuracies. One last thing that is worth mentioning regards the polarity of the velocity signals: in a tetrahedral microphone, W and X are in phase for a progressive plane wave coming from the front direction, while in an anemometric probe the x component of the velocity would have reversed polarity; this is correct, since the incoming velocity is directed towards the negative side of the x axis and vice versa. As a consequence, before performing comparative analysis, the polarity of the velocity measured with the tetrahedral microphones was reversed, and the W channel has been boosted 3 dB in order to maintain the correct relationship with the acoustic velocity.

Two similar tetrahedral microphones, a Soundfield and a Tetramic, were employed in our comparison. The Soundfield microphone is the direct descendant of the first Ambisonics microphone [Farrar (1979)]. The model employed is the SPS422B (2004), consisting of a tetrahedral microphone and an analog preamplifier/processor which receives the signal from the capsules and outputs the four channels of the B-format: the pressure W and the three

components of the velocity X, Y and Z. The capsules are mixed pressure - pressure gradient with a subcardioid pattern. This microphone is frequently employed in critical music recording tasks, therefore features a good SNR (the manufacturer declares an equivalent self-noise of 14 dB A-weighted).

The Core Sound Tetramic is a compact tetrahedral microphone which uses four small-diaphragm cardioid capsules [Tetramic (2007)]. The output is in A-format and can be converted to B-format using a few available software applications, either standalone or plug-ins. We used Tetraproc, a software that also performs equalization of the capsules' responses and provides B-format along with stereo output. The Tetramic is slightly noisier that the Soundfield (the manufacturer declares an equivalent self-noise of 19 dB A-weighted per capsule), but still usable for direct recordings.

3.2.3 Comparison in anechoic chamber

In this section, measurements in close proximity of a spherical source inside an anechoic chamber are considered, exploiting a condition of known impedance to check the accuracy of the relationship between pressure and velocity measured by the different transducers. For a monochromatic plane wave with frequency ω and wave number $k = \omega/c$ the spherical pressure field generated by a point source is

$$p(r,t) = \frac{p_0}{r} \exp\left(i(\vec{k}\cdot\vec{r} - \omega t)\right). \tag{3.28}$$

Applying the Euler equation gives the particle velocity

$$v(r,t) = -\frac{p_0}{c}\frac{1+ikr}{ikr}p. \tag{3.29}$$

The phase and amplitude relationship at a distance r from the source is given by the transfer function corresponding to the impedance Z:

$$Z = \frac{p}{v} = \frac{c}{\rho_0}\frac{ikr}{ikr+1}. \tag{3.30}$$

Explicit expressions for the amplitude and phase are derived by taking the modulus and the angle of the impedance. As is well known, for large r or k (at large distances compared to the wavelength or at small wavelengths compared to the distance), the impedance tends to a constant value, while for small kr the phase relationship between pressure and velocity tends to 90°, since the impedance tends to an imaginary value.

The first step was to measure the transfer function between pressure and velocity in the near field in an anechoic chamber. For all the measurements we used the sine sweep technique to derive the impulse responses [Farina (2000)]. The loudspeaker used is a Genelec 8020 studio monitor. This loudspeaker is not omnidirectional and has a limited low frequency response; despite this, it can playback frequencies down to 40Hz, which is sufficient

for our purposes: even if the level is low and the response is uneven in such a low range, the impedance only depends on the relationship between pressure and velocity, not on the absolute SPL. Given the small dimensions of the speaker, in the order of 20 cm, it is reasonable to assume an omnidirectional radiation pattern well below the crossover frequency: our near-field results are presented in the range of 50 Hz to 2 kHz. Figure 3.9 shows the amplitude and phase relationship between pressure and velocity along the x axis, measured with the three transducers (Microflown, Soundfield and Tetramic) 0.5 m in front of the source. The probe and the microphones were located in front of the acoustic axis of the loudspeaker, with their x axes aligned towards the speaker. The measured impulse responses for each component have been convolved with pink noise before calculating the transfer function. The comparison with the theoretical curve shows that all three transducers render the correct phase relationship between pressure and velocity. In particular, this confirms that the B-format components output by tetrahedral microphones correspond to the acoustic pressure and velocity. Analogous results were obtained at distances of 0.1, 1 and 3 m: in all cases, good agreement was found with the theoretical curves.

The next step consists in measuring the radiation index η in frequency bands. The measurements have been performed at distances of 0.5, 1 and 3 m. The theoretical case has been calculated assuming that the field generated by the small loudspeaker is spherical, while the measured data have been obtained convolving the measured IRs with pink noise and applying third-octave-band filtering. As the distance and the frequency increase, the field tends to a progressive plane wave, the impedance tends to a constant real value and the radiation index tends to one. Figure 3.10 shows the results at different distances, in third-octave frequency bands. Up to 4 kHz, all transducers fit well with the expected result. The Soundfield yields a value that is 5 to 10 % lower than the others at all distances. No transducer is able to match the unity value at high frequencies: the Microflown rolls off starting from 4 kHz, while the tetrahedral microphones show a more abrupt decay from 8 kHz on. The reason for this behavior is two-fold. On one hand, in the case of the Microflown the SNR decreases with the frequency, so the presence of uncorrelated noise in the y and z channels causes a reduction of the radiation index, since noise is actually equivalent to the effect of adding diffuse acoustic reflections.

On the other hand, for tetrahedral microphones, due to the spacing of the capsules, spatial aliasing is observed above 8 kHz, which shows up as a deviation of the polar patterns from their ideal curve. In particular, the side-oriented figure of eight Y and Z channels pick up a considerable part of sound, while W and X reduce their value, therefore justifying the drop of the radiation index.

As a further characterization of transducers' behavior, let us consider now the difference in signal pick-up between the on-axis and off-axis velocity components. The transducers were placed with their x axes facing the

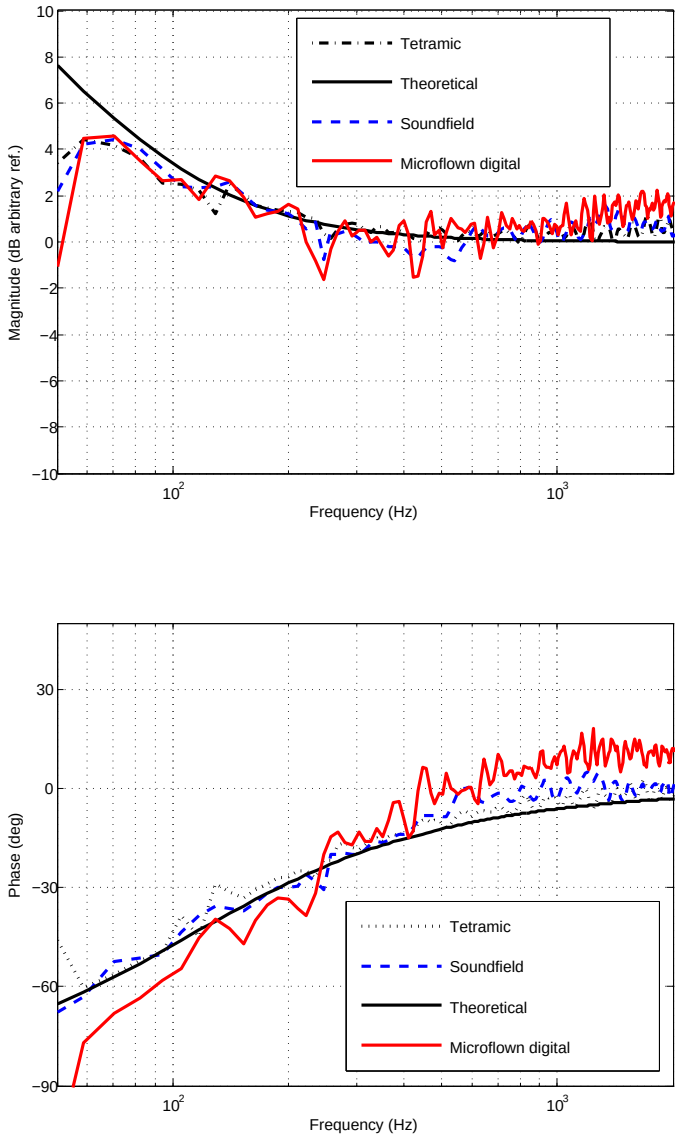

Figure 3.9: Amplitude (top) and phase (bottom) of the pressure-velocity transfer function measured with the three transducers 0.5 m in front of the loudspeaker, compared with the theoretical case.

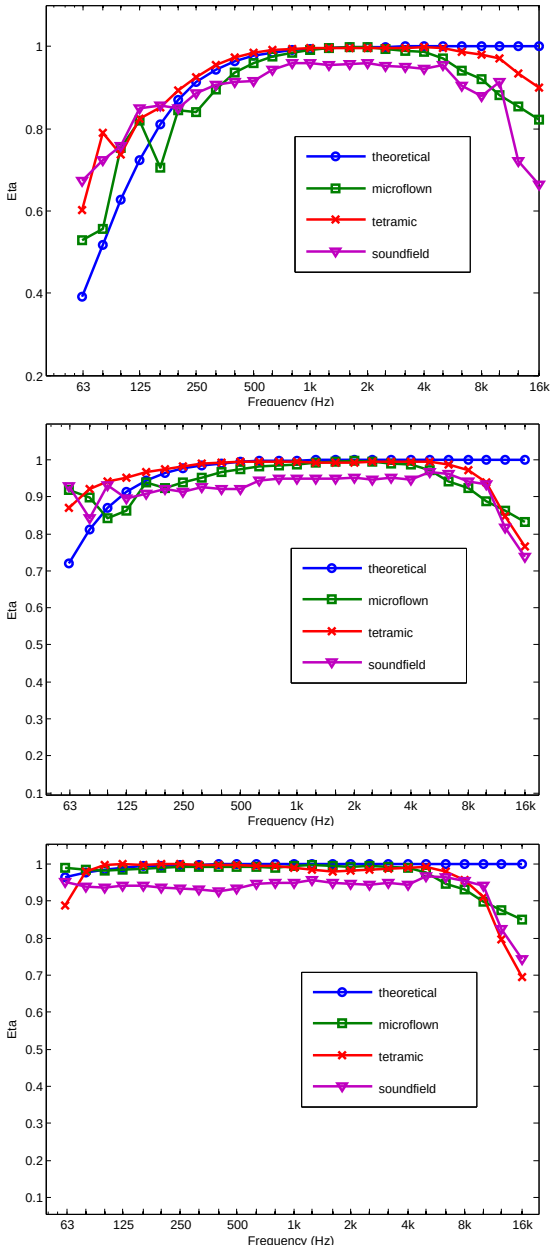

Figure 3.10: Radiation index for different distances, measured with the three transducers, compared with theoretical case. From top to bottom: 0.5 m, 1 m and 3 m.

loudspeaker at a distance of 1 m. The orientation was optimized by rotating each transducer to maximize the x component of the velocity. The analysis was done by convolving the measured IRs with noise and calculating the average RMS over a period of 1 s. The spectra of the X, Y and Z components for the three transducers are shown in Figure 3.11. The Microflown features the best rejection of orthogonal signals, with a difference of more than 20 dB from 300 Hz to 10 kHz. With the tetrahedral microphones, the difference is limited to approximately 15 dB, and greatly decreases below 100 Hz and above 10 kHz. As previously mentioned, the reason for the high frequency "crosstalk" is mainly the spatial aliasing due to capsule spacing, which causes deviation from the ideal figure of eight patterns.

The possible causes for the observed crosstalk are the following:

- Spatial aliasing at high frequencies in the tetrahedral microphones
- Uncorrelated thermal noise at high frequencies for the pressure-velocity probe
- Uncorrelated electric noise at low frequencies for all transducers
- Inaccuracies in the polar patterns at low frequencies for tetrahedral microphones
- Inaccuracies in the calibration and correction of the p-v probe at low and high frequencies

Table 3.2.3 reports the difference in dB, both A-weighted and un-weighted, between the on-axis and off-axis components measured with the transducers. This measurements show that the Microflown is more accurate in separating the components of the velocity.

	Microflown	Soundfield	Tetramic
$\Delta(X-Y)$	16.4	13.5	12.4
$\Delta(X-Y)$ A	20.8	16.8	15.9
$\Delta(X-Z)$	14.1	12.6	9.3
$\Delta(X-Z)$ A	20.7	14.1	13.8

Table 3.2: Side rejection of transducers compared, both unweighted and A-weighted. Values in dB. Higher values indicate better side rejection.

As a last step in the free-field comparison, the polar plots of the intensity vector on the horizontal plane are considered. The measurements were done with the microphones positioned 3 m in front of the loudspeaker, with the x axis aligned towards the speaker. The components of the intensity vector are obtained by multiplying the pressure and velocity components of the measured IRs. The vector is plot during a time interval of 20 ms, which

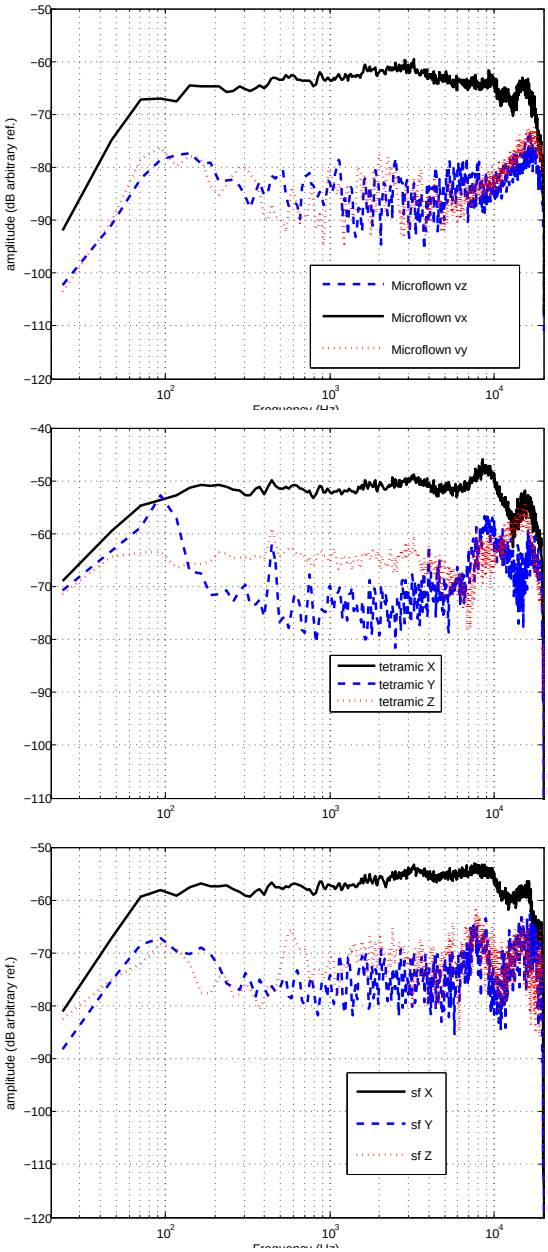

Figure 3.11: Side rejection for signals coming from the x direction. From top to bottom: Microflown, Tetramic, Soundfield. Ideal transducers would show no signal in the Y and Z components.

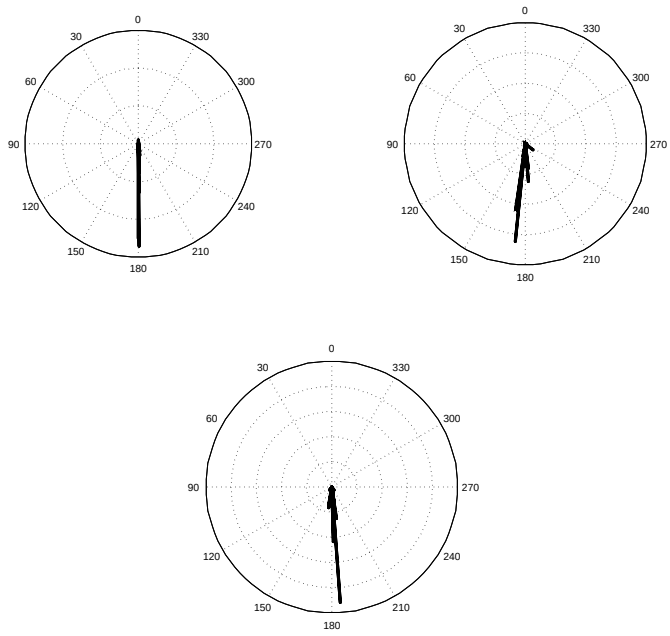

Figure 3.12: Polar plot of the projection of the intensity vector on the horizontal plane, broadband; anechoic condition, transducers 3 m in front of the source. Top left: Microflown; top right: Tetramic; bottom: Soundfield.

basically includes just the direct sound, given the anechoic conditions. In the ideal case, the intensity flows along the x axis pointing to the 180 degrees direction, since the energy propagates outwards from the speaker; this should result in a straight vertical line in the plot. The broadband plots of the intensity vector are shown in Figure 3.12; the Microflown is the most accurate in measuring the direction of the intensity vector. The deviations from the straight line measured with the tetrahedral transducers are directly related to the inferior side rejection examined in the previous section. The broadband accuracy of the tetrahedral transducers in measuring the direction of the intensity vector is approximately 10°. The analysis in frequency bands shows that the biggest inaccuracies and the "spread" of the intensity vector take place in the low and high frequencies. For example, as shown in Figure 3.13, in the third-octave band centered around 63 Hz there is a bias towards the right side, while in the band centered at 10 kHz the tetrahedral microphones show various spikes, indicating that different frequencies are encoded with a slightly different angle. At these frequencies, the Microflown still delivers accurate intensity measurements, thanks to the reduced crosstalk.

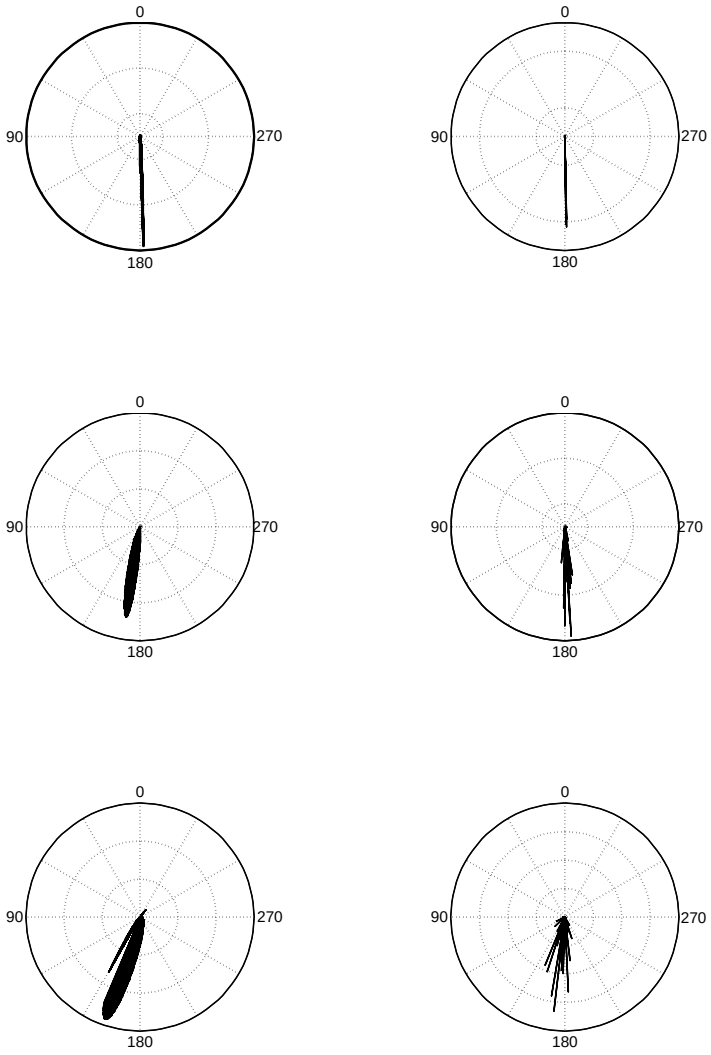

Figure 3.13: Polar plots in frequency bands; anechoic condition, transducers 3 m in front of the source. Left plots: 63 Hz; right plots: 10 kHz. From top to bottom: Microflown, Soundfield and Tetramic.

3.2.4 Comparison in reverberant environments

In this section, the Soundfield and the Microflown are compared in reverberant environments. Unfortunately, the Tetramic was left out of the comparison, since it was not available for this measurement. Our first measurements were done inside a reverberation chamber. Since the reverberation chamber was being used for absorption coefficient measurements, ten square meters of sound-absorbing foam were located on the floor in the center of the chamber when we performed the measurements. We took advantage of this configuration to see how the two transducers compare in measuring different positions of the field. The source used is a dodecahedral loudspeaker, placed near a corner of the room. The microphones were placed in the center of the room, approximately 1 m above the floor, and then in the opposite corner of the room. Near the corner, a diffuse field condition is expected, with low values of the radiation index and a scattered intensity, while in the center we expect part of the energy to flow towards the adsorbing material, without being bounced back, therefore leading to intensity vector plots with spikes toward the floor. In this case, the usual analysis of the radiation index did not show significant difference between the two points, therefore a one-dimensional approach has been adopted, by considering only the contribution of the vertical component of the velocity. The results, shown in Figure 3.14, indicate that this unidimensional radiation index is significantly higher in the center of the room starting from 1 kHz, the frequency where the foam is effective as sound absorber. Both transducers measure a similar curve, although discrepancies in the order of 50% occur between 500 Hz and 1 kHz.

An indication of the diffuseness of the field is given by the plot of the intensity vector during a period of a few tens of milliseconds. To evaluate the diffuseness and the effect introduced by the foam, we consider the plot of the intensity vector in a vertical plane, in this case the plane yz, for a duration of 80 ms of the IR, excluding the contribution of the direct sound. The results are shown in Figure 3.15. Both transducers detect the presence of a predominant energy flow towards the floor in the center position, while in the corner the radiation appears diffuse. However, the plots in the same condition are not really comparable, because the differences that were spot in Section 3.4 appear here spread over the whole angle and frequency range.

As a last comparison, the three transducers were used for the measurement of three-dimensional IRs in a variety of halls, including large spaces with reverberation time greater than 2 s. A dodecahedral loudspeaker was used for all the measurements. In this section, the resulting radiation index in third-octave bands is presented in four situations:

- a large empty corridor, with length greater than 100 m;
- a medium size hall;
- the same hall as above, with the source occluded by positioning the microphones in a coupled anteroom;

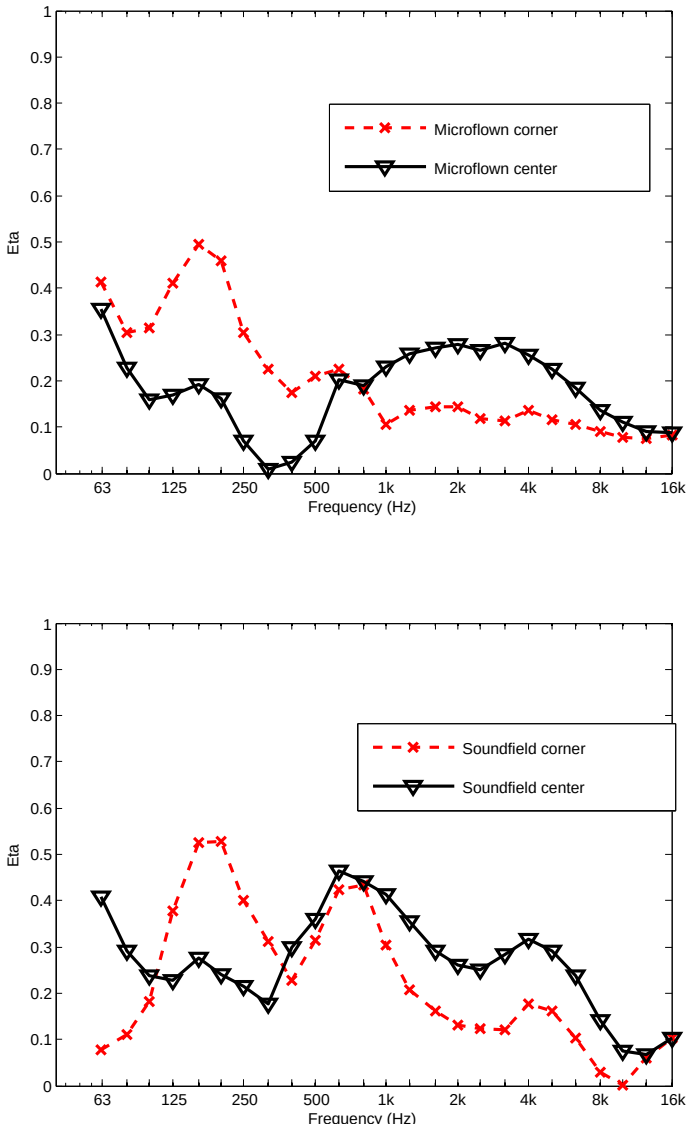

Figure 3.14: One-dimensional radiation index using the vertical component of the velocity, measured in third-octave bands in the center and corner of the reverberation chamber. Upper plot: Microflown; lower plot: Soundfield.

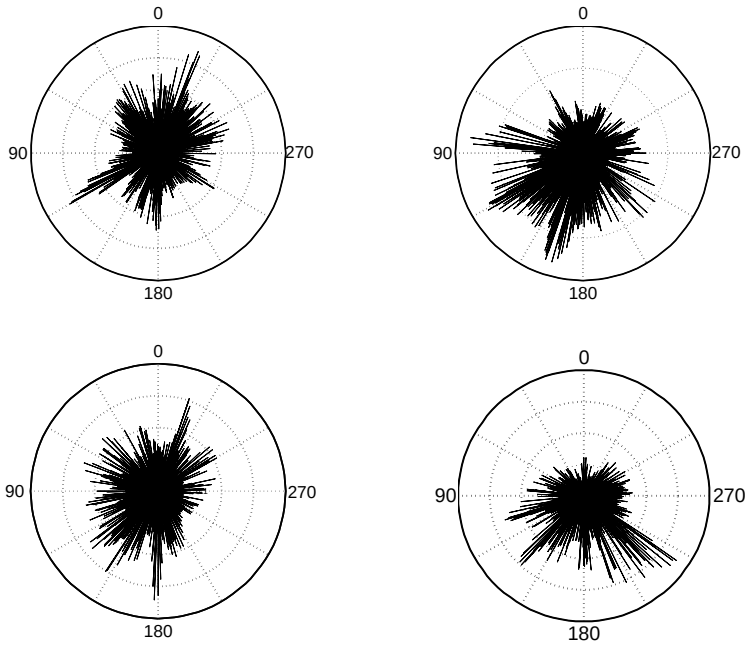

Figure 3.15: Polar plot of the projection of the intensity vector on the vertical plane yz. Top left: Microflown corner; top right: Microflown center; bottom left: Soundfield corner; bottom right: Soundfield center.

- a music studio with short reverberation time (300 ms) and diffuse decay.

The resulting plots are shown in Figures 3.16 and 3.17. The different transducers yield similar results; in particular, in each environment all the curves show the same peaks and dips, which are related to the presence of stationary waves corresponding to the distribution and overlap of room modes.

Bigger differences appear at high frequencies, namely above 5 kHz, where the results of the tetrahedral microphones are systematically lower than the anemometric probe. As discussed in Section 3.2.2, the behavior of tetrahedral microphones above 8 kHz is of doubtful accuracy, since effects of spatial aliasing compromise the polar patterns. For different reasons, discussed in Section 3.2.1, neither the Microflown results are accurate at such high frequencies. The Microflown was found to underestimate the radiation index in a radiating sound field; however one cannot conclude that the results in a reverberant field are underestimated. In order to draw a conclusion about which transducer is more accurate at high frequencies, the comparison should

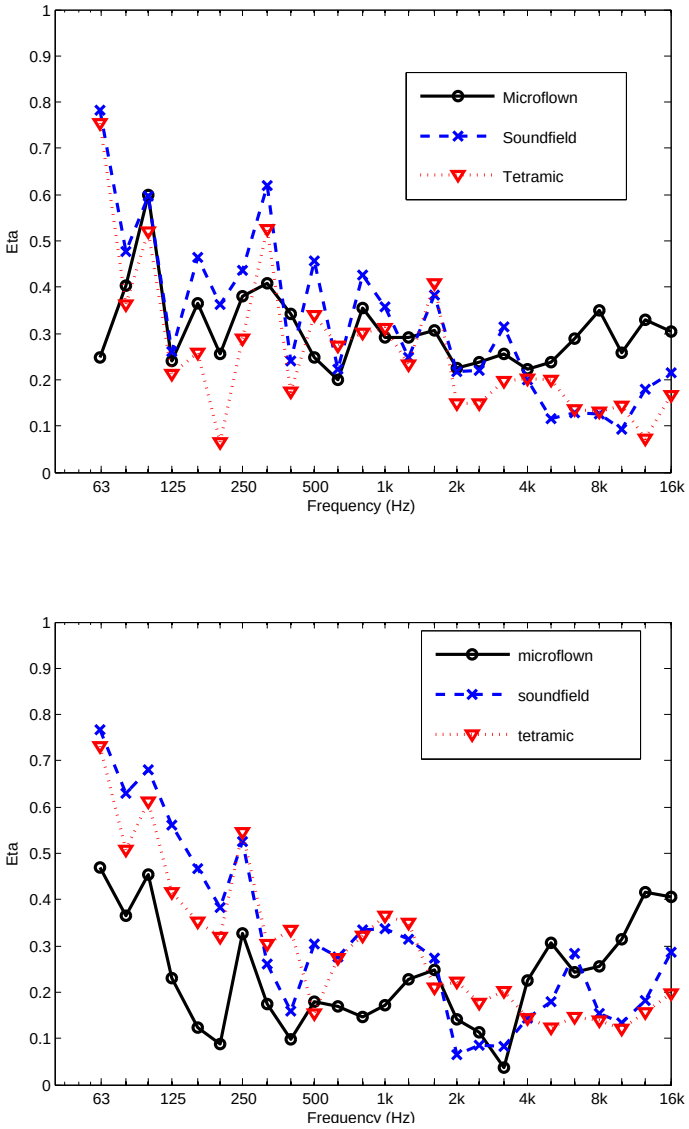

Figure 3.16: Radiation index in rooms measured with the three transducers; top: large corridor; bottom: medium studio.

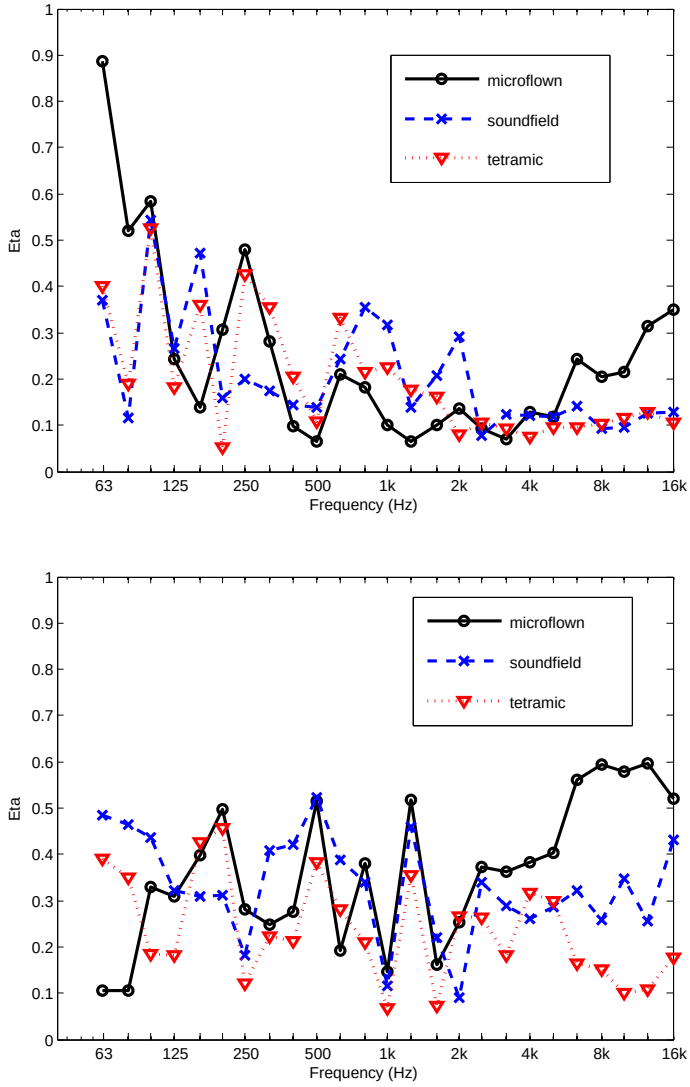

Figure 3.17: Radiation index in rooms measured with the three transducers; top: large hall (visible source); bottom: large hall (occluded source).

be repeated in a controlled field condition, possibly comparing with a p-p probe.

The measurements presented here serve as a comparison of tetrahedral microphones and anemometric pressure-velocity probes for applications related to sound intensity and the characterization of the spatial and energetic properties of acoustic fields. All transducers have proven suitable for sound intensity measurements, although some restrictions apply. Both topologies, if properly calibrated, capture the correct amplitude and phase relationship between acoustic pressure and velocity. For real time use and analysis, the tetrahedral microphones have a higher SNR and are therefore preferred. For IR measurements with the sine sweep method, the resulting SNR can be controlled by increasing the length of the excitation signal. The Microflown's velocity transducers have a better figure-of-eight pattern, which translate into a superior rejection over the whole spectrum for sounds coming 90° from the sides with respect to each velocity transducer, providing higher accuracy in determining the intensity vector and the energy flow paths; for audio applications, this might translate into a more accurate render of the acoustic environment after convolution with dry sounds, especially for the early reflections, which are known to be important perceptually. Tetrahedral microphones fail inevitably at high frequencies due to effects of spatial aliasing, while the Microflown is prone to the effects of noise. The Soundfield was found to slightly underestimate the radiation index in the anechoic measurements, although all transducer are accurate within 5% in the midrange. Thanks to its reduced size, the Microflown is certainly more suitable for measurements in small or difficult to reach spaces.

3.3 Second-order Ambisonics device with first-order transducers[2]

During recent years, prototypes of higher-order microphones have been built using pressure transducers located on a sphere [Bertet et al. (2006); Farina et al. (2007); Laborie et al. (2004)]. The higher-order signals are deduced by weighting each transducer with the projection of the given harmonic on its direction [Abhayapala and Ward (2002)]. Spherical microphone arrays require a relatively large amount of transducers. Because of the spacing between transducers, spatial aliasing affects the performance and the accuracy of the reconstructed harmonics at any order, as will be discussed later in this chapter. In spherical microphones, the effects of the body that holds the capsules is usually taken into account by applying a rigid sphere diffraction model and compensating its effects. One noteworthy exception is the approach used by Farina et al. (2011), who drop any model assumption, measure the responses of the spherical microphone in a large number

[2]This section is based on Cengarle et al. (2011)

Label	Polar pattern	Pressure derivatives
W	$1/\sqrt{2}$	$1/\sqrt{2}p$
X	$\cos(\theta)\cos(\varphi)$	$\partial p/\partial x$
Y	$\sin(\theta)\cos(\varphi)$	$\partial p/\partial y$
Z	$\sin(\varphi)$	$\partial p/\partial z$
R	$(3/2)\sin^2(\varphi) - (1/2)$	$(3/2)\,\partial^2 p/\partial z^2 - (1/2)\,p$
S	$\cos(\theta)\sin(2\varphi)$	$2\partial^2 p/\partial z \partial x$
T	$\sin(\theta)\sin(2\varphi)$	$2\partial^2 p/\partial y \partial z$
U	$\cos(2\theta)\cos^2(\varphi)$	$\partial^2 p/\partial x^2 - \partial^2 p/\partial y^2$
V	$\sin(2\theta)\cos^2(\varphi)$	$2\partial^2 p/\partial x \partial y$

Table 3.3: Representation of second-order Ambisonics signals in terms of polar patterns and sound pressure derivatives. θ is the azimuth and φ the elevation.

of directions and compensate the effects of the assembly and the capsules' responses by calculating a set of inverse filters from the measured data.

Thanks to the introduction of velocity transducers based on the principle of hot wire anemometry, described in Section 3.2.1, the acoustic velocity vector can be measured directly with high accuracy. Current state of the art probes employ three anemometric transducers and a pressure microphone that allow the simultaneous and coincident measurement of the sound pressure and the acoustic velocity vector, a set of quantities substantially equivalent to the first-order Ambisonics. As seen in the previous section of the chapter, these transducers are of small size, and three of them can be assembled orthogonally within less than 10 mm^3 to guarantee true coincidence at high frequencies and low diffraction and interference with the sound field. Moreover their polar response is consistent over the whole audio frequency range, unlike the patterns obtained with tetrahedral microphones. These features make them suitable for accurate spatial measurements and for calculating differences among closely spaced pairs. In the following sections the use of sound intensity probes to derive second-order harmonics starting from the Euler equation will be described.

3.3.1 Theoretical framework

All spherical harmonics can be written in terms of Cartesian coordinates on the unit sphere. Correspondingly the coefficients of the Bessel–Fourier expansion can be obtained by partial derivatives of the pressure field along the Cartesian axes [Cotterell (2002)]. Table 3.3.1 lists the signals of the second-order Ambisonics set in both representations, using the Furse–Malham normalization scheme [Malham (2003)].

The Euler equation, which relates the pressure gradient to time deriva-

tives of the velocity,

$$\vec{\nabla} p(\vec{r}, t) = -\rho_0 \frac{\partial \vec{v}(\vec{r}, t)}{\partial t} \tag{3.31}$$

can be used to express second-order derivatives of the pressure in terms of velocity gradients. For example, the second-order derivative of the pressure gradient along the x direction can be written as (dropping the explicit space and time dependence)

$$\frac{\partial^2 p}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial p}{\partial x}\right) = \frac{\partial}{\partial x}\left(-\rho_0 \frac{\partial v_x}{\partial t}\right) = -\rho_0 \frac{\partial}{\partial t}\left(\frac{\partial v_x}{\partial x}\right) \tag{3.32}$$

Thus obtaining the second-order spatial derivative of the pressure in the x direction requires differentiating over time the velocity gradient measured in the same direction. The last equivalence is valid as long as the pressure field and its derivatives up to second order are continuous functions of space and time, in which case the partial derivatives commute. This condition is valid unless a sound source is located at the point of interest. Following the same procedure used in Equation 3.32, the second-order spherical harmonics can be expressed in terms of velocity gradients as

$$\begin{aligned}
R_0 &= 1.5\partial^2 p/\partial z^2 - 0.5p = -\rho_0 1.5 \frac{\partial}{\partial t}\frac{\partial v_z}{\partial z} - 0.5p \\
S_0 &= 2\partial^2 p/\partial z \partial x = -2\rho_0 \frac{\partial}{\partial t}\frac{\partial v_x}{\partial z} \\
T_0 &= 2\partial^2 p/\partial y \partial z = -2\rho_0 \frac{\partial}{\partial t}\frac{\partial v_z}{\partial y} \\
U_0 &= \partial^2 p/\partial x^2 - \partial^2 p/\partial y^2 = -\rho_0 \frac{\partial}{\partial t}\left(\frac{\partial v_x}{\partial x} - \frac{\partial v_y}{\partial y}\right) \\
V_0 &= 2\partial^2 p/\partial x \partial y = -2\rho_0 \frac{\partial}{\partial t}\frac{\partial v_y}{\partial x}.
\end{aligned} \tag{3.33}$$

A zero subscript has been introduced when naming the harmonics to highlight the fact that these are unequalized quantities (equalization will be discussed later in this section). It is worth noting that different combinations of signals can be used to obtain the same quantities, a fact that will be exploited in Section 3.3.4, to design layouts with optimal arrangements of transducers. This property follows from the commutation of partial derivatives along orthogonal directions. As an example, the second-order x–y partial derivatives of the pressure field can be related to velocity gradients via either

$$\frac{\partial^2 p}{\partial x \partial y} = \frac{\partial}{\partial x}\frac{\partial p}{\partial y} = \frac{\partial}{\partial x}\left(-\rho_0 \frac{\partial v_y}{\partial t}\right) = -\rho_0 \frac{\partial}{\partial t}\frac{\partial v_y}{\partial x}, \tag{3.34}$$

or

$$\frac{\partial^2 p}{\partial y \partial x} = \frac{\partial}{\partial y}\frac{\partial p}{\partial x} = \frac{\partial}{\partial y}\left(-\rho_0 \frac{\partial v_x}{\partial t}\right) = -\rho_0 \frac{\partial}{\partial t}\frac{\partial v_x}{\partial y}. \tag{3.35}$$

The quantities appearing in Equation 3.33 still need some treatment before complying with standard Ambisonics coding and decoding conventions because of the strong filtering they exhibit. Consider a progressive plane wave incident from an arbitrary direction, with wave vector $\vec{k} = \hat{k}\omega/c$ and pressure field

$$p(\vec{r}, t) = Ae^{i\left(\vec{k}\cdot\vec{r} - \omega t\right)} = Ae^{i\omega\left(\frac{\hat{k}\cdot\vec{r}}{c} - t\right)}. \tag{3.36}$$

Spatial derivatives bring a factor $i\omega/c$ in front of the exponential, whereas time derivatives bring a factor $-i\omega$. All the second-order spherical harmonics calculated from the pressure field require a double spatial differentiation of the pressure, and thus exhibit a frequency response proportional to $-\omega^2/c^2$. Note that the only exception is the term linear in p appearing in the R harmonic. On the other hand the calculation of the spherical harmonics from velocity signals makes use of the velocity field corresponding to the plane wave of Eq. 3.36:

$$\vec{v} = \frac{p}{\rho_0 c}\hat{k} = \frac{A}{\rho_0 c}e^{i\omega\left(\frac{\hat{k}\cdot\vec{r}}{c} - t\right)}\hat{k}. \tag{3.37}$$

Thus, given the fact that all second-order spherical harmonics derived from velocity gradients require one time derivative and one spatial derivative, they exhibit a frequency response proportional to ω^2/c, corresponding to a high pass filter with a 12-dB per octave slope. Equalization is required in order to compensate this frequency-dependent gain, flatten the response, and allow combinations of spherical harmonics of different order. Following the same approach used in Cotterell (2002), where equalized pressure gradient signals are expressed in terms of time integrals, which correspond to low-pass filters, the equalized second-order signals can be expressed as time integrals of the velocity signals.

$$\frac{\partial^2 p}{\partial x^2}_{equalized} = \rho_0 c \int\int \frac{\partial^2 v}{\partial t \partial x} dt\, dt'. \tag{3.38}$$

In the case of the spherical harmonics we have, for example,

$$S_{equalized}(t) = \rho_0 c \int_0^t \int_0^{t'} S_0(t'')dt'dt''. \tag{3.39}$$

However, we notice from Equation 3.38 that one time integral undoes the time derivative, allowing to express the equalized term as

$$\frac{\partial^2 p}{\partial x^2}_{equalized} = \rho_0 c \int \frac{\partial v}{\partial x} dt'. \tag{3.40}$$

The practical consequence of this is that it is possible to skip the time derivative in the calculation of the spherical harmonics and only perform one time integration. This way the frequency response of the unequalized harmonic is proportional to ω/c, thus requiring a filter with a slope of only 6 dB per octave instead of 12 dB per octave. Considering that the second-order quantities are derived from first-order transducers, this is in agreement with the general rule that differential microphones require a filtering of 6 dB per octave for every increase in order [Elko (2000)]. The other equalized second-order harmonics are obtained following the same procedure, except for the R, where the term linear in the pressure does not have to be included in the integral. From now on we will drop the subscripts and always refer to the equalized second-order quantities.

In standard sound intensity measurements the pressure gradient is measured by means of two closely spaced pressure microphones, and the velocity is derived using a finite-difference approximation,

$$\vec{v}(r,t) \simeq -\frac{1}{\rho_0}\int_0^t \frac{p(r+d,t') - p(r,t')}{d} dt', \tag{3.41}$$

where d is the spacing between the two microphones. This approximation is only valid at wavelengths a few times larger than the separation between the microphones. An analogous approach can be used to derive the velocity gradient starting from the measurement of the velocity at two closely spaced points, a method that was already suggested by Olson (1946). For example, the second-order derivative of the pressure along the x axis can be expressed as

$$\frac{\partial^2 p}{\partial x^2} \simeq -\rho_0 \frac{\partial}{\partial t}\left(\frac{\Delta v_x}{\Delta x}\right) = -\rho_0 \frac{\partial}{\partial t}\left(\frac{v(x_2) - v(x_1)}{\Delta x}\right). \tag{3.42}$$

The method can be applied to the calculation of all the second-order spherical harmonics described in Equation 3.33. The use of a finite-difference approximation introduces limitations in the accuracy of the result. First of all it imposes an upper limit to the frequency range, because the expression is only valid as long as the distance between the two points in space is smaller than the wavelength of sound. A reduction in the spacing increases the upper limit of the usable frequency range, but reduces the SNR. The latter is due to the fact that, as the distance is reduced, the signals at both microphones become more similar and, therefore, their difference becomes smaller and more contaminated by the transducer noise.

3.3.2 Simulation of a second-order device

The behavior of a system constituted by pairs of closely spaced velocity transducers was simulated numerically to assess the effects of spatial aliasing and noise. The phenomenon of spatial aliasing affects the reconstruction of spherical harmonics when spaced microphones are used. It shows in terms of deviations of the polar responses from the theoretical curves (skewness) and

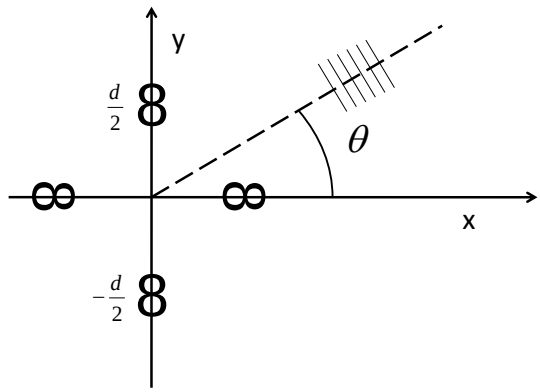

Figure 3.18: Simulated arrangement of velocity transducers equally spaced apart along the Cartesian axis in the presence of a monochromatic plane wave from direction θ.

the appearance of additional lobes, which are the result of the contamination of a given spherical harmonic with higher-order harmonics [Poletti (2005a); Daniel et al. (2003)].

To assess the effects of spatial aliasing and noise on the reconstructed harmonics, we ran a simulation of a second-order device using velocity transducers, following a procedure similar to the one described by Kolundzija et al. (2010). The simulated setup is shown in Figure 3.18. On the horizontal plane two ideal two-dimensional velocity probes are spaced apart a distance d on each axis. A monochromatic plane wave is used as input signal. The signal measured by the probes is simulated by applying the corresponding angle-dependent delay $t_{delay} = \pm d\cos(\theta)/2c$ on the x axis and $t_{delay} = \pm d\sin(\theta)/2c$ on the y axis. The signals at each probe are weighted with a cosine function to account for the bidirectional response of each transducer. The gradients are measured by taking the difference of the signals, and the spherical harmonics U and V are computed according to Equation 3.33. The polar plots are obtained by measuring the rms value of the harmonics over an interval of 1 s. The simulations were run with spacings of 20 mm and 50 mm, corresponding to the spacings that will be used later in the measurements, and an angular resolution of 5°. As reported in Table 3.3.1, the expected pattern of the spherical harmonic U is proportional to $\cos(2\theta)$, while the V one is proportional to $\sin(2\theta)$. These patterns have four maxima corresponding to full sound pickup and four directions where the signal pickup is completely absent. The resulting polar plots for the spherical harmonics U and V are shown in Figures 3.19 and 3.20. The plots are in decibels, normalized to an arbitrary value of 20 dB. The alias-free

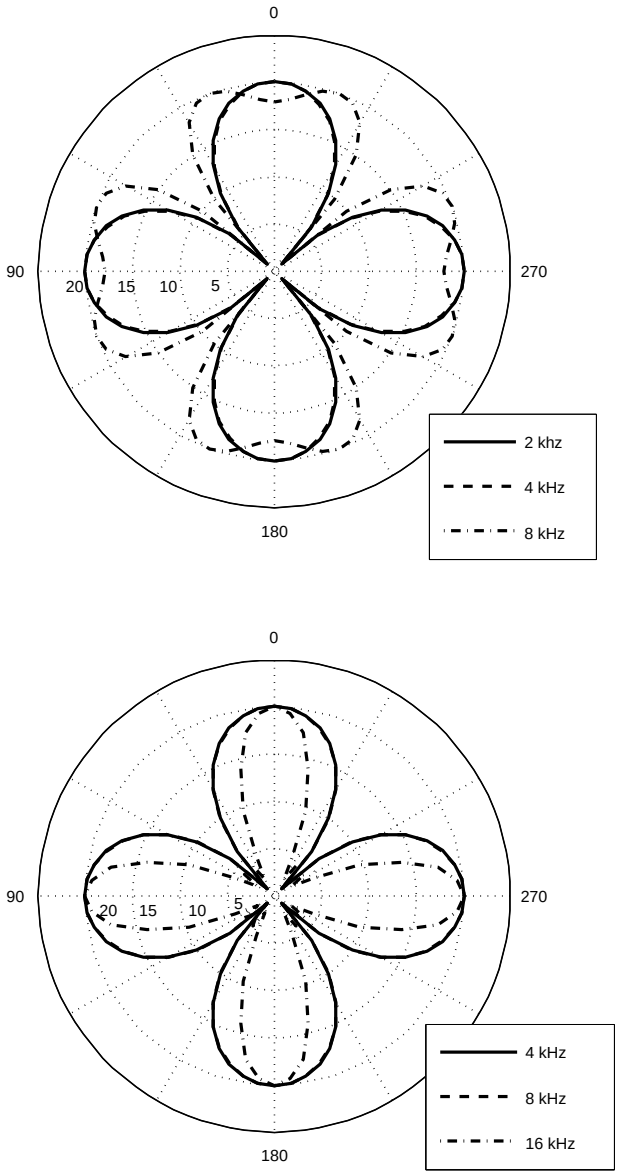

Figure 3.19: Results of simulation evidencing onset of spatial aliasing in the spherical harmonic U. Top: 50-mm spacing. Bottom: 20-mm spacing. Curves measured up to 4 and 8 kHz respectively differ by less than 1 dB from alias-free values.

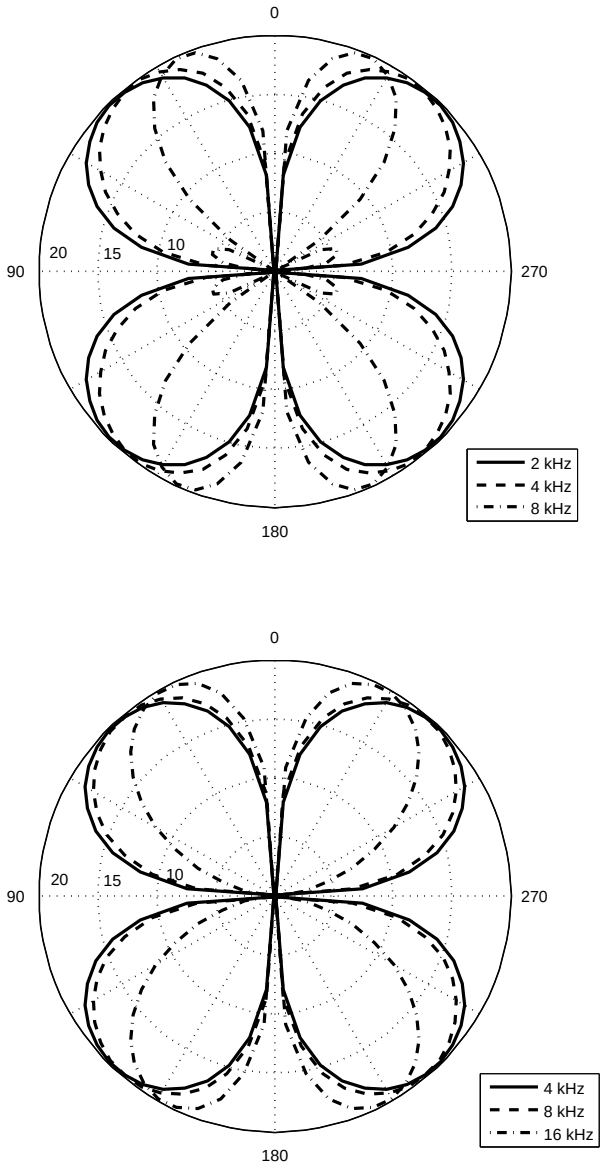

Figure 3.20: Results of simulation evidencing onset of spatial aliasing in the spherical harmonic V. Top: 50-mm spacing. Bottom: 20-mm spacing. Curves measured up to 4 and 8 kHz respectively differ by less than 1 dB from alias-free values.

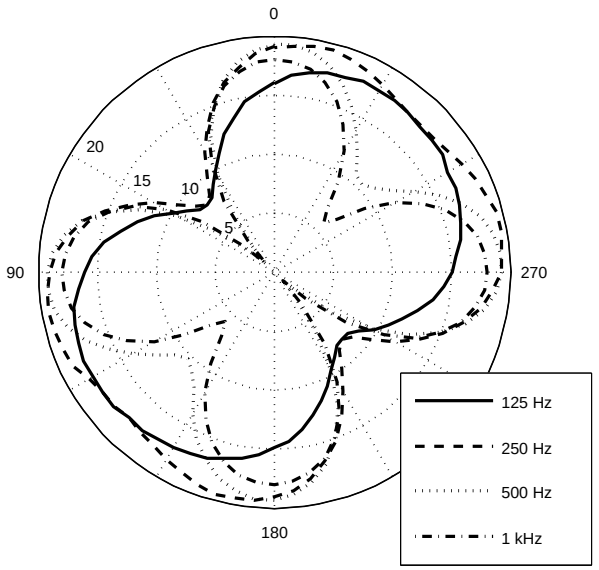

Figure 3.21: Polar patterns obtained for spherical harmonic U from simulation with 50-mm spacing, including noise at -20 dB with respect to signal and a sensitivity mismatch of 1.5 dB between probes on x and y axes. Resulting degradation appears in terms of a rotation of the position of the maxima and a reduction of the peak-to-minimum ratio. Response at lower frequencies tends to a first-order shape.

curves are not plotted as they overlap with the lowest frequency curve for each plot. If we consider the effects of aliasing acceptable as long as the spherical harmonic lies within 1 dB from the ideal, alias-free curve, then the plots show that for both harmonics a spacing of 50 mm is acceptable until nearly 4 kHz, while the spacing of 20 mm can be used beyond 8 kHz. The kind of higher-order contamination is slightly different for the two harmonics. However, the amount of discrepancy from the ideal values at any given frequency is similar. These results justify the rule of thumb according to which the differencing approach is valid as long as the spacing is smaller than half the wavelength. In ideal conditions, with noise-free transducers, the only limitation imposed by the finite spacing is an upper bound to the frequency range where the measurements fall within a desired accuracy: the smaller the spacing, the higher this upper bound. However, as mentioned previously, the presence of uncorrelated noise in the transducers worsens the performance toward the low frequencies, because the amplitude of the gradient is inversely proportional to the wavelength, thus the SNR will de-

crease. The SNR not only affects the general audio quality, but also the spatial resolution, giving rise to artifacts in the polar response. The larger the spacing, the lower the frequency where these artifacts appear. Phase and sensitivity matching between the transducers over the frequency range of interest are also critical for the accuracy of the results. To study the effects of noise and sensitivity mismatch, they have been included in the simulation, adding uncorrelated white noise in each transducer and changing their relative sensitivity as well. As an example, Figure 3.21 shows the results for the spherical harmonic U. Here noise was added with an amplitude of -20 dB referenced to the signal amplitude, also including a sensitivity mismatch of 1.5 dB between the two probes on the x axis and those on the y axis. The presence of noise reduces the difference between the peaks and nulls of the response, breaks the symmetry, and also gives the response a first-order-like shape, while the sensitivity mismatch causes the rotation, which is evident at 125 Hz. The plots confirm the increase of artifacts toward the lower frequencies. The choice of values for the amplitude of the noise and the sensitivity mismatch reflects the real-world behavior of the transducers, as discussed in Section 3.2.2.

3.3.3 Setup of a second-order device and measurement of its polar patterns

A device using two pressure-velocity sound intensity probes was set up and tested in order to derive the second-order harmonics. The functionality has been assessed analyzing the polar responses of the measured second-order components obtained in an anechoic chamber and comparing them to the theoretical values. The measured components were combined to produce second-order directional polar patterns. Two identical three-dimensional Microflown USP pressure-velocity probes were used in this experiment. They were set up in a face-to-face configuration, as shown in Figure 3.22. All the measurements were based on the impulse response (IR) technique using exponential sine sweeps as the excitation signal. The entire audio range (20 Hz to 20 kHz) was spanned during the measurements, leaving the possibility to limit the spectral band of interest by subsequent filtering at the data analysis stage. The sound source used is a Genelec 1030A studio loudspeaker. Every measurement was performed with two spacings of the probes, 20 and 50 mm, to evaluate the effect of the probe separation. The probes were installed on a rigid tripod with rotating head and protractor. The output of the probes was converted to digital by means of a MOTU 896 firewire interface and recorded on a PC. The data processing and analysis was carried out by means of Matlab routines.

The matching of frequency and phase responses of the transducers has a critical importance in the context of doing sums and differences of the signals. In order to obtain the desired matching, all the transducers were measured in a field condition of known impedance, using a spherical source

Figure 3.22: Two USP probes in face-to-face configuration. Protective caps limit minimum spacing to 20 mm, although they can be removed, allowing for a closer proximity.

calibration device, fitting the measured response curves according to the physical model that describes the functioning of the transducers and determining the parameters that characterize the response, to design suitable inverse filters for the compensation of each transducer's response and their matching, as discussed in Section 3.2.2. The inverse filters are stored as time domain impulse responses to be convolved with the signals to correct. The behavior of the probes in terms of frequency response and SNR improves significantly with the digital filters. However, like their analog counterparts, these are based on the fitting of the response obtained by the calibration and therefore do not take into account every possible discrepancy between the two probes, which might still exhibit mismatch at particular frequencies. To retrieve the spherical harmonics, the following steps were performed:

- Correction of each transducer's response by convolution with the corresponding inverse filters
- Calculation of the impulse response corresponding to the measurement position by convolution with the inverse sweep
- Removal of noise tail and harmonic distortion artifacts from the IR
- Convolution of the clean IR with the signal of interest on which to perform the operations: sine sweep, sine wave, or band-filtered noise
- Calculation of the spatial gradient by subtraction of the signals
- Multiplication by constants and combinations of signals as required by spherical harmonic formulas in Equation 3.33

- Equalization of second-order harmonics to compensate the low frequency attenuation effect described in Section 3.3.1

The retrieval of impulse responses from single transducers before combining the signals allows to greatly improve the SNR and to remove unwanted distortion artifacts due to loudspeaker and probe non-linearities [Farina (2000)]. The subsequent convolution with a sine sweep or band-filtered noise is equivalent to having measured that same signal with a harmonic-distortion–free signal chain with better SNR. This strategy was required to reduce the undesired effects of the transducer's self noise described in Section 3.2.1. This SNR has limitations for audio applications that require direct processing of real-time acoustic data.

The 50 mm spacing was used to obtain an SNR sufficient for a difference of about 10 dB between maxima and minima of the polar response at 500 Hz, as inferred from the simulation. Such a spacing, however, limits the higher bound of the usable frequency range to approximately 3.5 kHz. To further extend the range, all measurements were repeated with a 20 mm spacing, the smallest distance allowed by the mechanical design of the probes. This smaller spacing allows to extend the range up to 8 kHz. The equalization to compensate the second-order high-pass behavior (discussed in Section 3.3.1) was implemented in the frequency domain, taking the Fourier transform of the second-order signals and dividing each frequency component by the frequency itself to flatten the response. All the physical constants and the spacing were taken into account in order to keep the sensitivity of the different orders matched as well.

The measurements to verify the angular dependence of the spherical harmonics have been carried out in a small anechoic chamber (4.7-m length, 3.1-m width, and 2.8-m height, decay time RT30 < 0.1 s), where the arrangement of the two probes was set 2.2 m in front of the loudspeaker, at the same height as the acoustical center of the latter. The two facing probes were rotated in the horizontal plane with steps of 15°. This allows the measurement of the two spherical harmonics U and V, which do not require gradients over the vertical direction, as shown in Equation 3.33. Measuring the V spherical harmonic requires computing the difference of two lateral velocity sensors measuring v_y displaced along the x direction. However, measurement of the U component would require the addition of another transducer measuring v_y displaced along the y axis. Since a third probe was not available, the same arrangement rotated 90° was used to obtain the desired signal. Ideally the combination of non simultaneous measurements is allowed by the linear time-invariant characteristic of the sound field. Such implementation, however, does not take into account the mutual interference that would be caused by a third probe displaced laterally. The effects of such interference are expected to appear at wavelengths comparable to the dimensions of the probes, which corresponds roughly to frequencies above 5 kHz. The polar patterns that were measured for the spherical harmonics U and V are shown

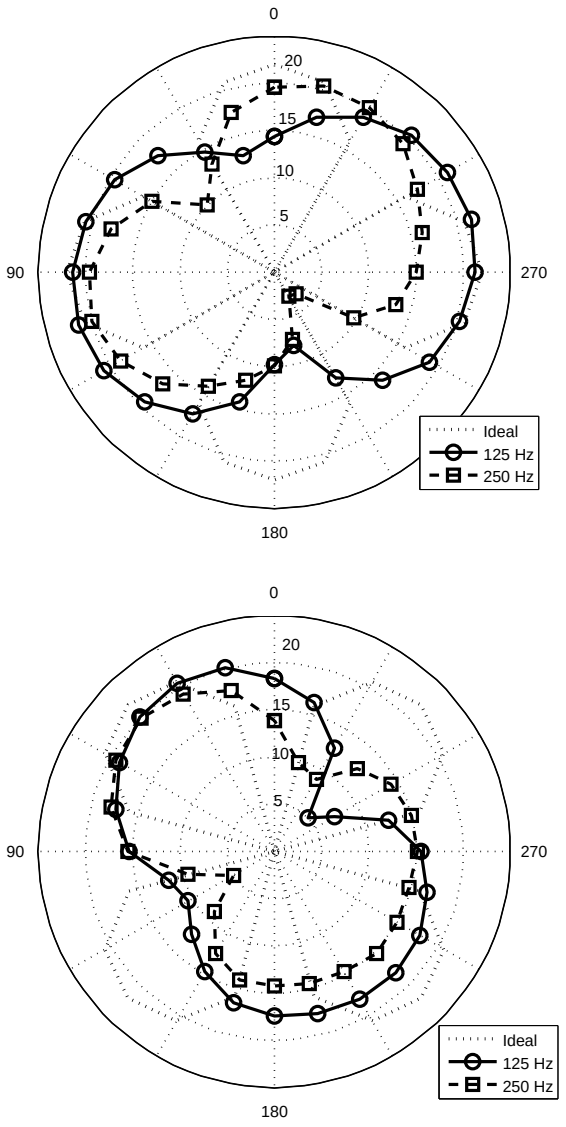

Figure 3.23: Polar plots of U and V spherical harmonics at low frequencies. Top: ideal curve, 125 Hz and 250 Hz octave bands for spherical harmonic U. Bottom: ideal curve, 125 Hz and 250 Hz octave bands for spherical harmonic V. Plots show poor system performance at low frequencies due to a small SNR, in agreement with simulation results.

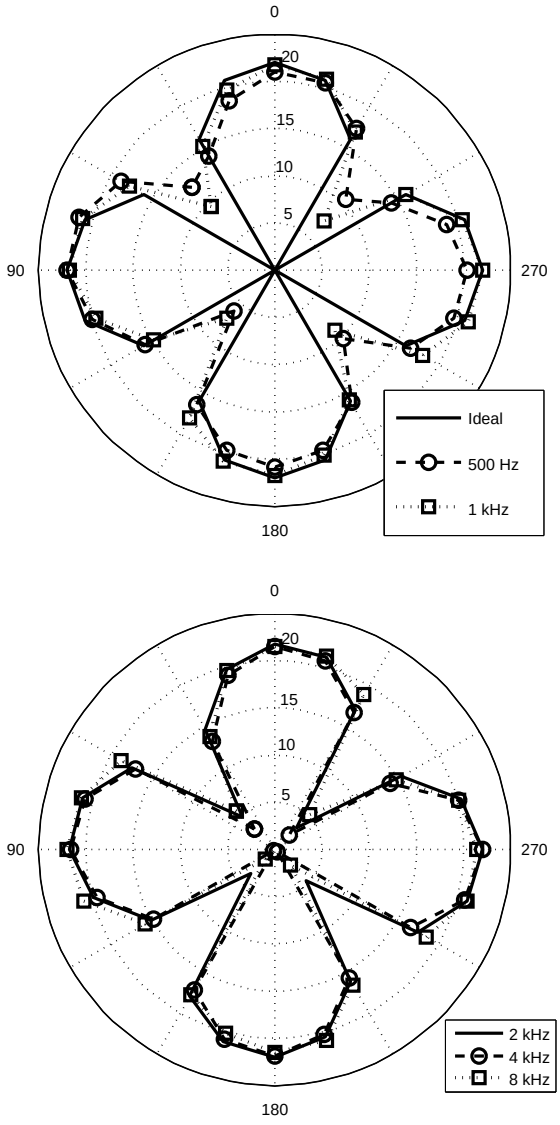

Figure 3.24: Polar plots of U spherical harmonic in frequency bands. Top: ideal curve, 500 Hz and 1 kHz octave bands. Bottom: 2, 4, and 8 kHz octave bands. Values are expressed in dB with arbitrary reference. Measured values follow expected shape; peak-to-minima ratios vary from 7 to 18 dB in different frequency bands.

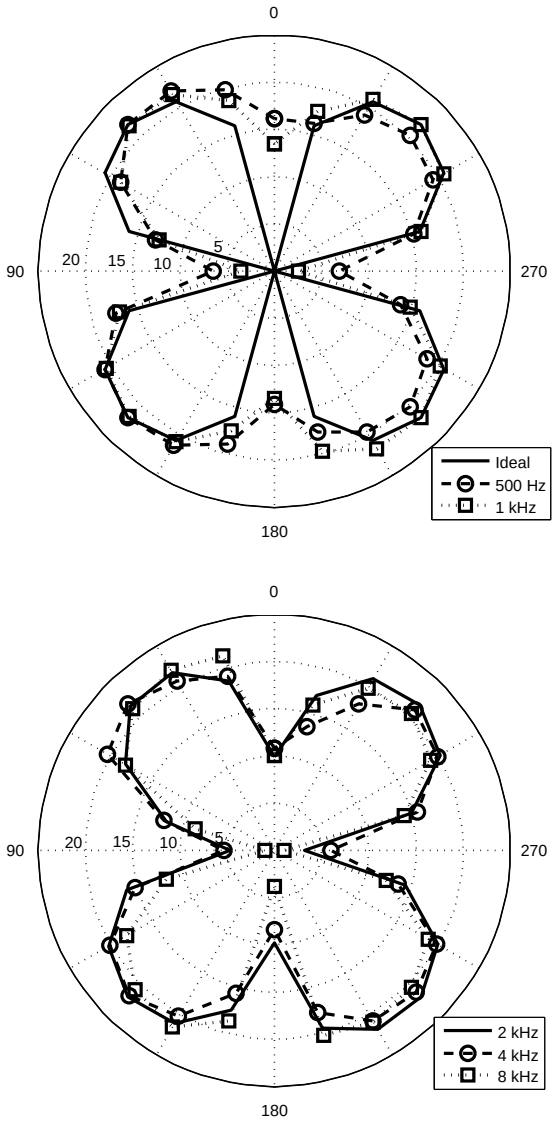

Figure 3.25: Polar plots of V spherical harmonic in frequency bands. Top: ideal curve, 500 Hz and 1 kHz octave bands. Bottom: 2, 4, and 8 kHz octave bands. Values are expressed in dB with arbitrary reference. Best results are obtained above 1 kHz, with peak-to-minima ratios in the order of 10 dB.

in Figures 3.23, 3.24 and 3.25, together with the ideal curves. Spacings of 50 and 20 mm were used below and above 4 kHz respectively. The plots show good agreement with the theoretical values, especially above 1 kHz, where the attenuation between maxima and minima is in the order of 20 dB. The system exhibited limitations below 500 Hz, revealed by a significant deviation from the expected pattern. As revealed by the simulation and previously discussed, this effect is caused by the presence of uncorrelated thermal noise in all the transducers, together with the fact that the signal amplitude obtained after taking the gradient is small due to the proximity of the transducers in each pair, thus falling below the noise floor for those directions where minima are expected. In particular we observe that the patterns below 500 Hz are compatible with the results of simulation, indicating that noise and a possible mismatch of the probes are a plausible cause for the inaccuracies of the results. In order to improve this behavior, a larger spacing such as 100 mm should be used. To quantify the discrepancy between ideal and measured values, the frequency- and angle-dependent error has been calculated according to the following formula:

$$Er(f,\theta) = 20 \times \left| log_{10} \left(\frac{m(f,\theta)}{i(\theta) + \epsilon(\theta)} \right) \right|, \tag{3.43}$$

where $m(f,\theta)$ is the measured value and $i(\theta)$ is the ideal one, which does not depend on the frequency. Both values are normalized to have a maximum value of 1 on the main axis. ϵ is a regularization parameter, which avoids a division by zero when $i(\theta)$ is very small; it was set to 10^{-2} in these experiments. This choice is equivalent to limiting the maximum attenuation of the ideal second-order patterns to two orders of magnitude, equivalent to 40 dB. Figure 3.26 shows the frequency-dependent error of the U and V harmonics for four different orientations of the sound source, measured with the 50 mm spacing. The values for each frequency band have been obtained performing a one-third-octave band filtering. The plots show that for every harmonic the error is larger in the direction of minimum sensitivity, and in every direction it increases at low frequencies. The plots indicate an increase of the error below 250 Hz, suggesting that this frequency represents the lower bound of the range where the setup can provide accurate second-order performance.

After obtaining the second-order harmonics it is possible to combine them with the pressure and the first-order components to obtain polar patterns with a higher directivity index. Two examples were examined:

- the second-order figure of eight, whose polar pattern is

$$S(\theta) = \cos^2\theta = \frac{1 + \cos 2\theta}{2} = \frac{P + U}{2}. \tag{3.44}$$

- The second-order cardioid, expressed as

$$S(\theta) = \frac{(1 + \cos\theta)(1 + \cos\theta)}{4} = \frac{1 + 2\cos\theta + \cos^2\theta}{4}. \tag{3.45}$$

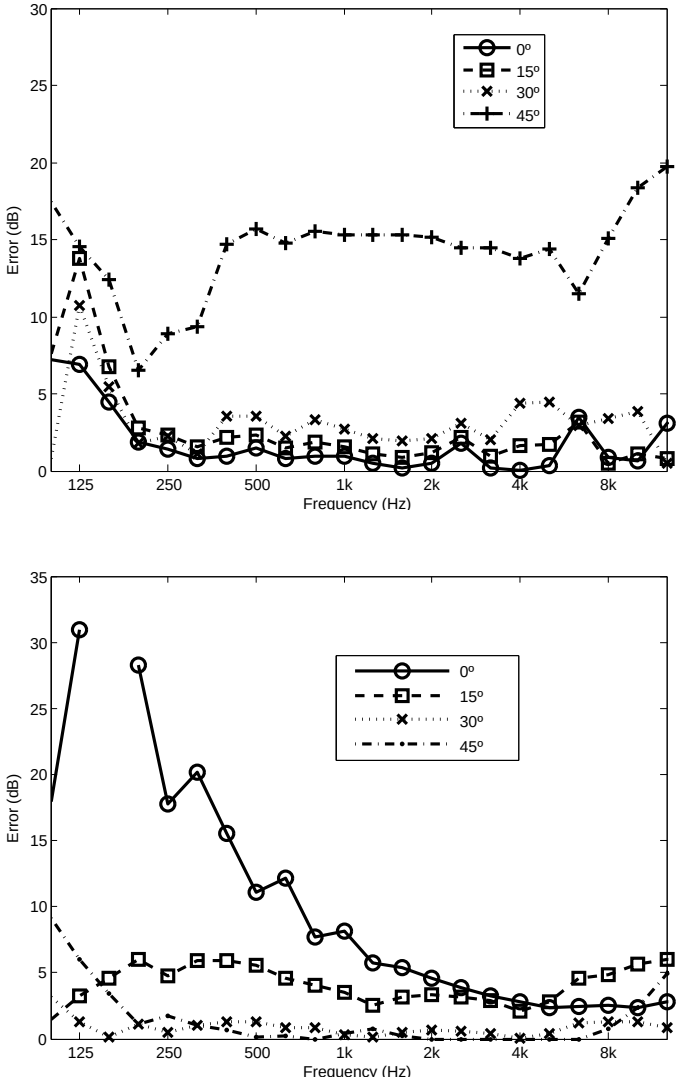

Figure 3.26: Error analysis, reporting discrepancy in dB between ideal and measured values of polar pattern in one-third-octave frequency bands using 50-mm spacing. Top: spherical harmonic U. Bottom: spherical harmonic V. Plots show that largest errors occur in the direction of minimum pickup and tend to increase toward low frequencies.

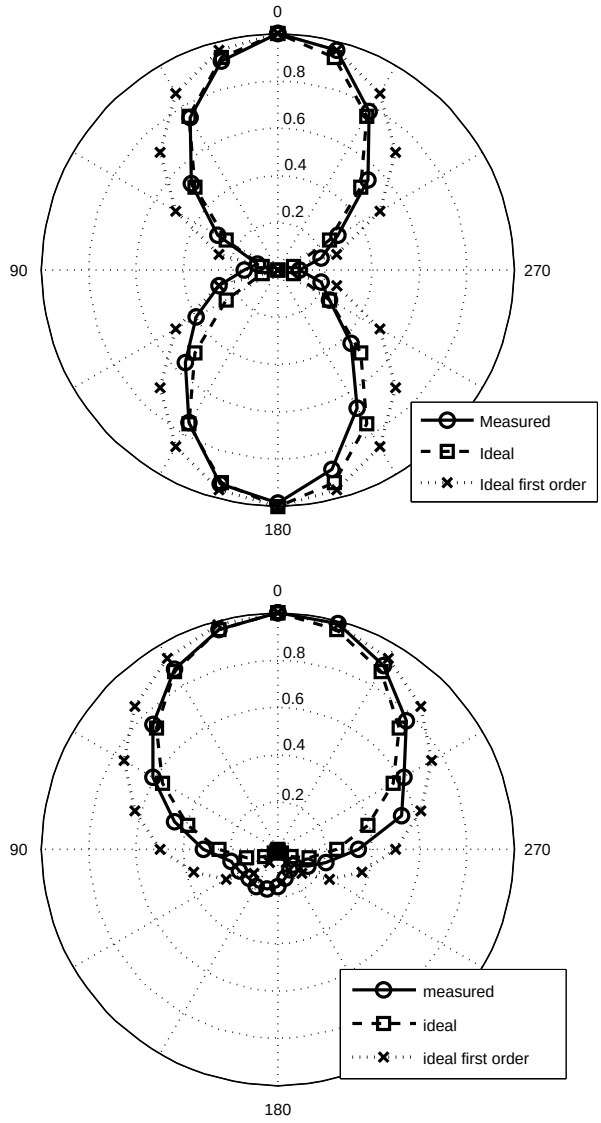

Figure 3.27: Polar patterns of second-order directional responses measured at 2 kHz and comparison of ideal and measured responses at first and second order. Top: figure of eight. Bottom: cardioid.

Using Equation 3.44 this pattern can be written in terms of spherical harmonics as follows:

$$S(\theta) = \frac{1}{8}\left(3 + 4\cos\theta + \cos 2\theta\right) = \frac{1}{8}\left(3P + 4X + U\right). \tag{3.46}$$

Both patterns are characterized by narrower lobes than their first-order counterparts. The patterns obtained are shown in Figure 3.27 together with the theoretical curves of the corresponding first- and second-order responses. These plots show the responses in linear scale that have been obtained in the one-third-octave frequency band centered around 2 kHz. The quality of the results obtained depends on the accuracy of the second-order components. For this reason the best results were obtained in the frequency band between 1 and 8 kHz, while outside these bounds the irregularities of the measured second-order harmonics caused severe deviations from the expected curves.

3.3.4 Proposal of a complete second-order Ambisonics device

Let us apply the approach presented in this chapter to the theoretical design of a second-order device based on velocity transducers. In order to measure the complete set of spherical harmonics up to order two, it is necessary to obtain the quantities reported in Table 3.3.1, which can be expressed in terms of the gradient of velocity, as shown in Section 3.3.1. In the approach described, these quantities are measured by means of pairs of anemometric velocity transducers. Since, as shown in Section 3.3.1,

$$\frac{\partial}{\partial t}\frac{\partial v_i}{\partial x_j} = \frac{\partial}{\partial t}\frac{\partial v_j}{\partial x_i},\ i,j = 1,2,3, \tag{3.47}$$

a pair of sensors measuring the velocity gradient along one axis can be replaced by another pair along a different axis, allowing for an optimization of the transducers used in the design. Figure 3.28 shows a particular configuration among all the possible ones capable of obtaining all the velocity gradients with a single measurement. It requires nine velocity transducers, plus one pressure transducer for the zeroth-order component. R is obtained by the two vertical velocity transducers displaced along the z axis; S uses the two x transducers along the z axis, and T the two z components displaced along the y axis. U requires the pairs of x and y transducers along the x and y axes respectively, while V uses the two y transducers displaced along the x axis. Equation 3.47 can be implemented in the design, anyway its further use would not reduce the number of sensors, but only change their arrangement. Note that, in particular, it is not possible to design a layout that does not require displacing sensors along three different axes in an attempt to make the device more compact. Although the authors did not have such a number of velocity transducers at their disposal, nonetheless a full set of second-order components was obtained in a room by repeating three times

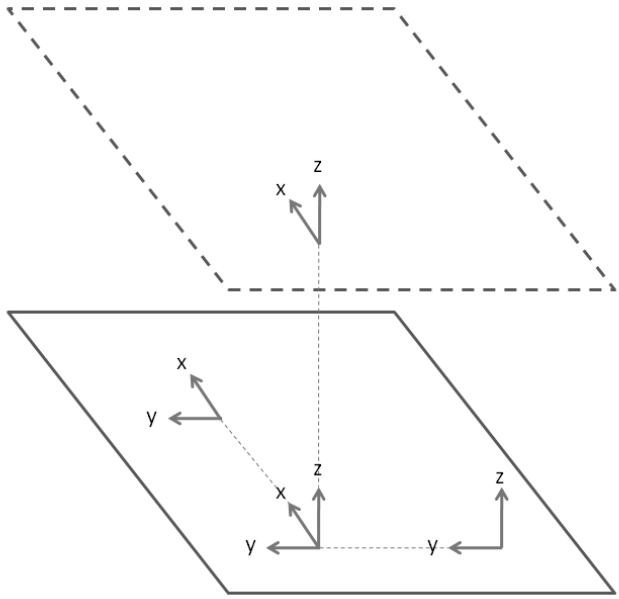

Figure 3.28: Draft of possible transducer layout for full three-dimensional second-order device.

the impulse response measurement with the probes in face-to-face configuration, aligning the probes each time along a different Cartesian axis, a legacy from the practice of three-dimensional sound intensity measurements with a single p–p probe.

The approach presented here allows the retrieval of full second-order components with nine velocity sensors plus a pressure microphone, a total of ten signals. The location and orientation of the transducers along the Cartesian axes facilitates the construction of the device. A recent proposal of a device for second-order recording using figure-of-eight microphones makes use of at least twelve velocity sensors in a dodecahedral arrangement plus additional transducers for the sound pressure [Craven et al. (2009)], while other proposals using pressure sensors on a sphere require the use of twelve transducers in a dodecahedral configuration [Gerzon (1973); Cotterell (2002)]. The anemometric velocity transducers are less than 6 mm long and 2 mm wide, and they do not require bulky electronics, such as transformers, in close proximity. Therefore many of them can be assembled in a closely spaced configuration. The main factor determining the dimensions of the device is the spacing between the transducers. A fixed spacing of 50 mm would allow measurements up to nearly 4 kHz, while including two pairs of transducers with a different spacing for each direction would extend the frequency range,

although with the inconvenience of using more channels or having to repeat measurements. Such an approach is further encouraged by the recent development and test of a process to incorporate various anemometric transducers in different orientations on a single wafer [Yntema (2008)].

3.4 Conclusions

In the first part, the concepts related to sound intensity were introduced, focusing in particular to the sound intensity vector and the radiation index. The sound intensity vector identifies the direction and magnitude of energy transfer in a sound field, and therefore includes information about the spatial properties of sound. The radiation index expresses the fraction of energy transfered by the acoustic field in a certain position and is useful for discriminating diffuse fields and stationary waves, where the energy is oscillating, from conditions where the energy flows along one direction, such as in progressive plane waves. Both the intensity vector and the radiation index depend on the phase relationship between pressure and velocity, therefore accurate transducers are required for their measurement.

In the second part of the chapter, anemometric transducers and tetrahedral microphones have been described and characterized in terms of their technical specifications and spatial resolution. The calibration procedures for the anemometric transducers have been discussed, and their limitations have been documented. The comparison between transducers has been carried out comparing their output in controlled conditions. The analysis was focused on the accuracy of the phase and amplitude response in sound fields of known impedance, evaluated by comparison of the intensity vector plots and radiation index values obtained with each transducer. The research evidenced that both transducers topologies provide the acoustic pressure and velocity signals with varying accuracy: the indirect measurement provided by tetrahedral transducers is affected by inaccuracies in the polar plots and in their variation within the frequency range. On the other hand, the direct measurement obtained with the anemometric transducers and the Microflown probes features accurate polar patterns and stable performance in the whole frequency range, but the low SNR affects the accuracy of results especially at high frequencies. While the anemometric transducers are not suitable for direct recordings, they turn out to be more accurate than tetrahedral microphones for measurements with techniques that improve the SNR.

In the last part of the chapter, an approach is proposed to obtain second-order components of the spherical harmonics expansion of the pressure field by means of velocity transducers based on hot-wire anemometry. Second-order partial derivatives of the pressure field can be expressed in terms of the gradient of air particle velocity, using the Euler equation of fluid dynamics. The approach was both simulated numerically and implemented

using pairs of anemometric velocity transducers in face-to-face configuration. Measurements performed in an anechoic chamber allowed to validate and test the accuracy of the device via analysis of the polar patterns of the reconstructed harmonics and comparison with the theoretical responses. Second-order components of the impulse response were used to synthesize second-order patterns corresponding to virtual directional microphones. One of the advantages of this technique is that the required layout of transducers can be defined easily and does not involve the problem of choosing a sampling of the sphere, as required when using pressure transducers for higher-order Ambisonics. The velocity transducers are small compared to acoustic wavelengths and they can be mounted on tiny rigid bars only a few millimeters wide, thus reducing the effects of wave diffraction and acoustic shadows. The zeroth- and first-order components alone, that is, the sound pressure and the three orthogonal components of the velocity, can be measured directly instead of being obtained by a combination of signals, by just considering four coincident transducers among the eight (or more) employed. Taking this subset of signals, true spatial coincidence with the second-order harmonics is lost, but a higher accuracy is obtained on the zeroth and first order, since they are not affected by crosstalk and higher-order contamination caused by spatial aliasing which typically appears at high frequencies when any spherical harmonic is obtained by many transducers spaced apart. For this reason the device can also be used directly for accurate standard sound intensity measurement purposes. However, for the second-order components the finite-difference approach brings limitations in the frequency range, because a close distance between sensors reduces the SNR at low frequencies, while a larger separation limits the applicability at high frequencies. The use of Microflown transducers, characterized by a low SNR, makes the device too noisy for direct recordings and limits its use to the measurement of impulse responses. One remark has to be made in this respect: as can be gleaned from the experimental results, the characteristics of the transducers that were used represent the limiting factor in the performance of the system. However, the results presented here serve as a validation of the proposed operating principle and as an assessment of the relationship between the accuracy of the results and the mismatch of the transducers, implying that the proposed device would benefit from an improvement in transducer technology. As regards possible future developments, the first step is the improvement of the performance in terms of SNR and frequency response. This can be addressed by selecting and combining two suitable spacings, with the goal of extending the usable frequency range and the regularity of polar patterns from 100 Hz to 10 kHz.

In the author's opinion, the second-order approach presented here is going to be surpassed in terms of performance and practicality by recent advances in spherical microphone array technology combined with availability of DSP processing. Nevertheless, due to the low number of transducers required, their small size and the intrinsic accuracy at first order, the proposed method

may find application in contexts where audio quality is not the priority.

4 Production and post-production

Audio post-production comprises all the steps that lead from the recorded material to its delivery to the audience in a final format. Editing and mixing are by far the most acknowledged steps, although many others such as mastering and authoring gain equal importance depending on the kind of production. The steps, procedures and techniques employed in post-production differ in the details between movie, music and broadcast applications, although many concepts are shared. One of these concepts is the choice of the destination format: once it has been established, the whole work-flow is adapted to it. For example, the panning tools depend on the number of channels and their layout, so doing a panning in stereo or in 5.1 requires a different algorithm, interface and range of controls. Dealing with an increasing number of output channels inevitably adds complexity in all stages; for example, the coverage of a live event with horizontal surround or 3D surround implies increasing the number of microphones and having more channels to control during the mixing stage. If the concept of channel based work-flow is maintained, then every 3D surround production will take a time proportional to the number of required output formats. The shift of paradigm from channel-based audio to channel-free, mentioned in Chapter 2, is a breakthrough toward simplification of the post-production practice, because it allows to make the whole work-flow unbound to the channel configuration.

This chapter presents the results of the research for facilitating the independence from the number of channels; it is divided into two parts: live broadcasting and cinema post-production. Traditionally, broadcasting has always adopted the available standard formats, for example stereo and 5.1. Now that 3D audio is gaining increasing importance and recording/playback systems are becoming available, it seems inevitable that 3D sound will land in the broadcasting market in the very near future. In order to make the transition as smooth as possible, simplicity and ease of use are the key. In case of live events, in particular sport events, the spatialization of the sound elements has to become as simple as possible, and possibly automatic. Our research takes a step along this direction: we present an algorithm and tool to control the levels of multiple microphones in real time and obtain a mix that maximizes the pick-up of sound from a desired point, given the po-

sition of the microphones; this idea has been developed in the context of surround sound for sport events and applied in the live mixing of football games. Regarding post production for multichannel audio, in the second part of the chapter we introduce a problem inherent in layout-independent audio production: the unknown signal levels at generic loudspeaker layouts, where the summing of signals during decoding can potentially give rise to clipping in some output channels; the proposed solution is an algorithm that dynamically searches for potential clipping using a suitable worst-case layout, so that absence of clipping in the worst-case layout implies absence of clipping in the others. The algorithm has been tested in *ad hoc* and realistic scenarios.

4.1 Assisted mixing of sport events[1]

In the recent years, the live capture of sound for the broadcasting of sport events has gained increasing importance and attention, thanks to the growth of technology and the desire to offer new features and an improved audio experience. The advent of surround sound in broadcasting has pushed for the evolution of microphone setups, consoles and mobile units to integrate 5.1 surround in an easy and practical way.

From the audio point of view, there are two main components that engineers want to capture in sport events: one is the sound of the action, produced by the players, that enhances the visual experience and brings the spectator closer to the game; the other is the sound of the crowd, that gives the impression of being part of the audience, especially when surround techniques are employed. In most cases, the sound of the action is the most difficult to capture, because of its relatively low level compared to the crowd sound and because of the practical impossibility of a close-miking approach. The typical situation for the sound coverage of a sport event is therefore to locate many directional microphones as close as possible to the field, to capture the action of the game, while additional microphones are added for the ambience and crowd sound. The following considerations are related to the specific case of sound recording of football games, although they apply to many other sports as well. When the game takes place in a vast area, such as in the case of football, the action happens in a region which is relatively small compared to the size of the field. A dozen of microphones are located at the side of the pitch, aiming towards specific areas of the field. Many limitations affect the positioning of the microphones, in particular: they cannot interfere with the movement of the player or constitute potential danger for the players; they shall not be positioned close to the benches, where they could pick up strategic or "confidential" information; last but not least, they cannot visually obstruct advertisement.

[1] This section is based on Cengarle et al. (2010)

In the current mixing procedure for football broadcasting, the work done by the sound engineer consists in rendering the sound of the action by raising only a few faders of the audio console at the time - or in some cases just one - while following the game on the video monitor. These faders correspond to the microphones that are closer to or aiming at the action. This procedure aims at isolating as much as possible the sound of the action from other undesired sounds. The levels of the ambience microphones for the crowd remain almost static during the whole game, provided a good balance is obtained at the beginning. Having to manually follow the action in real time, with no possibility of repeating a take, the sound engineer must therefore be accustomed to using a specific channel layout on the console, accepting a trade-off between number of channels and practicality. The scene-store-and-recall feature available in the majority of consoles is hardly used, because it implies discrete changes instead of the smooth transitions of a continuous mix. In such a scenario, the complexity of a live hands-on mix grows with the number of channels, which can easily surpass the dozen: although in many cases using more microphones could give a better sound quality for the action, using more than a dozen of channels would be too demanding for the mixing engineer.

In this section we introduce an application that simplifies the process of mixing the sound of the action by using a visual interface. Before going into to the details of the algorithm, it is worth introducing the concept of "point of interest", using the abbreviation *PoI*, to identify the region of space inside the field where the action is taking place. In the idealization considered here, the *PoI* is a circle whose parameters are the position of the center and the radius; it is a dynamic concept, meaning that its position changes over time and its size might too. To assist the sound engineer in the live mixing, we designed an application that provides a higher level of interaction and automates the movement of the faders to control the relative level of the microphones. The idea behind the application is to simplify the mixing process during a football match by changing the operation of the engineer from controlling various faders to just moving a point on a screen. Given a configuration of microphones around the field and a point of interest, the algorithm calculates in real time the gain factors to assign to the microphones in order to maximize the sound pick-up from the specified area. The input is the configuration of microphones, which is specified before the game, and the position of the point of interest, which is controlled in real time by the sound engineer. For each microphone, the gain that is calculated in real time is used to control the faders of the mixing console. The user interface is based on a visual representation of the field, where a circle representing the *PoI* is moved to follow the action. After the description of the application, we present reports on sound recording during a Spanish first division football match and the off-line testing of the application in the context of three-dimensional surround sound reproduction.

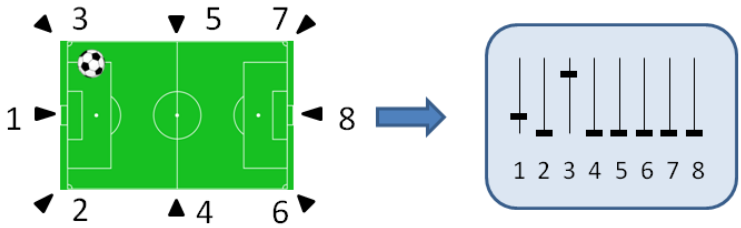

Figure 4.1: Mixing procedure for the sound of the action in a football game: given a configuration of microphones around the pitch, the sound engineer raises the faders of the console corresponding to the microphones that are close to the action, while lowering the others.

4.1.1 Algorithm

The idea behind the application is to simplify the mixing process during a football match by changing the operation of the engineer from controlling various faders to just moving a point on a screen. Given a configuration of microphones around the field and a point of interest, the goal is to calculate in real time the gain factors to assign to the microphones (or, equivalently, the levels of the corresponding faders in the mixing console) in order to maximize the sound pick-up from the specified area. The concept is shown in Figure 4.1, which illustrates how the position of the point of interest is mapped into a status of the console's faders. The input data is the configuration of microphones, which is specified before the game, and the position of the point of interest, which is controlled in real time by the sound engineer. For each microphone, the gain that is calculated in real time is used to control the faders of the mixing console. The user interface is based on a visual representation of the field, where a circle representing the *PoI* is moved to follow the action.

In order to calculate the gains, the following aspects were taken into account:

1. Given a source emitting a constant level, the output level resulting from the sum of the microphones shall be constant, not dependent on the position of the source.

2. The gain of each microphone shall be a monotonic decreasing function of the distance of the source; the closest microphone to the *PoI* shall have the largest gain (see Figure 4.2).

3. The user shall be able to choose a parameter controlling the number of microphones effectively participating in the sound mix, that is to blend

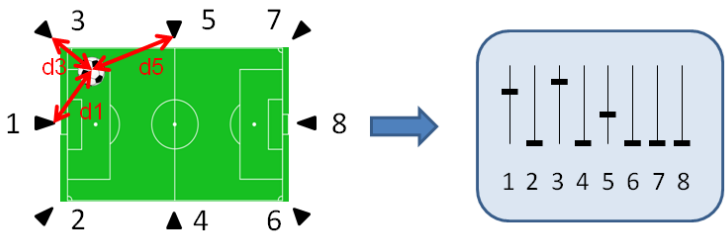

Figure 4.2: Only the close microphones participate to the mix. The farther the microphones, the lower the levels of the corresponding faders.

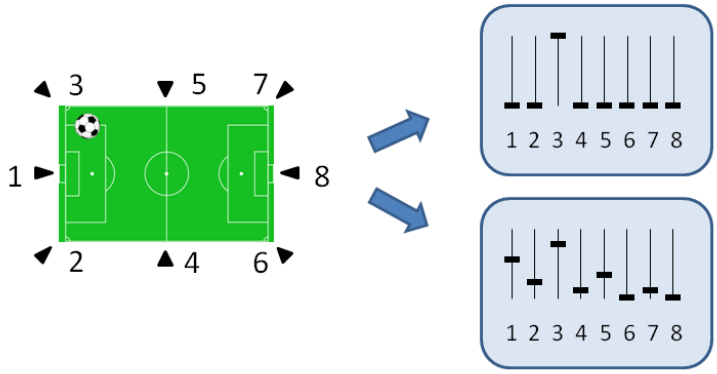

Figure 4.3: One parameter is required to control how many microphones participate in the mix. Ideally, it should allow any possibility between using only the closest microphone or using all microphones all the time, while maintaining the overall level.

between the "one microphone at a time" and the "all microphones at a time" configurations, as shown in Figure 4.3.

Moreover, it is assumed that the input gains on the console are set to give every microphone approximately the same sensitivity. The field is idealized in two dimensions as a rectangle, and a Cartesian reference frame is used. The input configuration specifies the positions of the microphones in Cartesian coordinates (x, y), the angle θ of their orientation in the horizontal plane and a coefficient A a specifying the polar pattern according to the following expression:

$$O(\theta) = (1 - A) + A\cos(\theta), \tag{4.1}$$

where O is the output level, normalized to one, as a function of the angle θ. Although the above formula is an approximation of first-order polar responses, it turned out to work correctly even for shotgun microphones, provided a hyper-cardioid response is used ($A \approx 0.75$). Every time the *PoI* is moved, the algorithm firstly reads the position of the *PoI* and computes the distances between it and each microphone. A quantity of 10^{-3} m is added to the distance as a regularization parameter to avoid it being zero in case the *PoI* overlaps with a microphone. The angle between the *PoI* and the direction of each microphone is also computed using trigonometric functions. Each microphone i is then assigned a gain $g_i(t)$ given by

$$g_i(t) = \frac{1}{(d_i(t))^{exp}} \frac{1}{(1 - A_i) + A_1 \cos(\theta_i)}, \tag{4.2}$$

where the subscript i refers to the $i - th$ microphone, d is the distance between the microphone and the *PoI*, A is the parameter that defines the polar pattern and theta is the angle between the axis of the microphone and the *PoI*. *exp* is the exponent that controls the decay with distance. After this calculation, the gains are normalized dividing by the sum of the squares of the gains of each microphone:

$$g_i(t) = \frac{g_i(t)}{\sqrt{\sum_{i=1}^{n} g_i^2(t)}} \tag{4.3}$$

This ensures that the global level, given by the sum of the squares of the gains, remains constant during time, independently of the position of the *PoI*. The rightmost factor in Eq. 4.2 compensates the off-axis attenuation due to the polar response of the microphones. The exponent *exp* affects the degree in which the *PoI* is synthesized by few or many microphones. A low value of the exponent will reduce the gain differences due to the distance, thus all microphones will contribute to the sound of the action with similar gains. The extreme case is $exp = 0$, where the gains do not depend on the distance. For $exp = 1$ the contribution of the microphones is in inverse proportion with their distance from the action. For high values of *exp*, only the closest microphones will contribute effectively. Engineers who are accustomed to raising one fader at a time will chose a high value of the exponent, while those who want to have most of the microphones involved will set a low value. When using a low exponent, all the microphones contribute to the sound; a consequence of this is that even audio events that are generated farther from the *PoI*, or in other locations, will be heard: conceptually, this corresponds to having a broader *PoI* region. Therefore, the exponent can be interpreted also as a parameter controlling the size of the *PoI*. The gains that are calculated so far are converted to decibels and sent in real time to an output port, to control an external device such as the mixing console. In case the console is controlled via MIDI and the mapping

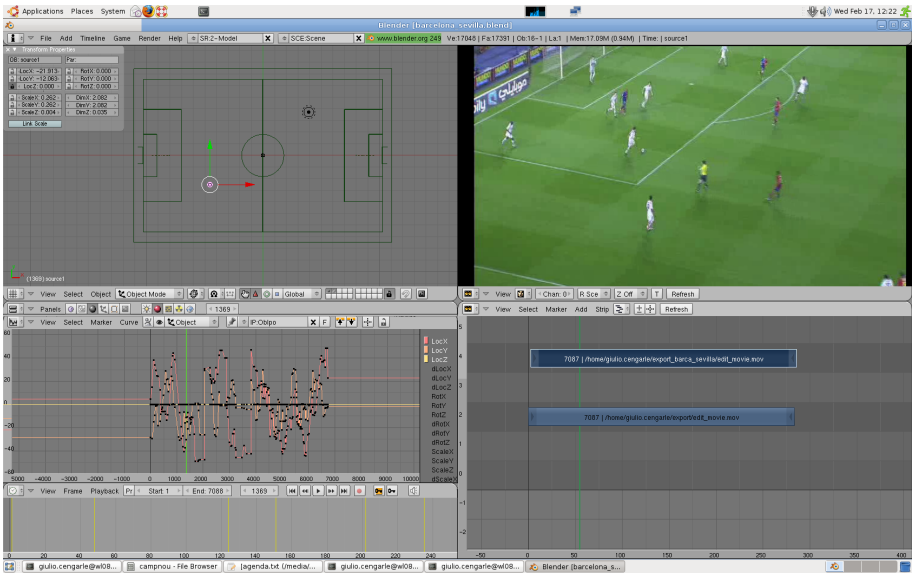

Figure 4.4: Screenshot of the Blender session used to control the *PoI*. On the top left the *PoI* is moved around the field in real time. The application can be used in a post-production environment with a video tab (upper right). Moreover, the coordinates of the *PoI* can be recorded and edited as keyframes in a timeline (bottom left).

is linear, the conversion to decibels is skipped. It is worth underlining that a device using this algorithm to control the console is not processing audio, but just controlling the faders. The integration in the work-flow is therefore straightforward and does not require any change in the signal path or in the common practice.

The algorithm was implemented and tested in Matlab, using previously recorded audio tracks for the testing. The first real-time working version was done in a PC using the 3D animation software Blender and the audio framework CLAM (2011). A Blender project was setup where the user could move the *PoI* with the mouse over an area representing the field, while the application sent open sound controls (OSC) over a specified port, that were received by CLAM. Within CLAM, a network and a processing were programmed to implement the algorithm and apply the gains in real time to a multi- track audio stream, either coming from a soundcard or being reproduced from a workstation. A screenshot of the Blender session is shown in Figure 4.4.

The next step was the creation of a standalone application with its graphical user interface. This was done using Python programming language and adopting a simple and effective approach in the design of the GUI; as can

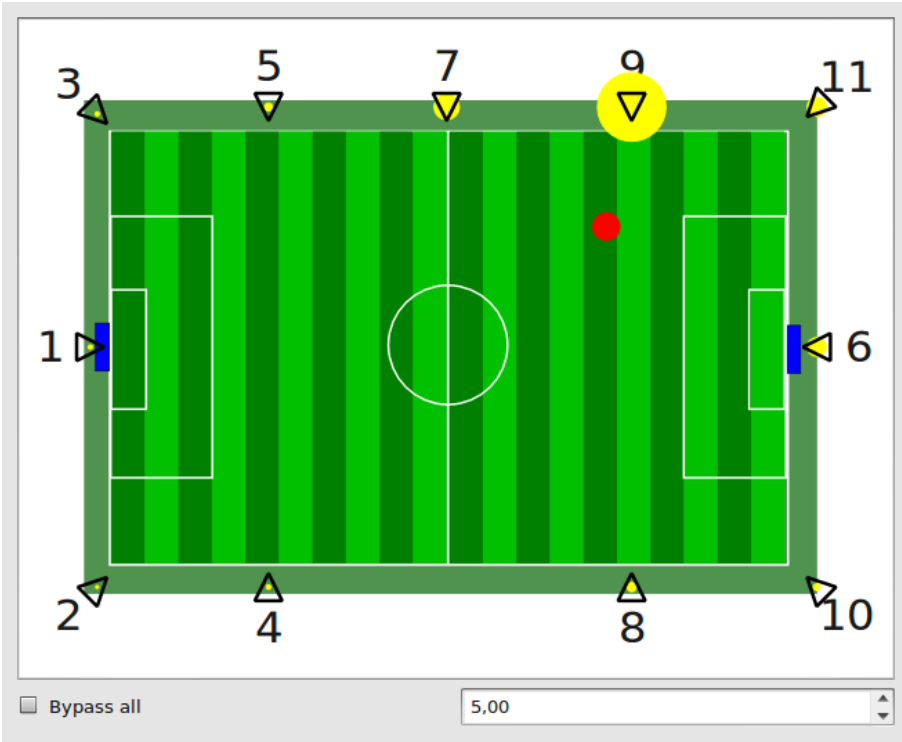

Figure 4.5: Screenshot of the python application. The circle in the field represents the *PoI*; the arrows around the field are the microphones, while their superimposed circles have a size proportional to the gains that are applied. The tab on the bottom right allows setting the distance exponent.

be seen in Figure 4.5, the screen represents a football field where the circle representing the *PoI* is moved by the user. The microphones that capture the action are represented as arrows oriented as the actual microphones. Every microphone has a superimposed circle, whose size is proportional to the gain that the application is computing for it. When the operator moves the *PoI*, one can see that the circles change their size in real time, getting bigger or smaller as the *PoI* approaches or moves away, respectively. This prototype application sends MIDI messages corresponding to the channels' gains through an output port, which could be for example a MIDI interface, to control a console, or to an internal port to control a digital audio workstation. A menu allows setting the value of the exponent for the size of the *PoI*. The application can be operated with a mouse or using the touch screen of a tablet; the latter was our choice, resulting in the compact device shown in Figure 4.6.

In order to configure automatically the MIDI ports and channels, the

Figure 4.6: The application integrated in a small tablet with touch screen.

application features a MIDI learn mode, in which a microphone is selected and listens for incoming MIDI controls; moving a fader of the audio console sends a midi message that is recognized by the application, so that the MIDI gains corresponding to that microphone are sent through the same channel. This allows a quick and intuitive setup of the microphone configuration.

An important issue to take into account is the total or partial overriding of the *PoI* control by the engineer operating the audio console. This feature is desirable for two reasons: one is the security in case there is a failure in the application; the other is to allow the engineer to trim or fine-tune the results of the automatic mix. First of all, a general bypass button that inhibits the sending of controls to the audio console has been implemented. Moreover, the status of each microphone can be switched at any time between:

- Active: corresponds to normal functioning and sending of controls.
- OFF: the microphone is removed from the configuration, its gain is set to $-\infty$ and a new configuration is computed taking only the remaining microphones into account. This mode addresses the case where a microphone is malfunctioning or is removed from the field.
- Manual: the application does not send controls for the selected microphone, which is controlled by the console, but the configuration is not changed to calculate the remaining gains.

In some cases, rather than changing the status of microphones in the application, the engineers want to act directly on the faders of the console. To avoid conflicts between the controls sent by the *PoI* and the faders, the application has been adapted to respond to changes operated in the audio console; if a fader is touched in the audio console, three options are available:

- The application moves the fader again as soon as the *PoI* is moved, or
- the application does not send controls to it for a chosen period of time – usually a few seconds – or
- the application does not move that fader until the difference between the computed gain and the actual gain exceeds a specified threshold.

Testing and feedback by professional users has allowed creating this extended set of features and finding the optimal parameters for controls such as the distance exponent, although they can vary according to the personal taste of each engineer.

4.1.2 Testing and integration with 3D surround

In order to test the application and to compare the output sound quality to the state of the art technique, before the live debut we carried out an off-line comparison of the two techniques, the standard manual one and the automated-assisted one, starting from a multitrack recording of a football match. The purpose was to record the signals of all the microphones separately, to play them back in a studio later, synchronized with the video, recreating a live-like mixing situation; this allowed experimenting with the rendering of the point of interest and doing a/b comparisons with the original broadcast audio, which was also recorded. Since the ambience microphones were recorded too, the same material was used later to produce a surround version.

Two football matches of the Spanish first division, Barcelona vs Getafe and Barcelona vs Seville, both taking place at Barcelona's Camp Nou stadium, were chosen. A configuration of eleven microphones around the field was used to capture the action, as this is a standard layout for these events. Shotgun microphones were employed, thanks to their rejection of audience sound and their high directionality, which offers a better pickup of distant sounds. These microphones were positioned in the lawn, as close as possible to the lines enclosing the game field, at a height of 0.5 m to 1 m. A stereo microphone Shure VP88, located in the center of the largest side, was used by the engineer to add stereo image to the action, as well as capturing some of the crowd sound. A Tetramic A-format microphone was located in the same position as the stereo pair, to be used for surround recording with Ambisonics technology. Two omni-directional microphones were put above the audience, hanging from the stalls, specifically for the crowd sound, and four

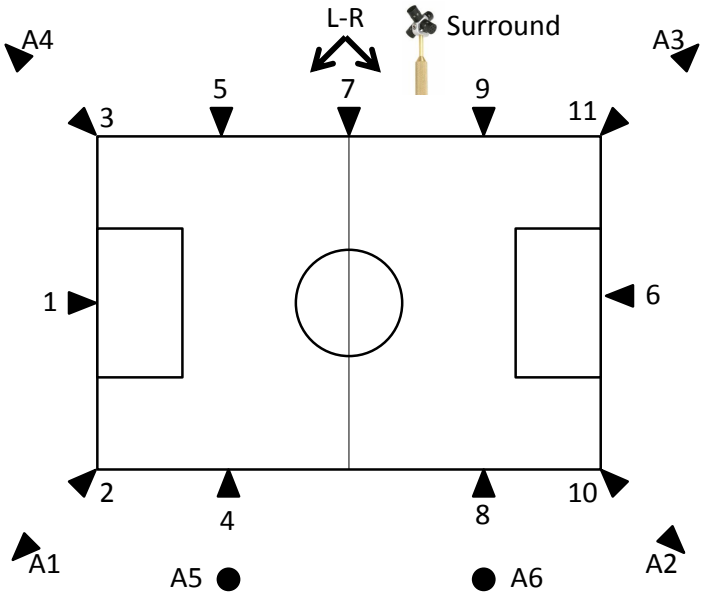

Figure 4.7: Microphone configuration in Camp Nou. Microphones 1 to 11 are shotgun used to capture the action. L-R and Surround are respectively a stereo and tetrahedral microphone located between the field and the audience. A1 to A6 are microphones used to capture the sound of the crowd.

directional microphones were put at the corners of the field aiming towards the crowd. The microphone configuration is shown in Figure 4.7. The signals from all the microphones were split and recorded separately in a multitrack recorder, while the sound engineer produced the live mix in the conventional way, consisting of a static 5.1 surround mix of the ambience combined with a mono mix of the action, done by moving the eleven faders of the shotgun microphones by hand. Some details of the installation are shown in Figure 4.8.

The recorded tracks corresponding to the microphones pointing towards the field were later reproduced in sync with the broadcast video, in order to test the application and compare it with the original mix. The first test was the proper functioning of the rendering of the *PoI*. The application was set to control via MIDI the gains of a DAW reproducing the multitrack recording. Listening to the mono mix, focus was put in particular on the rendering of the sound of the kicks, finding the application accurate in getting a clean sound, provided the *PoI* was set correctly. Comparing with the original production, some action sounds were recreated with higher presence. In a couple of cases,

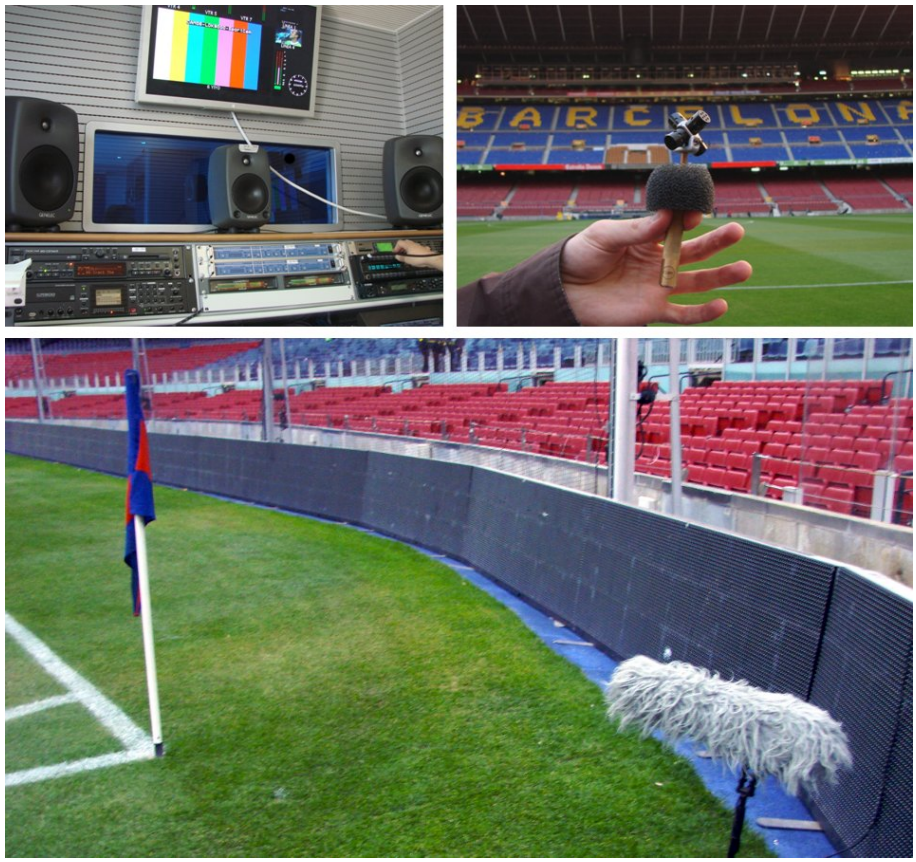

Figure 4.8: Recording setup at Camp Nou. Top left: the surround control room in the mobile unit; top right: tetrahedral microphone; bottom: a shotgun microphone in the corner.

the application brought out some sounds that were missing in the original mix, presumably due to a distraction of the sound engineer – a fact that is perfectly understandable due to the complexity of his job. Moving the *PoI* around the field to follow the action did not result in unwanted level jumps, as guaranteed by the constant level condition and the gain normalization that maintains a constant sum. The application turned out to be very easy to use following the video images, compared to the difficulty of moving the faders of the audio console.

After trying different values of the distance exponent, we opted for an optimum value of $exp = 5$, which resulted in a mix dominated by the closest microphone, guaranteeing a clear sound. The application was presented to and tested by some sound engineers and sport producers, who gave very

Figure 4.9: Three-dimensional surround listening setup. Twenty-two Genelec 8040 speakers are employed, together with two Genelec 7040 subwoofers.

positive feedback about its operation, ease of use and quality of the result. In particular, it was recognized that the chance of mistake in controlling the faders is greatly reduced using this technology, compared to operating the console. The tablet controller was carried to broadcasting mobile units and tested by sound engineers with a couple of broadcasting console models, using the prerecorded material to simulate a live feed. Again, the feedback from sound engineers was very positive in terms of easiness of use and quality of the results.

The simplification provided by the automatic control of the sound of the *PoI* turns out to be useful in the context of complex productions, such as is the case of surround mixing, where it leaves the sound engineer more freedom to adjust the ambience sound and to focus on the surround mix. The recorded material was used to present a three-dimensional immersive playback in an acoustically conditioned studio with a large screen, employing a setup of twenty-two loudspeakers, with ten speakers on the horizontal plane, four in a square in the corners of the floor, four in a square in the upper corners and four on top of the listener, as shown in Figure 4.9. A projector was used to display the video of the match. A PC running a digital workstation was used to reproduce the audio tracks. The automatic mixing application controlled the gains of the tracks corresponding to the microphones pointing to the field, and the resulting sound of the action was panned in mono

to the center of the screen. The Surround effect was rendered by decoding the Ambisonics microphone to our twenty-two loudspeaker setup. Since the sound provided by a single Ambisonics microphone was not enough to give an immersive sensation, the six ambience microphones were added to the mix. In order to locate the ambience sound around the listener, we opted for a surround panning of the audience microphones, displacing them slightly above the ear level. For the panning, both first-order Ambisonics encoding and amplitude panning were considered. It turned out that the Ambisonics encoding provided a better spread of the audience sound, giving a better immersive effect, in particular when there was a goal and the entire crowd was reacting. We found that the rendering of ambience sound in three-dimensional surround improves the sensation of involvement and participation to a live event; this is especially clear when switching back to 5.1 surround or to stereo. Moreover, while in a stereo production the action and the ambience would compete and mask each other, the use of surround allows for a higher level of ambience while still retaining all the details of the action, as happens in the real-life experience.

This approach to interactive mixing turns out to be very useful and intuitive in the current work-flow of live mixing for football, and will prove to be essential if the number of microphones is increased to more than a dozen. The successful testing and the positive feedback from the broadcasting industry and professional engineers have stimulated the development of ideas for future improvements and extension of features. For example, the use of the actual video of the game from a zenith camera as the background screen of the application would greatly improve the ease of use and the tracking of the *PoI*, because the engineer could just follow the ball with a finger. This case, however, is limited to sports where there is a zenith camera; in other cases it could be possible to generate the desired view by interpolating the images of other cameras. Although currently implemented and tested to work with consoles that feature MIDI control capability, it is possible to extend the compatibility to include communication protocols found in other popular sport broadcasting consoles. Also, the implementation of the whole process (the interface plus the software, the processing unit and the communication ports) in the form of a hardware device is desirable, to simplify the installation and facilitate its integration in sport broadcasting. One desirable feature to further automate the process is to slave the *PoI* tracking to the tracking of the ball or the players. Another desirable feature is the addition of more points of interest to work with simultaneously, and the possibility to link them or to make a stereo pair of *PoIs* for a wider rendering of the sound of the action.

To summarize, the application presented here is a graphical interface that allows controlling the levels of audio channels of a mixing console. The system automates the mixing based on the definition of a point of interest. It does not perform audio processing, it just controls the gains of audio channels, offering agile and safe operation in mixing environments with large

numbers of input channels.

4.2 Clipping detector for layout-independent 3D audio[2]

Since the advent of surround sound, various formats have appeared, such as 5.1, 7.1, 10.2 and 22.2, each one related to a specific layout of loudspeakers. The production work-flow for a given format has always targeted the destination layout, eventually accounting for formats which are subsets of the target either automatically, with down-mix, or manually, by re-mixing. However, with the growth of the number of formats and the complexity and cost of repeating the production of the same content for different formats, this channel-based work-flow becomes unpractical when many different formats are foreseen. Moreover, in the case of 3D sound, the lack of a *de facto* standard for playback is a further incentive for 3D audio production techniques to lean towards a format that is independent from the exhibition system. In this context, the concept of layout-independent audio scene has appeared, proposing that a soundtrack shall be encoded in a manner that does not depend on the number and location of channels in the playback stage. In this paradigm, the same soundtrack is delivered to different venues with possibly different layouts, and decoded differently for each of them once the layouts are specified. A scheme of this concept is shown in Figure 4.10.

Ambisonics was an early example of such a layout-independent work-flow. Object-Based audio scenes are other examples, whereby the encoding includes the audio essence of all the objects composing the scene and related metadata specifying their spatial properties: position in space, width, level, etc. Some implementations of the Object-Based Audio Format can be found in Hoffmann et al. (2003), Potard (2006) and Fascinate (2010). With object-based audio, the production becomes agnostic of the loudspeaker layout and independent from the playback format. In the last years, novel tools have appeared that implement such layout-independent work-flows in the post-production stage, such as immsound (2012), iosono (2012) and ssr (2012). The change in paradigm from a channel-based work-flow to a layout-independent one affects in particular the monitoring process: in the standard audio production practice for a given format, the engineers keep an eye on the output level meters, which are strictly related to the number of loudspeaker channels of the destination format; there is always a meter for each output channel, apart from the meters for each audio track; the output levels are kept under control to avoid digital clipping in the mix bus; sometimes dynamic processors, such as limiters and compressors, are employed in the output bus, where the signals they act upon are the direct outputs to the loudspeakers. In contrast, 3D audio layout-independent productions are typically played back on different loudspeaker layouts, which might be

[2]This section is based on Cengarle and Mateos (2012)

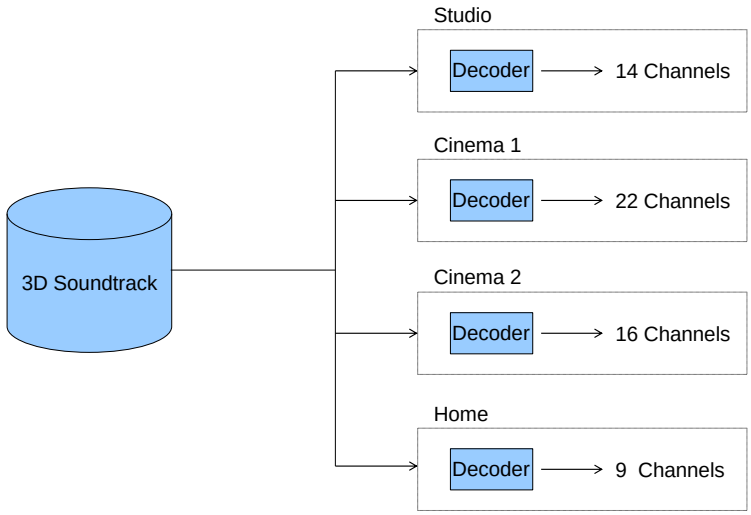

Figure 4.10: Concept of a layout-independent soundtrack decoded to different loudspeaker setups. The same content is delivered to each destination, where the in-house specific decoder performs the rendering.

even unknown during the production stage. Therefore, depending on the loudspeaker setup, the source signals will add in a different way before being sent to the loudspeakers. This can cause clipping in the playback system. To illustrate the problem in a practical case, let us consider the example of a stereo system, where the engineer wants to reproduce a sound source in the center (azimuth = 0°) at maximum level using amplitude panning. The stereo pair can be considered a subset of a larger multi-channel layout. In case of using two loudspeakers at ±30°, both speakers will playback the same signal, generating the illusion of a phantom source in the center, and the resulting sound pressure in the listening point increases 3 to 6 dB (depending on the degree of correlation in the arriving signals) with respect to the level produced separately by each loudspeaker. In a different layout where the loudspeaker pair happens to be rotated 30°, as shown in Figure 4.11, a typical amplitude panning algorithm (like VBAP) will use only one speaker to produce the same level as before. In the latter case, the center loudspeaker signal has to be boosted between 3 to 6 dB. If the levels in the first layout were already peaking near 0 dB digital Full Scale (0 dBFS), there would be clipping in the second layout. This simple example shows

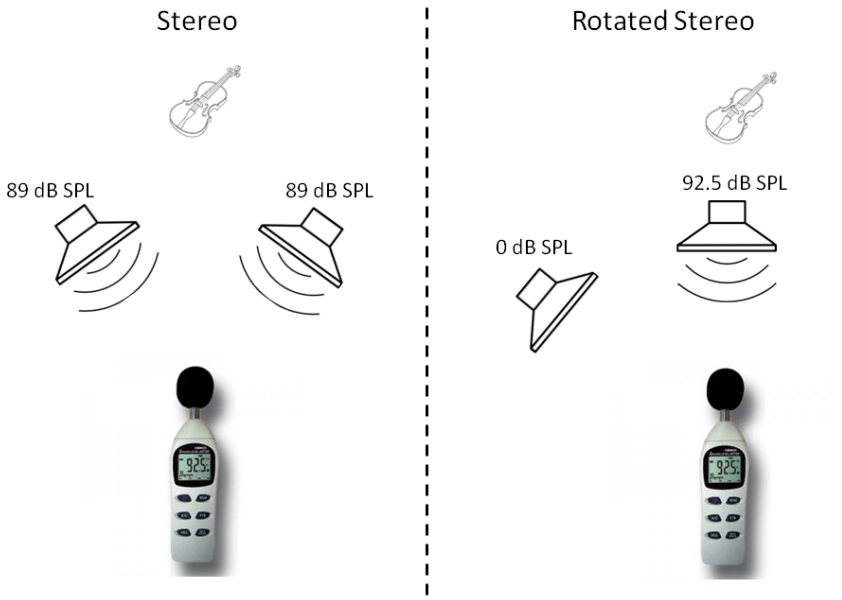

Figure 4.11: Decoding of a single source to a stereo and a rotated stereo system.

how clipping can happen when a single speaker has to do the job of many, typically because a loud source happens to be close to it.

This problem, inherent to layout-independent work-flows, is rather novel to the industry and, to the author's knowledge, it has not been tackled in the literature. In this chapter, we take an initial step along this direction by providing an algorithm that analyzes the soundtrack and estimates quantitatively the potential problems related to clipping. The algorithm is based on an initial definition of worst-case layout, and on a sample by sample rotation and decoding of the audio scene to the worst situation in the worst-case layout. Although the algorithm itself does not solve the clipping issue, it detects its occurrence and provides reassurance that there will not be clipping in any layout at least as dense as the defined worst-case layout. This might suffice in many practical situations. In Section 4.2.1 the strategies for tackling the problem of clipping detection are considered and the proposed algorithm is described, while Section 4.2.2 presents the validation and the results obtained.

4.2.1 Strategies for clipping inference

We consider a channel-free production work-flow based on a hybrid method which consists of amplitude panning for localized sources and Ambisonics for diffuse sounds and reverberation. A scheme of the processing blocks involved is shown in Figure 4.12. The audio tracks are sent to the spatialization plug-ins together with the associated position metadata. The amplitude panning plug-ins calculate the decoding coefficients in real time, according to a configuration file that specifies the position of the loudspeakers. The B-format signals are summed together and the resulting B-format bus is send to the Ambisonics decoder, whose coefficients have been computed according to the specified layout. The particular implementation used for this thesis is based on an audio session of the open source audio workstation Ardour (2012). The spatialization of amplitude panned sources and the decoding of first-order Ambisonics signals (B-format) are implemented in terms of custom made plug-ins using the LADSPA architecture, based on the VBAP algorithm for the former, and an in-house irregular Ambisonics decoder based on simulated annealing non-linear search [Kirkpatrick et al. (1983)] for the latter. The implementation of this hybrid approach has been possible thanks to the development of CLAM (2011), a framework for development of audio processing, where the audio spatialization plug-ins were built. In this layout-independent approach, once the mixing is done, the delivery format consists of all the audio tracks with the related gain and spatialization metadata, as well as the four B-format channels resulting from the mix of the Ambisonics tracks. Since we used the digital workstation Ardour with custom plug-ins, in our case the delivery format is represented by the Ardour session itself. The tracks levels have been balanced aesthetically during the mixing session and are supposed to be played back in any compatible system maintaining not only the position and the relative level, but also the perceived loudness. The sound system employed is calibrated for absolute loudness, such that each loudspeaker fed with a reference signal (pink noise with a level of -20 dBFS R.M.S. in the digital domain) produces a sound pressure level of 85 dB SPL C-weighted in the listening area. This practice has its origins in the movie industry, where both mixing rooms and cinema halls are adjusted to the same level, to avoid - in principle - volume control at the user end [Allen (2006); Katz (2000)]. When reaching the decoding stage, the actual layout configuration file, containing the information on the loudspeakers' positions, is used to configure the plug-ins, which compute the signals sent to each loudspeakers. If every playback setup is calibrated for the same reference level, the soundtrack and each individual track will produce the same loudness, regardless of the position of the speakers. The audio production is monitored through a suitable reference layout, although the configuration of the various final playback setups may be unknown. Even though preliminary mixes may be done even in stereo, proper monitoring of a 3D production requires the presence of a certain loudspeakers density covering the regions

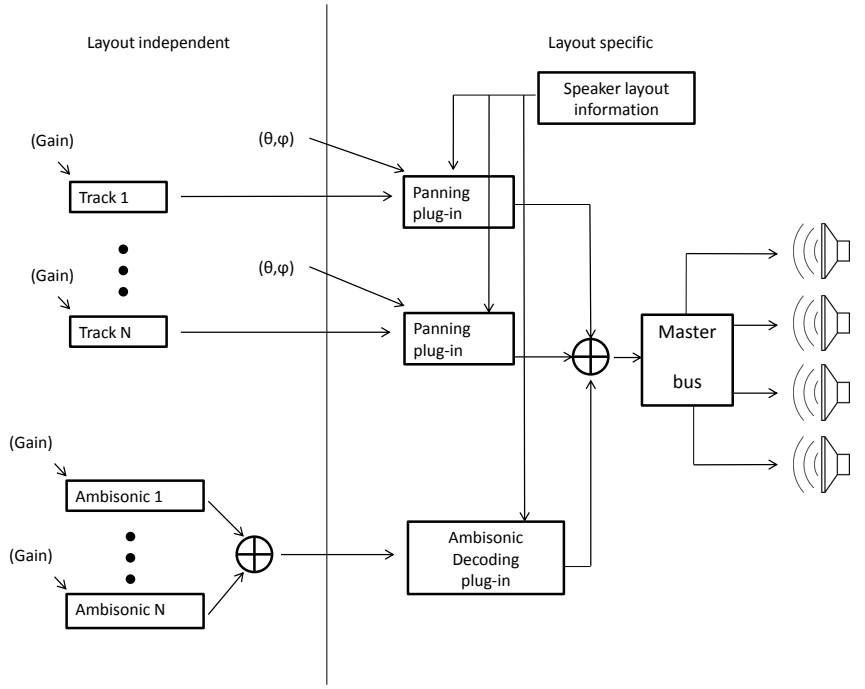

Figure 4.12: Scheme of the processing blocks of a layout-independent audio session for amplitude-panned and Ambisonics sources. Sound engineers act on the position and level of sound sources, while the decoders and panning plug-ins take care of reconstructing the audio scene according to the specific loudspeaker layout.

where the sources are going to be located. However, as shown in the simple stereo example of the previous section, absence of output clipping in the monitoring system does not give enough indication about what can happen with other layouts.

Avoiding clipping in an unknown playback configuration is not just a matter of keeping the levels on the safe side, for example not exceeding -3 dBFS. One may wonder why not just lowering the global level in the digital domain, to keep high headroom and compensate the loudness in the analogue output chain with a volume control. The reason is that in many cases, especially in the movie industry and in broadcasting, the content is produced considering absolute loudness, and is supposed to be played back at the same reference level. One trivial strategy to guarantee safe levels in most scenarios would be to sum all sources to mono and detect clipping. However, this would be too strict, since mono is not seriously contemplated as an exhibition format by producers of 3D audio content. Besides, it could even fail in tricky situations: consider two sources panned left and right in

a stereo system, where one of them is the exact copy, polarity reversed, of the other; in this case, their sum to mono would vanish, even though the tracks alone may be on the edge of clipping. Adding a third source on the edge of clipping to one channel would cause the stereo version to clip, but this would not occur for the mono version! Another strategy to circumvent the problem is to apply multichannel control of dynamics after the decoding stage. In case the decoding processing uses floating point arithmetic, it could be followed by a digital limiting stage with a threshold close to 0 dBFS. This option might not be welcome by audio engineers, because they need to be aware and hear the effects of any processing device in the chain before committing their decisions to the master version of a soundtrack. An algorithm is required that can predict the occurrence of clipping, issue a warning and optionally trigger real-time processing (e.g. limiting) to let the engineer choose between accepting the limiter's effects or retouching the mix to prevent the problem. The proposal discussed in this paper is based on the definition of a worst-case layout contemplated by the creator of the 3D soundtrack. The algorithm developed on top of it ensures that there will not be clipping in any layout with a higher loudspeaker density. The definition of the worst-case layout depends on many factors, including the type of content and the policy of the producers. For example, a 9.1 layout [Auro3D (2012)] with no loudspeakers on the ceiling, and only four on top of the L-R-Ls-Rs of a standard 5.1, might be suitable for reproducing a classical concert, but it might be defined as a worst-case layout for a spectacular sci-fi movie, with plenty of sound panning on top of the audience. Regular loudspeaker layouts are proposed as candidates for worst-case layouts, due to two main reasons: symmetry and hierarchy. Firstly, the symmetry of regular layouts allows for an understanding of their virtues and defects in terms of only one loudspeaker and its nearest neighbors, which are parameterized by a single angular distance (e.g. in a dodecahedron, 42°). Secondly, regular layouts can be classified hierarchically in terms of angular resolution, as the higher the number of loudspeakers, the lesser the angular distance among neighbors; in contrast, irregular layouts are difficult to compare, as the loudspeaker density varies among different reproduction areas. Once a suitable worst-case layout is defined, the algorithm proceeds on a sample by sample basis. For the sake of concreteness, the discussion presented here is restricted to amplitude-panned and Ambisonics sources. In a first stage, the algorithm selects the loudest object-based source, as the main suspect that might lead to clipping. The whole worst-case layout is rigidly rotated in order to consider the worst possible scenario from the point of view of clipping: the loudest source coincides with one loudspeaker L_1, and the second loudest source lies on an edge with one nearest neighbor L_2. In typical amplitude panning algorithms, the latter condition forces the second loudest source to be reproduced by only two loudspeakers, L_1 L_2, therefore giving a higher load to speaker L_1. This initial rotation is performed regardless of the Ambisonics sources. The reason is that they are mixed in the Ambisonics

bus before decoding. If the decoder is properly built and normalized [Daniel (2000)], absence of clipping in the Ambisonics bus guarantees absence of clipping in the decoder, regardless of whether the layout is regular or not. In the last stage, the content in the Ambisonics bus is decoded to the rotated worst-case layout and added to the amplitude panned sources. The following steps detail the complete algorithm:

1. Definition of the worst-case layout.

2. Clustering of sources: for each source, consider all others at angular distance less than the distance to the nearest neighbors of the worst-case layout. Exit if all clusters contain only one source.

3. For every cluster:

   (a) Read the post-fader level and position metadata, e.g. azimuth and elevation, of each amplitude-panned audio track.

   (b) Detect the loudest source in each cluster.

   (c) Check if the level of the loudest source is at least $-6 \log_2 N$ dBFS, where N is the number of sources within the cluster. Exit if the condition is not met.

   (d) Rotate the worst-case layout so that one loudspeaker, L_1, coincides with the direction of the loudest source (see Figure 4.13a).

   (e) Further rotate the worst-case layout about the line joining the origin and L_1, so that the second loudest source lies in the line between L_1 and one of its nearest neighbors, L_2 (see Figure 4.13b).

   (f) Decode the amplitude panned sources in the cluster to the loudspeakers of the rotated layout.

   (g) Decode the Ambisonics bus to the rotated layout and add to the former (Fig. 4.13c).

   (h) Detect resulting level g_1 in L_1 and use this value as a quantitative indicator of potential clipping. Trigger a clipping warning if $g_1 > 0$ dBFS, or possible dynamic processing based on the value of g_1.

The choice of $-6 \log_2 N$ dBFS in the algorithm is due to the fact that the maximum boost in level given by the sum of N sources is $6 \log_2 N$ dB, which happens in the extreme case where all signals are identical and located at the same spot. This algorithm is computationally heavy, because it searches for every cluster of sources; however, it can tackle the issue of clipping caused by a large number of quiet sources. For a reduced computational load, a simplified version can be obtained by examining only the single cluster containing the loudest source.

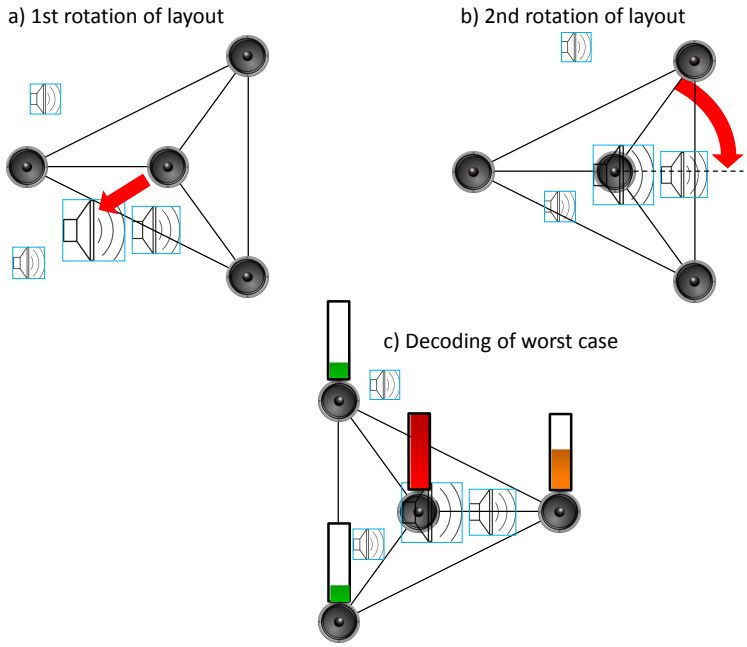

Figure 4.13: Scheme of the algorithm for amplitude-panned sources, showing how the layout is rotated to give the highest load to the center speaker and monitor its levels.

4.2.2 Application and validation

In this section, the approach proposed above is validated in practical examples where clipping occurs due to decoding layout-independent content to different layouts. The algorithm has been implemented using MATLAB. All amplitude panned sources are decoded using the VBAP algorithm. The Ambisonics decoder was chosen to satisfy the maxRe constraint [Daniel (2000)], which generates the highest level on a single loudspeaker, in case the source position coincides with it. The worst-case layout was set to a regular dodecahedron, for which the angle between two adjacent speakers as seen by a listener located in the center is approximately 42°. The following examples consider three practical situations of increasing complexity.

Stereo vs. LCR

The first example focuses on a simple enough scene that admits analytical treatment. Let us consider the two most standard industry layouts: a stereo system (L-R channels at ±30°), and an ITU 5.1 system, the latter containing

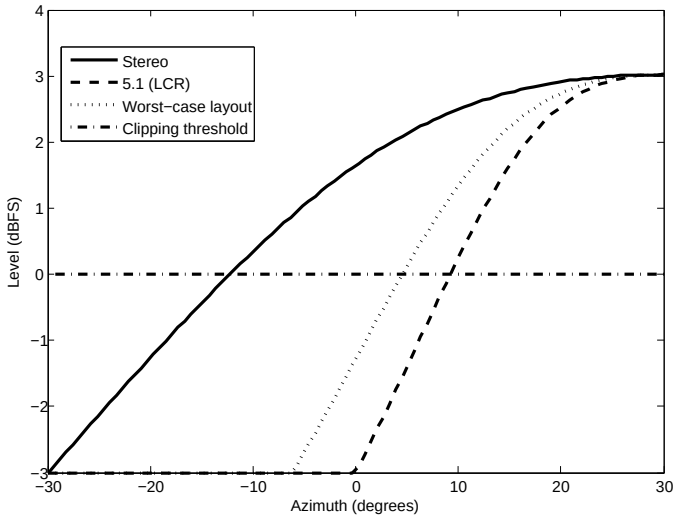

Figure 4.14: Levels at L-speaker in dBFS as a function of azimuth of source S_2, for two different layouts and the worst case.

an LCR subset in the frontal area. Consider the case where a sound source S_1 is fixed at 30° azimuth (L-channel) and another identical source S_2 is panned uniformly from −30° (R-channel) to 30°. Both sources have a peak level at -3 dBFS. Figure 4.14 shows the peak level in the digital domain in speaker L as a function of the position of S_2, using a VBAP decoding. The maximum level that can be produced by the panning law in such a situation is +3 dBFS, when S_1 and S_2 coincide at 30° azimuth. In the Stereo decoding, some signal from S_2 is added to the L-speaker as soon as the source departs from the R-speaker and moves towards the left, while in the LCR setup the signal is added to the L-speaker only when the source moves beyond the C-speaker. Having a speaker in the center allows a loud source to be panned in the right panorama without affecting the level of the L-speaker. Applying the algorithm, potential clipping problems are detected slightly before they show up in the LCR configuration. We also notice that the resulting level in the L-speaker in stereo is always higher than the level in our worst-case layout. The reason is that the angle between stereo loudspeakers is 60°, while our worst-case layout has an angular distance of 42° between speakers.

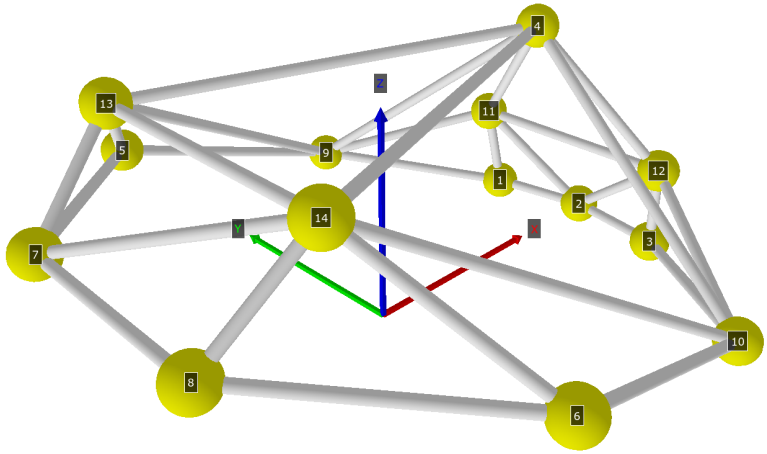

Figure 4.15: Scheme of the 14.1 layout. The screen LCR subsystem corresponds to the labels 1-2-3, respectively. Note the appearance of three independent channels (labeled 4-13-14) directly hung from the ceiling.

14.1 VS 22.2

Let us consider a slightly more complex example, comparing the decoding to two 3D loudspeaker layouts: a 22.2 as specified by Hamasaki et al. (2005), and a 14.1, depicted in Figure 4.15, which is similar to the ones installed in various cinema theaters in Holland and France [immsound (2012)].

These two layouts provide good examples of how clipping can occur in either setup for different reasons. The 22.2 layout has one loudspeaker on top of the sweet spot, while the 14.1 has three loudspeakers that cover the entire ceiling. On the contrary, the 22.2 shows a higher loudspeaker density in areas other than the ceiling. Thus, concentrating sources is likely to cause clipping issues in either layout, depending on where they are located. As a first example, consider two sources with a constant peak sample level of -4.4 dBFS located at elevation 80°, and azimuth 0° (front) and 180° (back). The decoding to the 22.2 layout leads to a sum which heavily loads the top loudspeaker, giving a total level of +1.1 dBFS. In contrast, in the decoding to the 14.1 layout the sum is spread over the three ceiling loudspeakers, leading to a maximum peak at -1.0 dBFS. The detection algorithm presented here correctly anticipates clipping, with a maximum level of +1.1 dBFS. Focusing on the front-top area, the results change. The decoding of the same two sources at elevation 45°, and azimuth ±10° (front-top-left and front-top-right) leads to absence of clipping in the 22.2 layout, but to a peak level of +0.7 dBFS in the top-front loudspeaker of the 14.1 system. Again, the

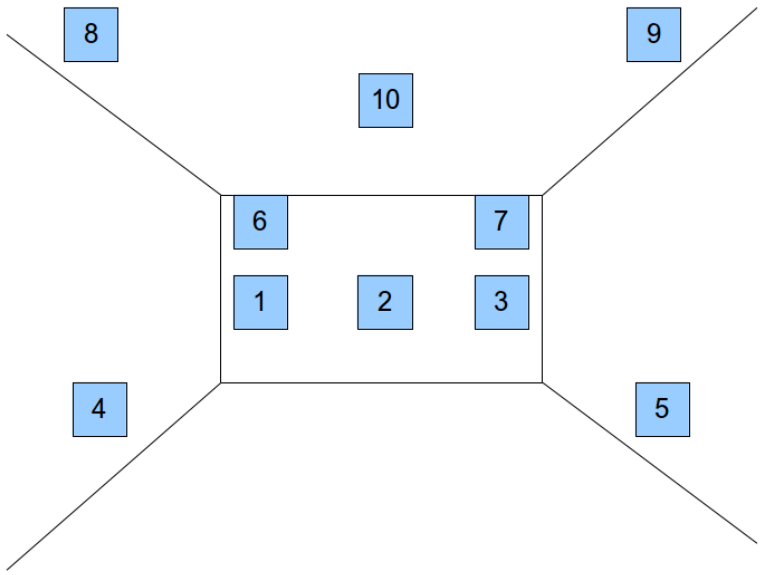

Figure 4.16: Scheme of a 10.1 layout obtained by the addition of a "voice of God" channel to the 9.1 in Auro3D (2012).

detection algorithm presented here reports clipping with a peak level of +1.1 dBFS. In both cases, the algorithm warns that clipping would occur with some layout, so the engineer can slightly reduce the level of the sources or apply some dynamic processing to tame the levels.

Complex soundtrack decoded to different layouts

As a more complex example, let us consider an excerpt from an experimental audio production created at the Barcelona Media 3D sound laboratory. The excerpt analyzed consists of a helicopter running circles in the ceiling above the listener, with rocket explosions happening straight above the listener and a wind-like ambient sound. The helicopter and rocket are panned using VBAP, while the wind is a B-format track that is rendered by the Ambisonics decoder. The helicopter and rocket tracks have peak levels close to full scale. The production has been monitored using the 14.1 loudspeaker layout described above. The resulting peak level in the output bus is -0.4 dBFS, with no clipping occurrence. However, decoding the soundtrack to a 22.2 layout results in a peak level of +1.1 dBFS in the loudspeaker on top of

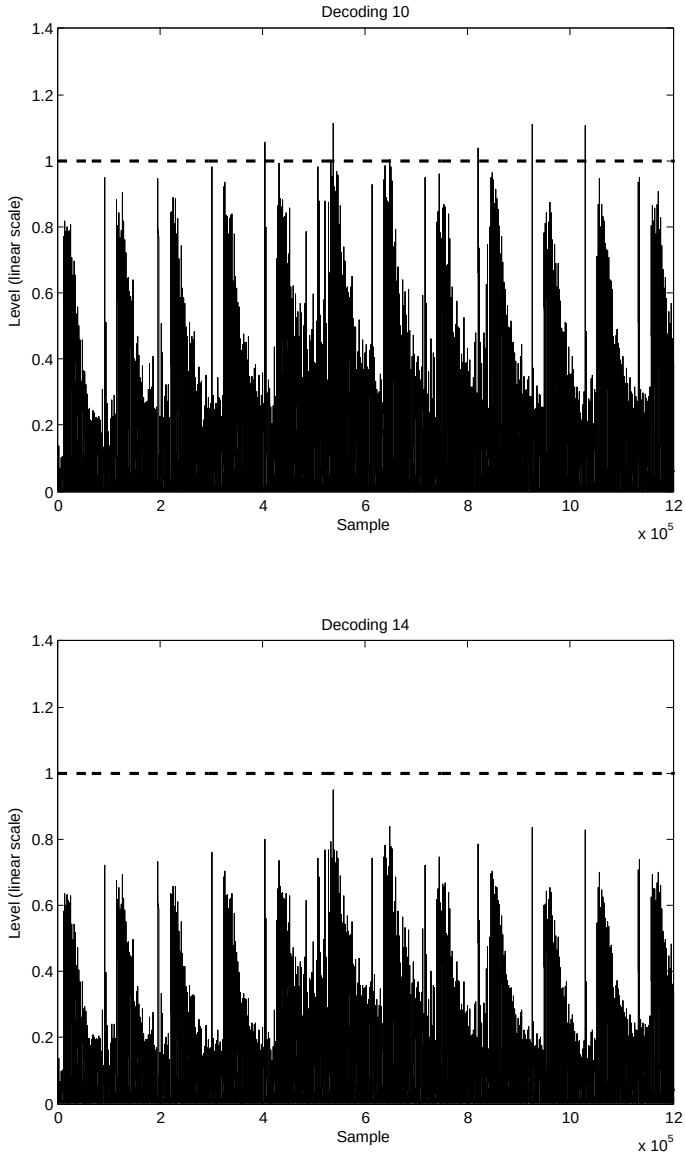

Figure 4.17: Levels in the most loaded speaker for the tested layouts. Top: 10.1; bottom: 14.1. Levels are plot in linear scale for ease of visualization.

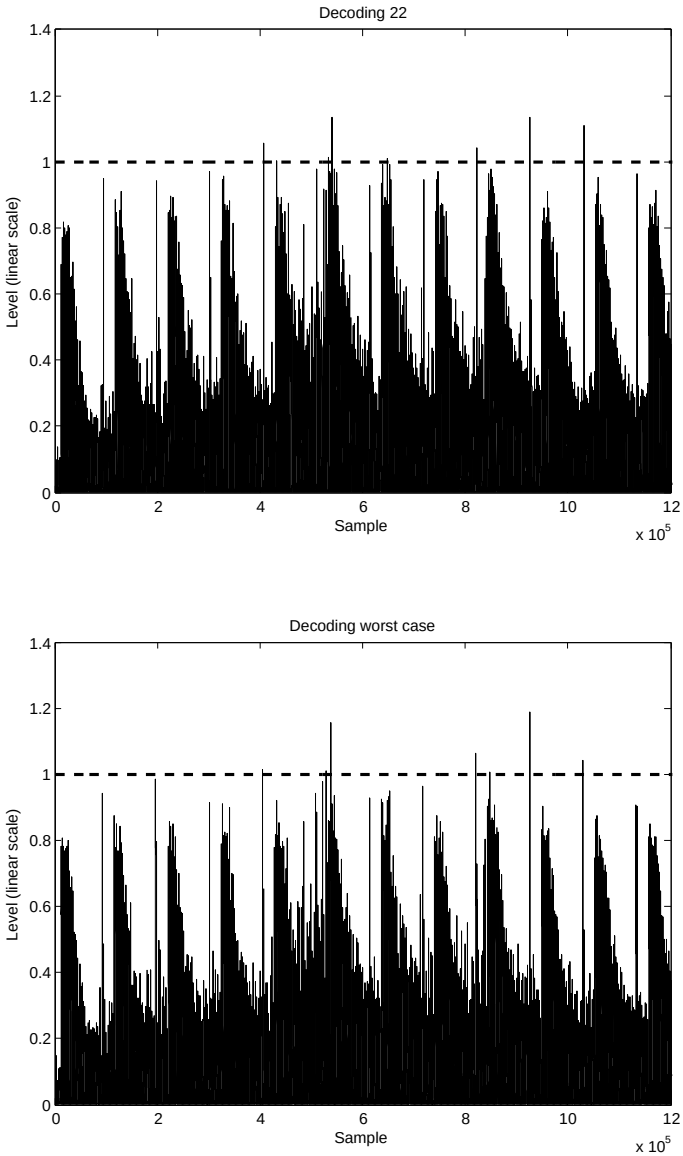

Figure 4.18: Levels in the most loaded speaker for the tested layouts. Top: 22.2; bottom: worst-case level reported by the algorithm using the regular dodecahedron. Levels are plot in linear scale for ease of visualization.

the sweetspot. Decoding to a 10.1 layout (Figure 4.16), obtained by the addition of a "voice of God" channel to the 9.1 layout, yields a peak of +1.0 dBFS in the "voice of God" loudspeaker. The decoding of this soundtrack to these three layouts yields a difference of 1.5 dB in the peak level of a single loudspeaker channel. The levels in the most loaded loudspeaker are shown in Figures 4.17 and 4.18 for a fragment of the soundtrack 25 seconds long. In this complex case, our clipping estimation algorithm correctly reports peaks above full scale, as shown in Figures 4.17 and 4.18, corresponding to the samples where clipping occurs in the 22.2 and 10.1 layouts. The value of the maximum peak reported by the algorithm, converted in decibels, is +1.1 dBFS, which corresponds to the actual peak levels encountered.

These examples validate the proposed solution as a practical strategy to detect and quantify the problem of clipping, which relies on i) an initial definition a worst-case loudspeaker layout suitable for the particular content and exhibition policy, and ii) a sample by sample set of rotations of the whole audio scene that renders the potentially most dangerous situation in the worst-case layout. The algorithm reports the excess levels (in dBFS or sample values) that might cause clipping in different layouts, and the sample at which it occurred, allowing the engineers to act accordingly. The choice of worst-case layout can be adapted to situations where the soundtrack is designed for a higher or lower density of speakers, without loss of generality.

5 Subjective effects of 3D audio

Audio-related technology has one ultimate target: the listener. After introducing 3D audio technology and making it available to the end user, the most important question is: how is 3D sound perceived? Also, what improvements does 3D audio bring to the listener that 2D audio is missing? As already anticipated in the introduction, the main advantages are that 3D sound can get closer to a faithful representation of reality, and unleashes great potential for creative effects. In this chapter we discuss how certain aspects related to the impact of 3D sound on the audience can be evaluated and what results are found.

While the algorithms and the technical aspects themselves are measurable in an objective way and give predictable results, the impact of the same technology on the listener is subjective and therefore more difficult to assess and quantify. The judgment of audio quality is a subjective matter, which is easily influenced by secondary aspects such as the artistic content and the peculiar preferences of the subject.

Our goal here is to study some specific aspects of audio perception comparing 3D versus other standard formats. The first aspect that has been addressed is if 3D audio has more emotional impact than 2D surround; in other words, the goal here is to assess if the "immersiveness" that is associated to 3D audio can be quantified. A basic response can be obtained by answers to questionnaires by listeners experiencing and comparing different audio formats, although on a deeper level the emotional impact is related to psycho-physiological data that can be measured; the first methodology allows to obtain information on the conscious effects, while the second can reveal unconscious aspects. In the first part of the chapter we present a study in which we have collaborated, carried out by the Perception and Cognition group of Fundació Barcelona Media, where psychophysical data such as the hearth rate and the electrical conductivity of the skin are measured while subjects watch audiovisual material with 5.1 surround or 3D audio. The same subjects were asked to fill a questionnaire, so that the perception of 3D content in terms of immersion and emotional impact could be evaluated both on a conscious and an unconscious level.

The next aspects that were evaluated are related to the reverberation and the spatial masking effect. Both in real life and in audio production,

reverberation has the important role of supporting the dry sound of the sources and making them blend together, but its balance with respect to direct sound is critical to avoid "muddiness" and loss of intelligibility. Our perception of reverberation in real environments and our aesthetic taste for recordings are quite different: for example, the best listening seat in a concert hall is hardly in close proximity of the stage, while the best recording position in the same hall is hardly far in the reverberant field. There is general agreement that recording from the best listening spot in a hall gives too much reverberation in playback, with subsequent loss of definition and articulation. The preferred position of the microphones is always within the reverberation radius. The reverberation radius, also known as critical distance, is the distance from the source where the level of the reverberation equals the level of the direct sound. Within the reverberation radius, the direct sound dominates, while beyond the critical distance the level of reverberation is higher than the direct sound. Assuming a model where the direct sound decays as the inverse square of the distance, and the reverberation is a diffuse field where the level does not depend on the position, an approximation of the critical distance is $d_c \approx 0.057\sqrt{V/RT}$, V being the volume of the room and RT its reverberation time. Typical values of volume and reverberation time of concert halls give critical distances in the order of a few meters, which implies that most of the audience is immersed in the reverberant field. Nevertheless, the timbre, articulation and position of sound sources are perceived satisfactorily in good halls. When playing back a recording, two effects come into play regarding reverberation. Firstly, the perceived reverberation is the sum of the one that is present in the recording and the contribution of the listening room, although standard studios and listening rooms have such a low reverberation that the effect can be neglected. The second and, in the author's opinion, most important effect is the fact that the original three-dimensional diffuse reverberation is now collapsed in the area (or line) where the speakers are located, and competes with the direct sound for intelligibility. In the case of stereo, the whole reverberant field is shrunk to an area not wider than 60° and concentrated in the front stage. In 5.1 surround, the reverberation is at least distributed over the whole horizon, but it is only in 3D that the spread of reverberation, if properly captured, can be faithfully reproduced.

Not only the reverberation benefits from a three-dimensional spread: real life situations can present the listener with dense soundscapes, with spread sources and many different elements. Creating a rich soundscape in stereo or traditional surround requires sound designers and mixing engineers to accept trade-offs between how many sound sources can be included in the panorama and how many of them will actually be heard instead of being masked by the dominant ones. In 3D, a sound engineer can locate sounds on a sphere (at least in the upper hemisphere) instead of a line, and has therefore more freedom to fill the space with sound, to create reach, dense scenes. In many conversations of the author with mixing engineers and sound designers, the

topic of how to take advantage of the spatial distribution of sound was raised. The opinions and personal experience of sound engineers and the author himself seem to support the idea that with 3D sound one can use higher levels of reverberation and a higher number of sound elements (if properly distributed!) before they mask each others causing loss of intelligibility. Whether or not an increase of density or reverberation is desirable from an artistic point of view is part of the aesthetic choices of content creation and is entirely dependent on the context and taste, but with our experiments we want to provide some objective support to this hypothesis.

The perception of single sources in the presence of reverberation and dense soundtracks is related to the phenomenon of spatial masking. Masking is often considered in terms of frequency, where two sounds within the same critical band compete with each other and, in case they are different in level, the louder sound makes the softer one inaudible. Masking also depends on the direction of arrival of the simultaneous sounds: when the masker and masked sound coincide in space, they are more difficult to separate compared to the case when they come from different directions [Hawley et al. (2004), Kidd et al. (2005)]. To asses the benefit of 3D audio rendering compared to stereo and 2D surround, we designed a psychoacoustic experiment where the perception threshold was measured for a hidden tone masked by bandwidth filtered noise, comparing the case where the noise is spread over the stereo panorama, the horizontal plane and the whole 3D sphere. Results confirm that the same level of masking noise gives less masking in 3D compared to 2D or stereo. The psychoacoustic experiments consisted in measuring thresholds by means of A/B comparisons with two-alternative forced choice design, fitting the results with a Weibull psychometric function.

A final experiment was carried out to assess the subjective perceived level of reverberation in 3D and standard formats (stereo and 5.1). In this experiment, based on the method of adjustment, subjects were asked to match the reverberation level of a spoken voice to a reference reverberation applied to the same signal, with the purpose of finding if there are significant perceptual differences in reverberation level among 3D and 2D formats. The experiment is repeated so that each audio format is used as a reference against which to compare the others. Results indicate little difference in perceived level among formats, suggesting that once a reverberation level is chosen, it would be perceived similarly in any format.

5.1 Emotional impact of 3D audio

In describing the 3D audio listening experience with good audio content, most end users use the words “realistic” and “immersive”; content creators who use a 3D format do so because they feel it can convey higher emotional involvement and it can match better the expected direction of arrival of sounds as intended in the script: for example, helicopters approaching from

out of the screen are supposed to sound overhead; a crackling noise to provoke fear in the dark scene of a thriller can be located behind the audience, and so on. 3D audio can impact the listener on a conscious or unconscious level. The conscious reaction to 3D audio can be assessed by having subjects listen to versions of the same audio track with both 3D and a standard format and asking them suitable questions. Besides, whether or not they are conscious of the sonic differences, the reaction of listeners to different types of audio content reflects in variations of psycho-physiological data if the audio formats evoke more or less emotional arousal. This section presents the results of a test aimed at revealing the emotional differences caused by 3D audio versus traditional surround. The test has been carried out within the European Project *2020 3D Media*, and the results that are presented in this section have been published in an internal project report.

Two audiovisual productions were chosen for this task:

The library, a short movie about six minutes long

Sintel, a short animation movie about four minutes long

Both shorts are intended to evoke emotions such as fear and suspense, with clear peaks of such emotions in specific parts. The audio for both movies has been produced using a layout independent approach, allowing to easily render two versions of each short: one with standard 5.1 surround audio and one with 3D immersive audio. Both shorts feature audio content that includes sharp, well localized sounds in and out of the screen, together with enveloping reverberation and ambient sounds. The room used for the audio/video playback is an acoustically conditioned sound studio, approximately 7 m x 5 m, with a 4 m x 2.5 m perforated screen and a 3D loudspeaker system with 23 channels and one LFE channel. Standard 2D projection is used for the video. The audio decoding ensures loudness matching between the 5.1 and 3D versions, the 5.1 version being essentially the projection of 3D sound onto the horizontal plane, using only five channels plus the LFE.

The test is a half mixed design, with the following variables:

i Audio format: 5.1 or 3D

ii Movie: *The library* or *Sintel*

iii Order of presentation of the movies

24 subjects (12 male, 12 female), aged between 25 and 51, participated in the test. The participants were not expert in audio, and most of them had not experienced 3D audio before. Each participant watched both movies, each one of them in a different audio format. The audio format in which a movie was presented alternated between participants, so did the order of presentation. The following dependent variables (psycho-physiological measurements) were recorded by means of a polygraph:

- Electro-dermal activity (EDA)
- Facial electromyography (EMG)
- Heart rate (HR)
- Respiratory sinus arrhythmia (RSA)

EDA reflects changes in the electrical conductivity of the skin due to the activity of the sweating glands induced by the parasympathetic system. It is related to the degree of emotional activation and it can also be an indicator of cognitive effort.

EMG measures muscular activity and is often employed to measure facial muscle activity. Depending on the registered muscle and stimuli conditions, it can indicate the emotional valence (the positive or negative characteristics), discomfort and attentional capture.

HR and RSA indicate attentional responses as well as general stress levels.

For each participant, the experiment begun with the recording of baseline activity, measuring the aforementioned variables during one minute at rest, without doing any activity. The participant then watched both movies, each one in a different audio format, without being informed about it, while the psycho-physiological data were recorded. After the movies, a cognitive effort test was performed, recording data during one minute while the subject had to perform some arithmetic operations. Finally, the subject was asked to fill a questionnaire.

The psycho-physiological data are interpreted within the *arousal-valence* model [Bradley and Lang (2007)], according to which every emotion is represented as a point in a two-dimensional space formed by the arousal dimension, which expresses the intensity of the emotion, and the valence dimension, which expresses its positive or negative connotation. In this respect, EMG and EDA measurements are often used as emotional and attentional indicators of the level of immersion induced by audiovisual material [Sanchez-Vives and Slater (2005)]. Various studies have demonstrated that an increase in the EMG activity registered in the corrugator supercilii muscle (situated just above the eyebrows) is a good indicator of the presence of negative emotions, and that variations in the EDA levels reflect variations in arousal [Larsen et al. (2003)], which can indicate increases in emotional activation or cognitive effort, depending on the conditions.

The individual differences between subjects in psycho-physiological data were bigger than the differences within subjects between conditions, so it was necessary to standardize the measurement values in order to make them able to be compared. Two ways for normalizing data have been tried:

1. Z-scores: within a subject, the mean value of all measurements is subtracted from each value, and the result is divided by the standard deviation of all values. By doing so, the new values are expressed in stardard deviations from the mean for each participant.

2. Baseline subtraction: the mean value of the baseline is subtracted from each measurement. The result is in the same unit of the measured data (e.g. mV, $m\Omega$, etc.).

The analyses have been carried out with both methods, in order to explore which of them could provide better results. Each movie has been divided in epochs 10-second long. Since the duration of the two movies is slightly different (The library has duration of 350 s, which amounts to thirty-five 10-second epochs, while Sintel has duration of 210 seconds, corresponding to 21 epochs), it was necessary to exclude some epochs of the first movie in order to compare the values of each subject in each epoch of each movie. We present here the results of a within-subjects analysis, in which values for each subject in each audio condition have been compared, regardless of the "movie" variable. This analysis has been carried out using standardized data, applying the two aforementioned standardization methods. Standardization of data by z-scores has been carried out as follows: for each subject, the mean value and the standard deviation between the first 21 epochs of each movie have been calculated. The z-score value of each epoch in each participant has been calculated by subtracting the mean of all epochs in both movies and dividing the result by the standard deviation. By doing so, the different values in each subject are expressed in standard deviations from the participant mean. The baseline subtraction standardization method has been carried out by subtracting from each epoch the mean value of the baseline period of that participant. Both standardization methods have given very similar results, although the baseline subtraction method gives a higher standard deviation per epoch and condition, which reduces the significance of the results. Figure 5.1 shows the EDA z-scores mean values of each epoch for the 24 participants between the two audio conditions. As can be seen in the graph, the lines representing each audio quality have a parallel path. This path reflects the effect on EDA of both movies content, that is similar between both audio conditions. The difference between both lines in Y axis shows a higher EDA activation in 3D audio condition, and this difference is statistically significant ($N=21$; $T=14.9$; $p<0.001$), suggesting a higher arousal provided by 3D audio.

Within-subjects analysis results also show a higher statistically significant EMG activity in 3D audio condition ($N=21$; $T=7.54$; $p<0.001$), as shown in Figure 5.2. These results suggest the presence of more negative emotions provided by audio 3D condition. As both films are expected to produce negative emotions (such as suspense or fear), we can consider that the better spatial localization enhances the valence of the experienced emotions, at least in the case of this kind of negative emotions.

Results of heart rate analysis are not so robust as EDA and EMG data are; furthermore, its interpretation is rather more difficult. With the z-scores standardization, statistically significant higher values for the 3D audio condition were obtained ($N=21$; $T=3.08$; $p=0.006$), as shown in Figure 5.3

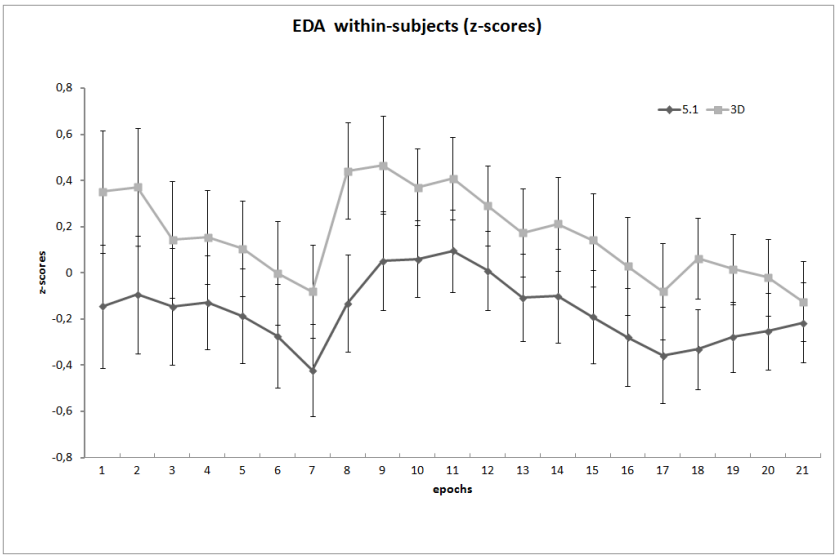

Figure 5.1: Electro dermal activity within subjects comparing 5.1 and 3D audio, as a function of the time epoch of the movies.

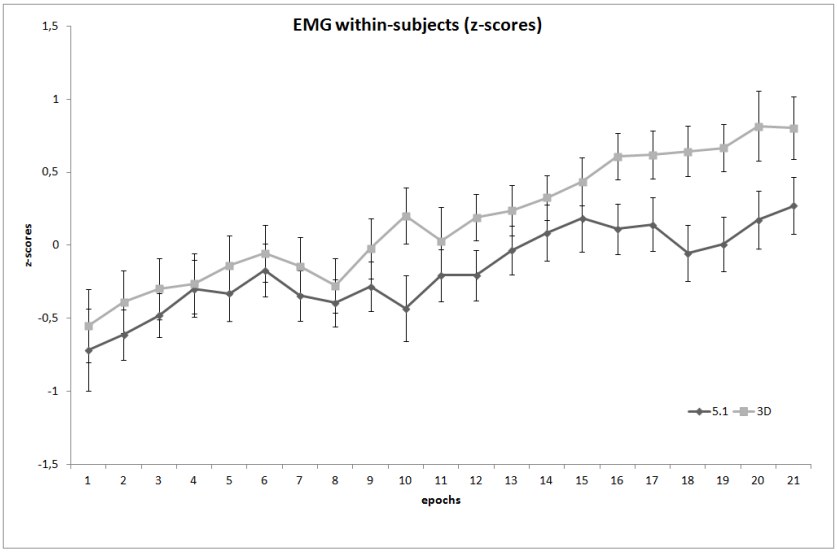

Figure 5.2: Facial electromyography within subjects comparing 5.1 and 3D audio, as a function of the time epoch of the movies.

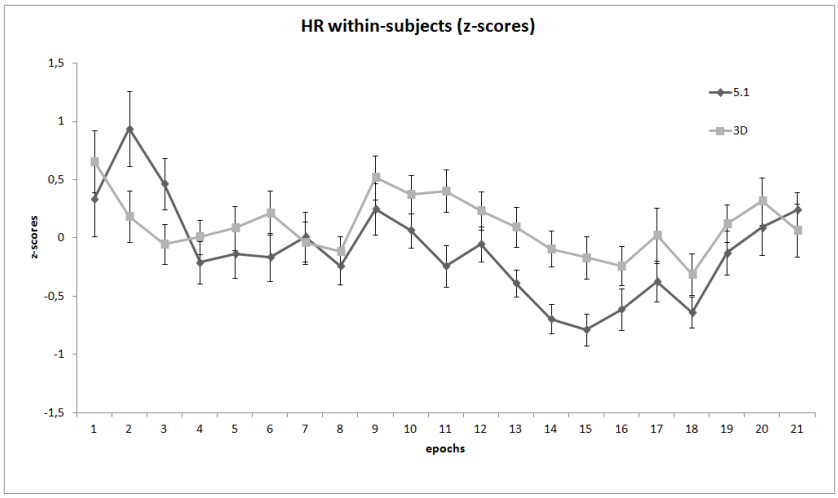

Figure 5.3: Heart rate within subjects comparing 5.1 and 3D audio, as a function of the time epoch of the movies.

Literature reviewed shows evidences that HR can be used as an index of both attentional engagement and of emotional responses, but studies on this field have found contradictory results. Related to the use of HR as a stimulus valence index, Ravaja (2004) says: "In particular, during perception (e.g., media viewing) HR is virtually useless as a measure of arousal. Although, in general, HR decelerates in response to both unpleasant and pleasant pictures, study participants exhibit relatively more HR deceleration when viewing unpleasant, compared to pleasant, pictures, suggesting that HR is sensitive to the valence of stimuli during perception". The higher HR level found in 3D audio condition could be interpreted as the presence of more positive (or less negative) emotion than in the 5.1 audio condition, which is contradictory with EMG results. Given that the EMG results are quite robust and do not present this ambiguity, interpretation of heart rate results as an index of less negative emotion provided by 3D audio condition should be rejected. Also cited by Ravaja (2004), Palomba, Sarlo, Angrilli, Mini, and Stegagno found HR accelerations in threatening films, showing that some particular kind of emotional stimuli could give a different HR reaction path. As long as we can considerer the movies used as a stimuli as threat eliciting, this could be an explanation for the higher HR level in 3D audio condition, meaning that the threat felt by participants would have been increased. This explanation is consistent with data from EMG, so it should be taken into account. No

significant results have been found in the RSA epochs analysis, with any standardization method.

At the end of the session, each participant had to fill a questionnaire, answering the following questions:

1. What is your overall impression of the first movie?
2. How comfortable was the visual experience in the first movie?
3. How much did the visual aspects of the first movie involve you?
4. Please rate the quality of the audio in the first movie.
5. How much did the auditory aspects of the first movie involve you?
6. How compelling was the sense of movement in the first movie?
7. How realistic did the space in the first movie feel?
8. How well could you identify sounds in the first movie?
9. How well could you locate sounds in the first movie?

The same questions were repeated for the second movie. Each question could be answered by a rating from 1 to 10 in increments of 0.5, 1 being the negative extreme and 10 the positive one. Some of the participants were asked explicitly if they perceived a difference between the audio of the two movies: in most cases, the answer was negative. The responses to the questionnaire were used to rate two aspects: *quality* and *immersion*. The analysis was done by taking the average and standard deviation of the ratings for each audio format. Only slight differences in terms of immersion and quality were observed between the 5.1 and 3D conditions: the quality rating was 8.3 for 3D and 8.0 for 5.1, while immersion was rated 8.0 for 3D and 7.9 for 5.1; not only the average ratings are too close, but the standard deviation is in the order of 2 for each group, which indicates no evidence of significant difference between the ratings. The only conclusion we can draw is that the high level of rating for both audio formats indicates that the overall quality and immersion was perceived as being high in all cases.

The results of the comparison of 3D versus 5.1 content evidence that the arousal caused by the 3D audio is greater than the one caused by 5.1 audio, although the questionnaires do not indicate such evidence, meaning that the difference is mostly on a subconscious level. In particular, higher EDA values in the 3D audio condition show an increased arousal of the emotion provided by this format, while the higher EMG values indicate a more negative quality of this emotion. The genre of the contents used as stimuli in the experiment is expected to elicit negative emotions as fear of suspense, so we can consider that the valence of the emotion elicited by the movie on viewers has been increased by 3D audio condition. Consequently, the increased arousal and

valence must be interpreted as a greater emotional experience provided by 3D audio compared with 5.1 audio.

One limitation of this study is related to the kind of contents used as stimuli. Media contents can elicit a number of different emotions, and in this experiment only emotionally negative contents have been tested. Further investigation should analyze if the increase of emotional experience provided by 3D audio is also produced in emotionally positive contents, or even in emotionally neutral contents.

The different emotional impact was not reflected in the answers to the questionnaire, confirming that indirect measurements of users' reactions provide a more complete picture of users' experience.

As a last remark, we want to stress that the participants were not audio experts and approached the task as normal consumers and movie-goers. No indication was given that they should pay special attention to the audio. On the other hand, self reports of audio professionals who have worked with both standard and 3D surround indicate that the conscious perceptual difference between the two formats is quite noticeable. Following this hint, further work in this field should include the same test with professional users, to see to which extent the different audio formats affect their conscious and unconscious response.

5.2 Evaluation of spatial masking

The next step in the research on perceptual aspects of 3D audio regards the evaluation of masking and how it is affected by the spatial distribution of sound. Masking is the phenomenon where an otherwise audible sound is not perceived in the presence of another simultaneous sound (masker). The masking effect depends on the relationship between loudness and frequency of the masked sound and the masker. The phenomenon has been studied extensively for monaural sounds, typically pure tones masked by band-filtered noise; in this context, masking is related to the hearing's frequency resolution: when two tones fall in the same critical band, the loudest one can make the softer one inaudible. However, thanks to our hearing's ability to separate sounds coming from different directions and focus on the desired sound source (the cocktail party effect), we can effectively use spatial hearing to reduce the masking effect in those cases where the masker and the masked sound sources have different directions of arrival. Some researchers have found that there is advantage in the detection of the masked sound if the sources are spatially separated [Hawley et al. (2004); Kidd et al. (2005)]. It seems reasonable to state that if the sources that compose a soundtrack are distributed in the 3D space, they are easier to identify for a listener, compared to when they are collapsed to the horizontal plane or to the front area. In order to support this statement, we designed an experiment to detect the masking threshold of a hidden tone masked by band filtered noise

for different spatial distributions of masking noise. Three playback formats were considered: stereo, 5.1 and 3D. For each audio format, two hidden sounds were used: pure tones at 500Hz and 1kHz, with a duration of 100 ms. As masking sounds, filtered white noise has been employed, spanning 400 Hz to 1.6 kHz to mask the hidden tone at 1 kHz and 200 Hz to 800 Hz for the hidden tone at 500 Hz. In the case of stereo, the left and right channels are fed with two independent, uncorrelated bursts of noise each one with an attenuation of $1/\sqrt{2}$ (3 dB). For 5.1, the hidden tone is played from the center channel, while the uncorrelated masking noise is played from channels L, R, Ls and Rs. In 3D the hidden tone is still played from the center channel, while the masking noise is played through sixteen loudspeakers, quasi-isotropically distributed in the upper hemisphere. In the 5.1 and 3D case, each uncorrelated filtered noise channel is attenuated by a factor $1/\sqrt{n}$, where n is the number of loudspeakers that participate in the playback of the noise, respectively four and sixteen for the 5.1 and 3D cases. These attenuation factors correspond to 3 dB every doubling of the number of channels, which is in agreement with the fact that summing two uncorrelated sources with equal level gives an increase of 3 dB. The test has been performed in a soundproof, acoustically conditioned studio approximately 7m x 5m x 3m, with a decay time less than 0.3s in the whole audio spectrum. The loudspeakers used for playback are Genelec 8040; they are equalized and their level and time of arrival in the listening position has been adjusted by means of a Trinnov Optimizer digital audio processor. The equal loudness of the noise produced in the three conditions has been verified by means of an SPL meter; the level of the masking noise in each audio format during the experiment was set to 70dB C-weighted. The duration of the masking noise is 300ms, while the hidden tone has a duration of 100ms and is centered with respect to the length of the masking noise. All sounds are onset and offset with a fade by means of two raised cosine onset and offset gates of 10ms. In this experiment, for each condition a fixed level was used for the masking noise, while varying the level of the hidden sound. The level of the hidden sound is expressed in dB relative to the masking noise level. The purpose is to determine the audibility of the hidden sound by measuring its detection threshold. The experiment has been conducted using a two-alternative forced choice design (2AFC): in each trial, two noise bursts are played back sequentially, one of them containing the hidden tone. The order of the presentation of the hidden tone is chosen randomly at each presentation. The subject has to indicate if the hidden tone was contained in the first or in the second noise burst. The level of the hidden tone is varied adaptively in each trial according to the previous response of the subject, following a staircase procedure. The experiment was implemented in Matlab, adapting existing functions of the MLP routines [Grassi and Soranzo (2009)]. After a certain number of trials, the fraction of correct answers was computed as a function of the level of hidden sound. As in many perception experiments, this curve is the psychometric function of the subject: it expresses the probability of

detection of the stimulus as a function of its intensity. At levels well above the threshold, the fraction of correct answers is one (neglecting user's mistakes), while well below the threshold one assumes that the answer is given by chance, therefore the probability of obtaining correct answers is $1/n_{FC}$, where n_{FC} is the number of choices in the experiment. For a large number of trials and a 2AFC experiment, the fraction of correct answers below threshold tends therefore to 0.5. The psychometric function expressed in terms of logarithmic intensity follows the typical behavior shown in Figure 5.4. The plot shows the fraction of correct responses versus stimulus amplitude, for a 2AFC experiment. A good analytic representation of typical psychometric functions is given by the Weibull function [Weibull (1951); Watson (1979)]:

$$w(x) = 1 - (1 - g)e^{-(kx/t)^b} \tag{5.1}$$

In this representation, x is the stimulus level; the parameter t represents the threshold, which is defined as the abscissa where the function value corresponds to 80%; b represents the slope of the function and g is the performance at chance (0.5 in the case of 2AFC); k is a free parameter. The threshold of perception is usually obtained by fitting the experimental data to a Weibull function and obtaining the parameter t. Figure 5.4 shows the Weibull fitting function overlapped with the data.

Three subjects participated in the experiment; subject A reported extensive previous experience with spatial audio, while subject B reported limited experience and subject C no experience. Each subject completed four sessions with one hundred and sixty trials in each run, for each of the three audio formats and each of the two tone frequencies. This corresponds to a total of 640 responses for each frequency and audio condition. The order of execution of the sessions with respect to the audio format was the same for each subject: the three audio formats were interleaved. The tests at 500Hz were performed after all the tests at 1kHz. Data were collected and analyzed by means of Matlab scripts. Two types of analysis have been performed:

- between subjects: for each subject, thresholds were calculated for each frequency and audio condition, joining all the answers with the same frequency and audio condition;
- between conditions: threshold were calculated by joining the three subject's data for each condition and frequency.

Less than ten answers, admittedly recognized as mistakes by the subjects as being wrong responses for perceived stimuli well above the threshold, were removed from the data. In fitting the data with the Weibull psychometric function, the minimization algorithm provided by Matlab's function *fminsearch* was employed, using the threshold and slope of the Weibull function as search parameters and obtaining them as results of the fitting. As already discussed, the threshold is the value of the stimulus where the probability of

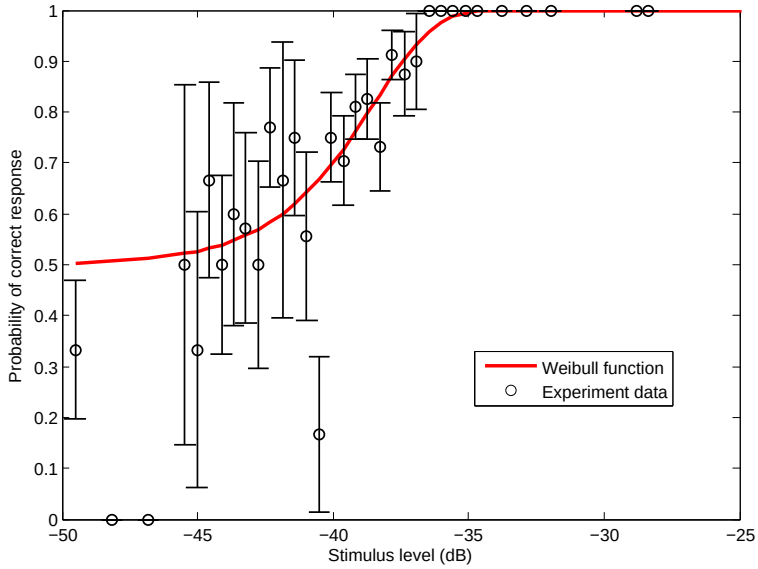

Figure 5.4: Experimental data of a generic threshold detection experiment for a single subject and corresponding fitting of psychometric function with the Weibull curve. The threshold is the abscissa where the function value is 0.8 (in this case -38 dB).

a correct answer is 80%. This value is used in the following discussion as the threshold of perception for the hidden tone. The slope of the psychometric function is an indication of how well the threshold can be determined from the experimental data: a steep slope indicates a sharp transition between certain perception and guess, meaning that the uncertainty of the threshold spans a relatively small interval. On the other hand, data which present a mild slope indicate higher uncertainty in the value of the threshold. In order to give an estimation of the uncertainty associated to the threshold, the bootstrap method was applied to the analysis of the data [Foster and Bishof (1997)]. Given an experiment with N trials and the associated answers, the boostrap method consists in resampling the data M times, each time choosing N samples with repetition among the original data. For each of the M new samples, the threshold is calculated in the usual way and the average and standard deviation of thresholds are presented as the resulting threshold and confidence interval. More precisely, the confidence interval is

Subject	Condition	Threshold (dB)	C. I. (dB)
A	stereo	-34.13	-35.02; -33.13
A	5.1	-37.92	-38.87; -36.56
A	3D	-40.05	-40.85; -38.86
B	stereo	-30.55	-31.42; -29.61
B	5.1	-34.11	-35.12; -32.89
B	3D	-39.95	-40.67 -39.15
C	stereo	-32.04	-32.68 -31.34
C	5.1	-33.29	-34.76; -30.25
C	3D	-34.69	-35.93; -32.77

Table 5.1: Masking threshold and confidence interval for each subject and audio format at 500 Hz.

the stimulus interval where the threshold falls with 95% probability. For the bootstrap analysis we chose 500 samplings of each data group. This choice is a trade-off between accuracy in the estimation of the confidence interval and the time needed to run the algorithm. Beyond five-hundred samplings, the distribution of the results was not improving significantly.

Figure 5.5 shows an example of the analysis that was done for each condition: the top graph shows the fraction of correct answers versus stimulus level for a given subject and audio condition; the red curve is the fitting function of the overall data, from which an estimation of the threshold can be obtained; the plot in the middle shows the number of trials that have been performed for each stimulus level, showing the efficiency of the staircase method in concentrating the trials near the threshold; finally, the bottom plot shows the distribution of five-hundred thresholds calculated from bootstrap sampling of the original data. In this example, the threshold is $-41.69\ dB$ and the 95% confidence interval is $[-42.28\ dB,\ -41.05\ dB]$.

Figures 5.6, 5.7 and 5.8 show the measured data and corresponding fitting functions for each frequency and audio condition for subjects A, B and C respectively. The number of trials per stimulus level and the bootstrap distribution of thresholds have been omitted from the plots, since they always looked similar to the example of Figure 5.5. Tables 5.2 and 5.2 report the resulting thresholds and confidence intervals for 500 Hz and 1 kHz respectively.

The threshold level is expressed in dB relative to the masking noise level. The lower (more negative) the value, the lowest the detection threshold of the hidden sound, therefore the lower the masking effect. For ease of visualization, the thresholds and confidence interval for each subject and audio condition are shown in the histogram of Figure 5.9. The results show that for each subject the masking threshold is lower when the masking sound is spread in 3D, while the highest values correspond to stereo. For subjects B

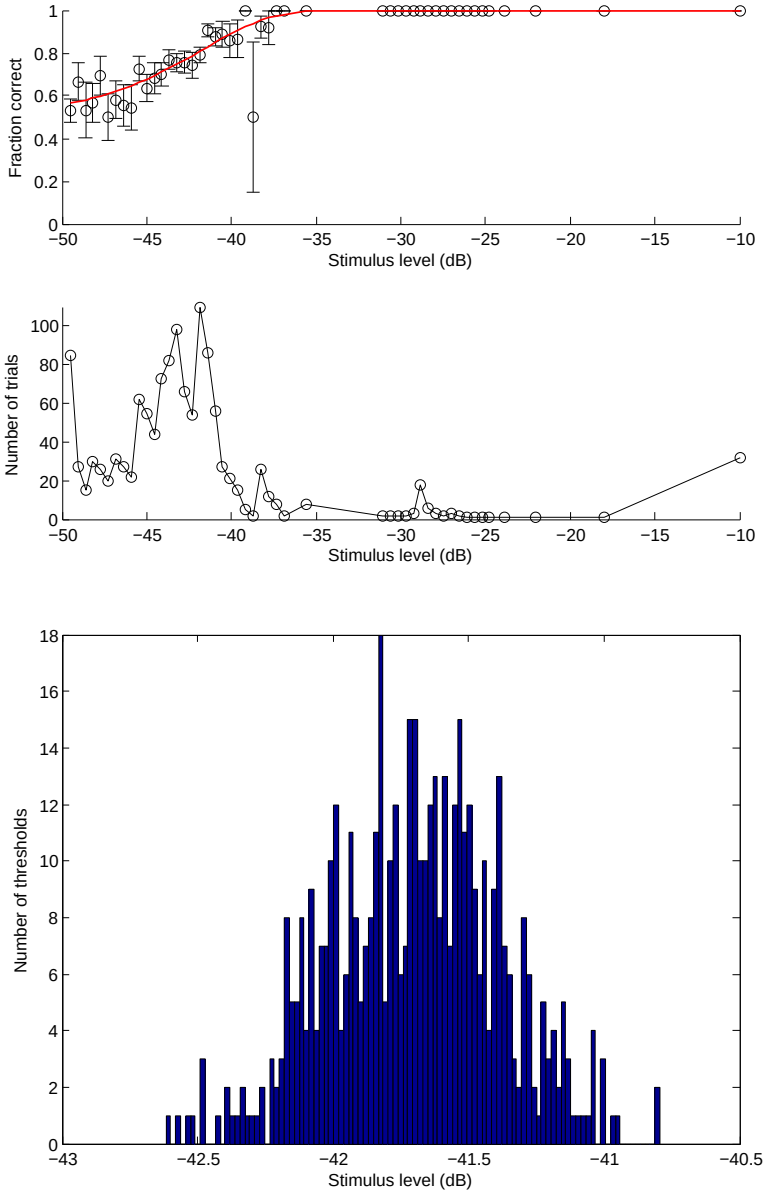

Figure 5.5: Example of data and results of threshold detection experiment for the 3D audio condition; top: experimental data and fitting function; middle: number of trials per stimulus level; bottom: distribution of thresholds after bootstrap sampling.

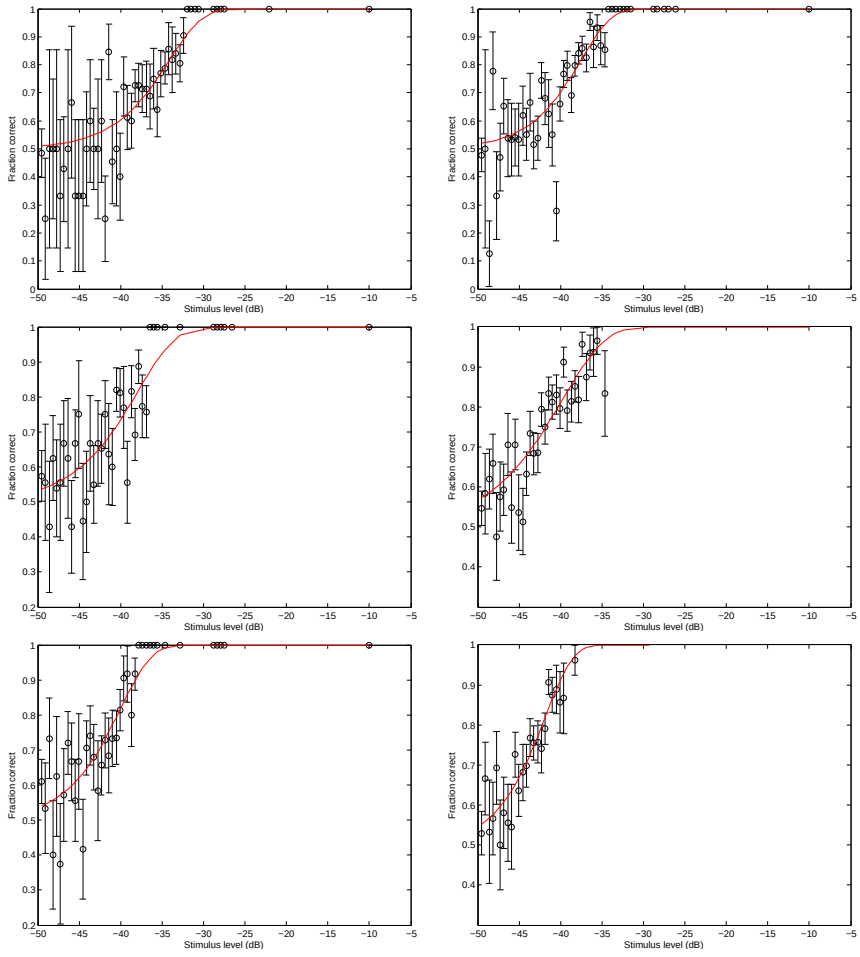

Figure 5.6: Psychometric functions for subject A at 500 Hz (left) and 1 kHz (right); from top to bottom: stereo, 5.1 and 3D.

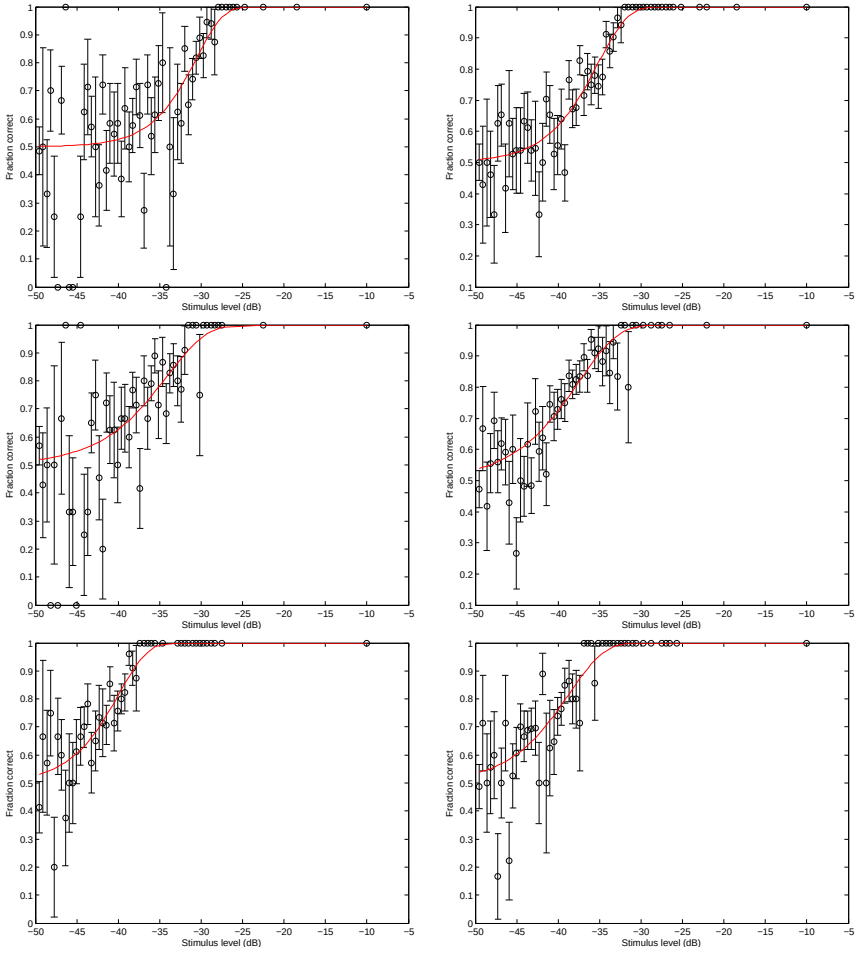

Figure 5.7: Psychometric functions for subject B at 500 Hz (left) and 1 kHz (right); from top to bottom: stereo, 5.1 and 3D.

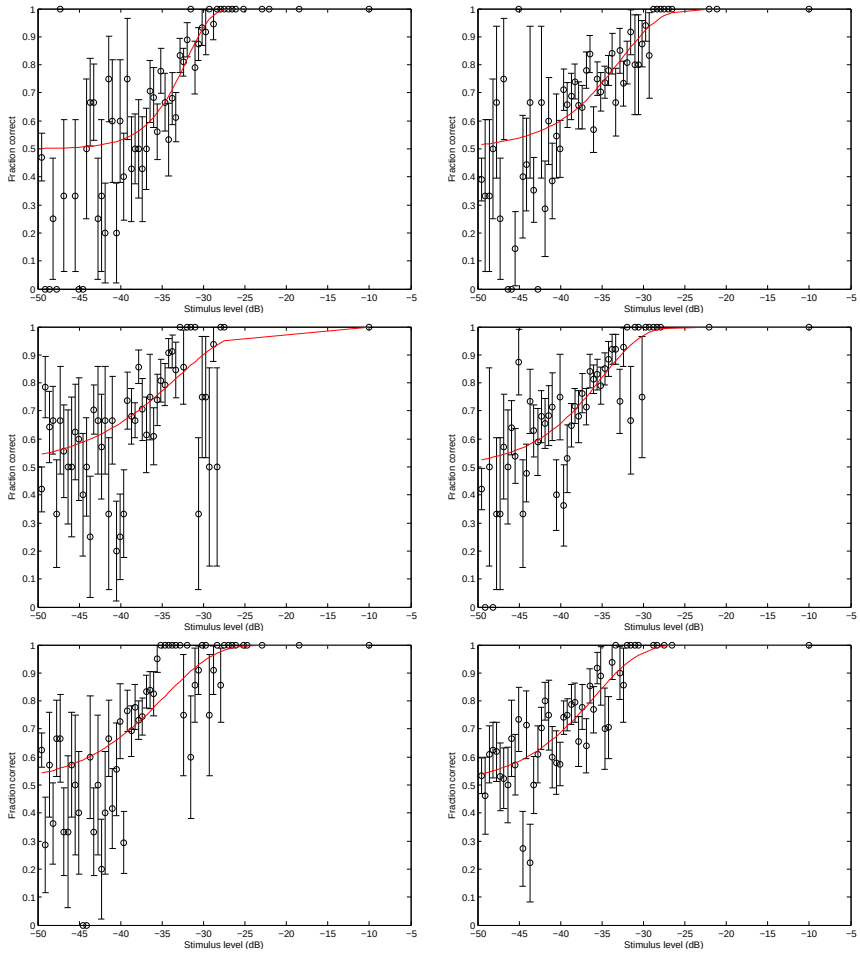

Figure 5.8: Psychometric functions for subject C at 500 Hz (left) and 1 kHz (right); from top to bottom: stereo, 5.1 and 3D.

Subject	Condition	Threshold (dB)	C. I. (dB)
A	stereo	-37.56	-38.09 -36.97
A	5.1	-39.83	-40.51 -39.19
A	3D	-41.93	-42.41 -41.36
B	stereo	-35.22	-35.74 -34.64
B	5.1	-37.52	-38.19 -36.75
B	3D	-38.71	-39.64 -37.72
C	stereo	-32.79	-33.59 -31.57
C	5.1	-35.08	-35.89 -34.06
C	3D	-35.88	-36.85 -34.72

Table 5.2: Masking threshold and confidence interval for each subject and audio format at 1 kHz.

and C there is an overlap in the confidence interval between 5.1 and 3D at 1 kHz; otherwise, the results are significant, and the thresholds are unambiguously separated. Overall, users experience a difference in masking threshold of approximately 1.5 to 2 dB between each audio format. Although the shift in thresholds caused by the audio condition is similar between subjects, the absolute position of the thresholds varies according to the experience of users: the center of mass of thresholds is lower for subject A and higher for subject C; this likely reflects the different level of ear training between the subjects. An overall shift in the thresholds is also present between 1 kHz and 500 Hz, showing that the perception of the hidden tone improves at 1 kHz. This is probably related to the increased sensitivity of our hearing to tones in this range. Despite the absolute position of the thresholds, the results indicate that each subject can discriminate the hidden tones better if the other sound sources are spread in 3D.

Let us now consider the joined data of all subjects: figure 5.10 shows the data and psychometric functions obtained joining the responses of all subjects in each condition, while the threshold values and confidence intervals are reported in Figure 5.11. The confidence interval is now smaller, thanks to the increased number of trials considered. In this analysis the threshold is independent from the subject, and the differences between audio conditions are still close to 1.5 dB at 500 Hz, while at 1 kHz the confidence interval does not allow to clearly draw a winner between 5.1 and 3D. The reduction in the difference between thresholds is due to the fact that now both experienced and inexperiences users are mixed.

These results show that there is a significant difference in the perception of hidden tones if the masking noise is distributed over different portions of the panorama, keeping its level constant.

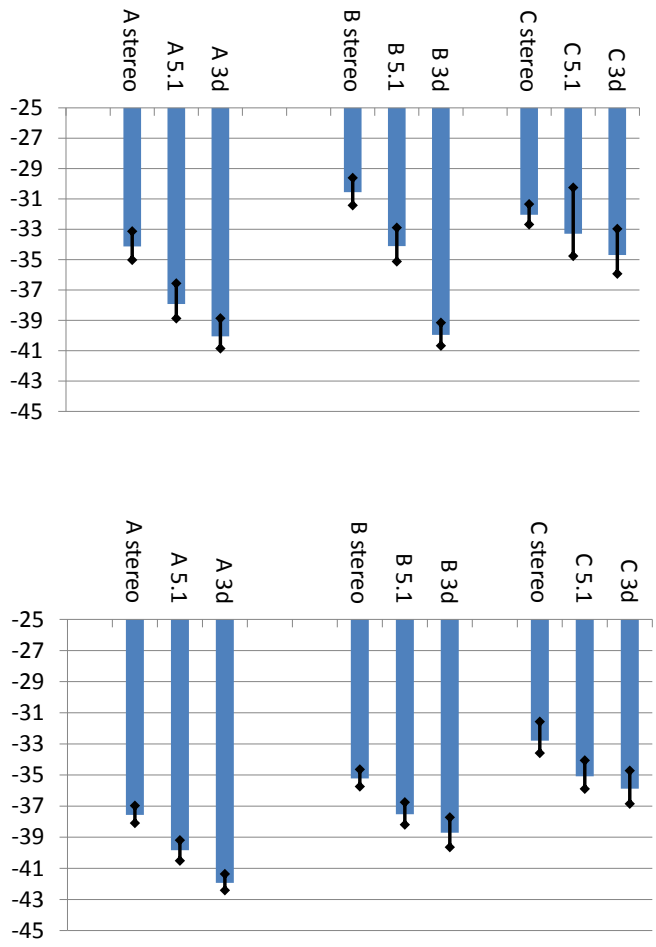

Figure 5.9: Histogram of thresholds and confidence intervals for each subject and audio condition. Top: 500 Hz; bottom: 1 kHz.

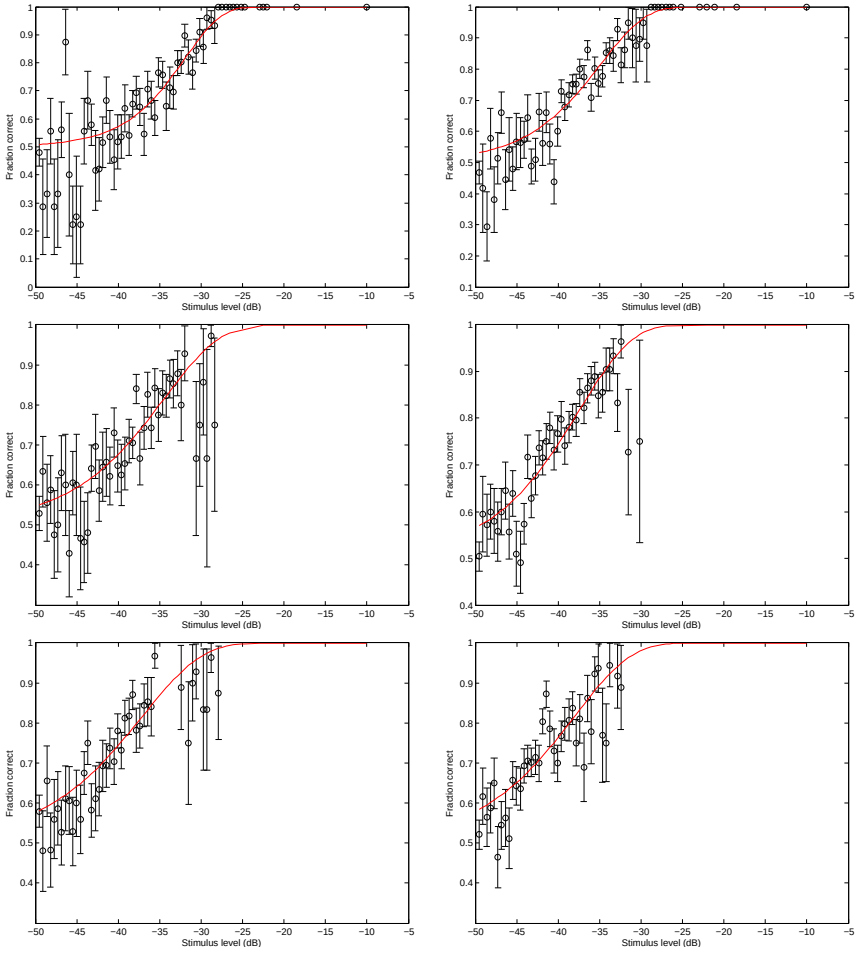

Figure 5.10: Psychometric functions of all subjects, joined, at 500 Hz (left) and 1 kHz (right); from top to bottom: stereo, 5.1 and 3D.

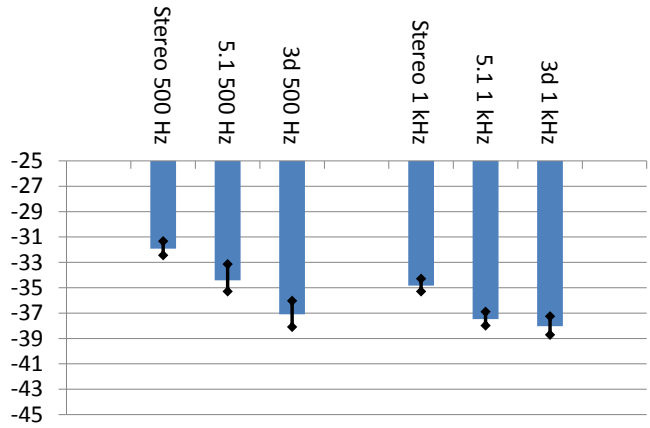

Figure 5.11: Histogram of thresholds and confidence intervals for joint data in each audio condition.

5.3 Subjective adjustment of reverberation levels

The last experiment presented here is a test to evaluate if the perception of the reverberation level varies with its spatial distribution. Reverberation is often added to audio tracks in music and movie post-production, to blend sources together and help create a realistic and compelling sense of space. Sound engineers choose the reverberation according to the characteristics of the space they desire, manipulating parameters such as reverberation time and density of early reflections. Once the reverberation is chosen, they mix it with the dry audio tracks, and their relative level determines the level of the reverberation that is perceived. In a relative way, the level of reverberation can therefore be defined as the level of the reverberation channel compared to the dry source. The level of reverberation is chosen according to aesthetic criteria, however it is often bound between certain limits, to avoid loss of intelligibility for the dry sources. Having seen that the perception of 3D sound and the way distributed sources interact or mask each others is different than in 5.1 or stereo, we wanted to verify if the level of reverberation is perceived differently in each audio format. For this purpose, an experiment was designed where the subjects were asked to adjust the level of a reverberation track to make it match perceptually the level of a reference reverberation track. For this experiment we chose a dry recording of a speech, about ten seconds long, spoken by a male actor in English, with the tone and rhythm of a normal conversation. A reverberation track was generated using a commercial convolution reverberation software which provides

high quality impulse responses recorded with various surround microphone setups. Two similar but uncorrelated surround reverberations, of four channels each, were used, convolving the dry audio with each channel, to obtain a total of eight uncorrelated reverberation tracks with similar acoustic characteristics. The chosen impulse responses were recorded in large churches, with spaced microphones positioned in the far field, therefore the degree of correlation between the channels is very low. Assuming that diffuse field reverberation is uncorrelated when recorded with spaced microphones, we used the channels of the reverberation to reconstruct a virtual 3D reverberation by placing each channel in a different position of space using a VBAP panning algorithm. The same studio and equipment used in the previous experiments was employed. Three formats were chosen for testing: 3D, 5.1 and stereo. For 3D, the eight channels of reverberation were distributed uniformly over the upper hemisphere and decoded to the studio's twenty-three loudspeaker setup. In 5.1, the sources were projected on the horizontal plane using a decoding algorithm that preserves the energy compared to the 3D case, while in stereo the channels were assigned to the left and right speakers. The verification of the loudness matching of reverberation when played back in the different formats was done by measuring the SPL produced by the reverberation applied to stationary pink noise and decoded to the three formats. In all cases the reverberation level produced in the listening position matched within a half dB.

Our experiment consist in having the subjects match perceptually the level of reverberation in a given format to a reference reverberation level in another format. The dry voice was send to the center channel in all cases, using a center speaker for stereo as well, to make sure the level of the dry voice was not affected by the stereo panning law of the workstation employed. In order to establish a common reference, the reverberation in 3D and the dry voice were mixed to the taste of the author, to give the impression of being in a large, reverberant space without loosing the intelligibility of the speech. This level was chosen as the reference level for 3D, and the same level applied to the 5.1 and stereo reverberations established the reference level for the respective formats. Let A, B and C be 3D, 5.1 and stereo respectively. Five subjects were asked to match the level of A vs B, A vs C, B vs C and all the inverse combinations, where the first format in each comparison represents the fixed reverberation and the second is the one adjustable by the user. Following this scheme, a total of six experiments, each one repeated five times, was performed by each subject. The test was carried out using the method of adjustment, letting each subject adjust the fader of the variable reverberation to match perceptually the reference. The resolution of the fader is 0.1 dB. The participants were not informed about the difference between the conditions and were not able to see the numerical value of the fader's position. Besides, a hidden gain trim was used to introduce a shift in the fader position at each experiment, to avoid subjects memorizing the position of the fader from previous trials. No time constraints were given to

the subjects. On average, each subject was able to perform each trial in less than two minutes. When the subject concluded each comparison, another person recorded the position of the fader, without letting the subject know the value. Three of the subjects had previous experience with audio, while two had limited or no previous experience.

Figure 5.12 shows the mean and standard deviation of the adjustment for all subjects in each comparison, after having compensated the trim shift. The value of the average in dB is the difference with the reference reverberation level in each test. If no difference were perceived, all the scores would be 0dB. As we can see in the histogram, the difference between the various conditions is very small, and the standard deviation does not allow to interpret the differences as clearly significant. Certainly, compared to the previous masking experiments, the differences in the perception of reverberation in the three formats are smaller. Still, the data indicate a trend in the perceived reverberation: the adjustment of 3D reverberation has always resulted in higher level than the 5.1 and stereo references; stereo has always required less level than 3D and 5.1; finally, 5.1 required more level than stereo but less than 3D. These clues can be summarized saying that the higher the number of channels, the softer the actual reverberation level is perceived. However, the difference we found is small, in the order of half a dB between each pair of conditions, and the standard deviation does not allow to draw definitive conclusions. In any case, the difference between formats is so small that the results seem to support the hypothesis that the reverberation is perceived with equal loudness in the three conditions.

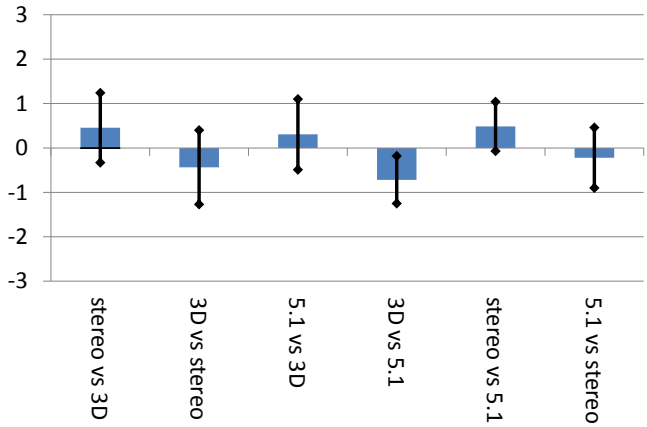

Figure 5.12: Histogram with mean and standard deviation of adjusted levels of reverberation for each comparison.

5.4 Conclusions

Three experiments have been performed to study perceptual aspects related to the differences between 3D audio and standard formats, namely 5.1 surround and stereo. Firstly, the psycho-physiological data collected while subjects watched short movies with 3D or 5.1 audio indicate that the intensity and valence of the emotions provoked by the content are amplified when 3D audio is used. This can be interpreted as the subject being more immersed into the audiovisual experience. Curiously, the difference between the two audio formats was not high enough to make them consciously distinguishable to the ears of the subjects, as indicated by the responses to questionnaires where they had to evaluate the audio quality and spatial accuracy of the two formats. The subjects used for this test were not familiar with audio technology, therefore the results give a good description of the effects of 3D audio for the non-technical audience.

The second experiment was carried out to study how the audio format and the spatial distribution of sound affect the masking phenomenon. The measurement of the psychometric function for the perception of a hidden tone masked by a spatially distributed noise in stereo, 5.1 and 3D has evidenced that the threshold of detection of the hidden sound is lower when the noise is distributed in 3D compared to stereo and 5.1. The difference between each format is about 2 dB for both experienced and inexperienced subjects. Roughly we could quantify this difference as an improvement of 2 dB per dimension, since stereo is one-dimensional and 5.1 is bi-dimensional. The level of ear training of the subjects is reflected in the values of the thresholds for each condition, but the differences between conditions within each subject are comparable. The results of this experiment support the subjective impression that sound sources are easier to perceive, identify and separate when they are located in separate directions compared to when they collapse on a plane or on a line. Therefore, when mixing in 3D the sound engineers can make richer soundtracks and more subtle effects than what is normally allowed in stereo or horizontal surround.

The last experiment we presented studied the perception of the level of reverberation in the three aforementioned formats. By means of the method of adjustment, subjects with or without audio experience were asked to match the level of a reference reverberation by adjusting the faders of a workstation. The experiment included cross comparisons between all the formats. The results are rather difficult to interpret, due to the relatively large dispersion of the data, in the order of 1.5 dB. Considering the mean scores, there is a slight difference, around 0.5 dB, between 3D to 5.1 and 5.1 to stereo; this indicates that the broader the spread of reverberation, the higher the level needed to make it perceptually comparable with a reference. However, this difference is so small, especially compared to the differences in masking thresholds, that we could as well assume that no significant difference is found in the perceived level of reverberation between formats. This result,

together with the previous finding on masking threshold, could be interpreted in the following way: the perceived level of reverberation has little (if any) dependency on the playback format, but its masking effect on the direct sound of sources decreases with 3D audio; therefore, one can exploit 3D audio to increase the level of reverberation where it is required, thus delivering a higher perceived level without reducing the intelligibility of the sources. More data from additional subjects is needed to clarify this point.

6 Conclusions and future developments

The topic of this thesis is the broad subject of 3D audio technologies, and the aspects we have addressed here are some specific research problems that fit within different aspects of the work-flow. Three areas have been considered: firstly, the recording part, with the research on transducers for capturing the spatial characteristics of a sound field; then, the post-production, where we have introduced tools for the control of mixing consoles and for the detection of clipping in layout-independent systems; finally, we have studied some perceptual effects of 3D audio in comparison to other standard formats. Doing research in all these aspects has allowed us to get an overview of the whole picture, seeing how many different technologies play a role towards a common goal: bringing a three-dimensional audio experience to the listener. Together with the overall overview of a relatively new field, the author also became aware that many of the aspects dealt with are open to further development or different approaches. In this last chapter we present a brief summary of the results of our research, together with a discussion on possible future developments in each topic.

In the recording part, one of the main problems is obtaining a simple and compact microphone solution with higher directionality. Given the promising features of anemometric probes in terms of compactness and accuracy of polar patterns in the whole frequency range, our goal was finding a way to employ them for higher-order Ambisonics recording. The research has begun with a study and comparison of the characteristics and performance of tetrahedral microphones and anemometric pressure-velocity probes, with the aim of characterizing the strengths and limitations of each transducer topology by comparing their output in various fields of known characteristics; the purpose of this effort was to assess the suitability of anemometric probes for accurate measurements: these transducers have in fact been employed in the main part of the chapter to implement a device that uses the Euler equation to derive second-order harmonics of the acoustic field by finite differences of first-order signals. This method, if implemented with ideal transducers, allows to obtain the first-order signals directly, without inaccuracies caused by spatial aliasing or transducer's mismatch, and to obtain the

second-order signals with a reduced number of transducers compared to the most employed alternatives. Our results show that the approach is valid and feasible, although the intrinsic inaccuracies and the low SNR of anemometric transducers represent the limiting factor for the system's performance. A more comprehensive comparison and characterization of the transducers would require including a standard p-p sound intensity probe, which is still considered the reference method for sound intensity measurements. However, some works have appeared in literature, such as Jacobsen and De Bree (2005) and Tjis et al. (2009), that have validated the direct p-v approach for sound intensity measurements. Regarding the second-order Ambisonics device, in recent years spherical microphones have undergone serious improvements thanks to both the commercial availability of professional microphone solutions and the development of new techniques for combining the capsules and obtaining the desired signals [Farina et al. (2011)]. In this regard, the next step would be comparing the proposed second-order device with the state of the art of spherical microphones. We expect that spherical microphones would be the winners, both in terms of SNR (thanks to the professional-grade microphones employed) and in terms of accuracy of the polar patterns at all frequencies, thanks to the higher number of transducers employed. Still, the proposed approach could find application in contexts where audio quality is not the main concern, while compactness and maybe resistance to hostile operating conditions are the main requirements.

In post-production, we have focused on two aspects: live broadcasting and channel-free post-production. In live sports broadcasting, one of the main obstacles to the evolution of the audio technology towards 3D is the added complexity and the limited resources that already keep the sound engineer fully occupied to just deliver a stereo feed. Our research here tackled the simplification of the mixing process, automating the capture of sound from the point of interest. The result is the algorithm and device for automated mixing of multichannel live events, which brings a novel way of controlling the faders of a mixing console in situations where the movement of the faders is predictable. The algorithm allows to maximize the pick-up from a certain area of the stage or field, given the position of the microphones employed. Its implementation in a tablet PC connected via MIDI to an audio console has been tested in the context of live broadcasting of football games. The application has turned out to be robust and functional, allowing to greatly simplify the live mixing process. Two main development paths are possible: one regards the extension of the microphone setup employed in football games; while with the current manual procedures adding more microphones is not feasible, using automated control many more microphones could be installed for the benefit of a more uniform coverage: arrays of microphones could even be installed in the field and be mixed with a fingertip! The second possible development is the application of the same concept to other sports. Field sports similar to football could employ exactly the same technology, with little or no adaptation required. Different

sports, such as track and field, track cycling, bob sleigh and others, can be tackled introducing automation in the control point by making it slave to camera tracking or snapshots. The second problem considered in the post-production chapter is the unknown gain change in the output signals of a playback system when the loudspeaker layout is changed: relocating loudspeakers and applying any decoding algorithm to the new configuration results in a different sum of the soundtrack elements to the speakers, which may cause clipping in some channels. This problem was found when decoding actual soundtracks to similar speaker configurations with slightly different positions. Clipping due to decoding is a novel problem which is a by-product of the channel-free production concept. This problem cannot be solved unambiguously without introducing too-strict restrictions, therefore we have proposed a practical method based on the definition of a minimum acceptable loudspeaker density. The proposed algorithm decodes the soundtrack to a reference worst-case layout, defined in terms of the minimum loudspeaker density, and warns if clipping is found. While the proposal has been validated in some ideal and real-world cases, further tests are required to check if the strategy delivers good results in most situations. As regards the implementation, the goal is having the algorithm running in real time for instantaneous feedback to the engineers during a mixing session.

The last part we have considered is the perception of 3D audio compared to other formats. 3D audio has just been integrated into the professional work-flow and is now beginning to reach the audiences. Being a new format, its possibilities have not been thoroughly explored yet. As with other fields of multimedia, the introduction of a new format usually pushes for the development of a new aesthetic language, which depends on its peculiarities and the effects it provokes on the audience. To give an example of how a format can deeply influence artistic aspects, let us consider the changes brought by 3D image over 2D: after initial experimentation, it became clear that with 3D it is not possible to employ quick changes of plans and editing a movie for fast sequences, because the spectator's vision can not adapt as quickly to each change of plan as it does in 2D. Movie sequences have become therefore longer and "slower", while at the same time a great depth of field is employed to bring the whole scene in focus and let the eyes of the audience focus on the subject and blur the background. It is possible that analogous choices will affect the 3D audio production, but firstly the effects of 3D audio on the audience have to be assessed. We started from the opinions and expectations of people who have worked with 3D audio to focus on particular aspects that were constantly brought up in conversations about the advantages and impact of 3D audio. The first claim we considered is that 3D audio is "immersive", in the sense that it brings more involvement to the listener. In order to verify it, we participated in an experiment where subjects watched short movies with 5.1 or 3D audio, while psycho-physiological data such as heart rate, facial electromyography and electro-dermal activity were recorded. These data correlate with the intensity and valence of the emo-

tions provoked by audiovisual content, and the experiment has demonstrated that higher emotional arousal is provoked when 3D audio is employed. Curiously, the effect is subconscious, since non-expert spectators were not able to discern the differences between audio formats, as shown by their responses to suitable questionnaires. The second claim we wanted to verify refers to the ability to perceive better certain separate sources when they are spread in the 3D panorama. Our experiment consisted in measuring the detection threshold of a hidden tone masked by a filtered noise, comparing the cases where the noise is presented in stereo, 5.1 or 3D. The results indicate that the perception of the hidden tone improves with the spread of the masker, and the difference is noticeable for both expert and non expert listeners. This result supports the initial claim. Further experiments could be done employing more subjects and trying different frequencies and directions of arrival of the hidden tones. The last experiment that was considered assesses the perceived level of reverberation between the different formats. Here the results are difficult to interpret due to the high variation of the data; however, the results seem to indicate that the difference, if any, is small and definitely less than what emerged from the masking tests. In order to draw a conclusion, the test shall be extended to a higher number of subjects. However, if we accept the conclusion that no difference is perceived, this supports the hypothesis that a higher level of reverberation can be used in 3D: for example, pushing the reverberation by 3 dB will make the listeners perceive 3 dB more of reverberation, as they would perceive in stereo or 5.1, but will cause less masking and therefore less clarity loss. Overall, the psychoacoustic tests evidence advantages in using 3D audio over standard formats; the field is open to fresh experimentation, since many of the aspects that have been studied in monaural conditions can now be treated in comparison between different formats. Besides, further studies should consider auditory aspects together with the visual ones, to determine for example the mutual enhancement or interference between 2D and 3D audio and image formats.

To conclude, it is worth remarking that, although the field of 3D audio is based on old and well-known principles, now that it is reaching widespread diffusion it is being reviewed with a fresh approach, and many of its facets are being discovered and polished. The technology is maturing and the implemented solutions are growing fast; at the same time, the artistic potential is promising but still largely hidden and unexplored. Under these premises, 3D audio is still a fertile field for research in the near future.

Bibliography

Abhayapala, T. and Ward, D. (2002). Theory and design of high order sound field microphones using spherical microphone array. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages 1949–1952. IEEE.

Allen, I. (2006). The x-curve: Its origins and history. *SMPTE motion imaging journal*, 115(7-8):264–275.

Ardour (2012). The ardour workstation. www.ardour.org. [Online; accessed 14-May-2012].

Auro3D (2012). auro3d. http://www.auro-3d.com/. [Online; accessed 18-June-2012].

Baker, S. (1955). An acoustic intensity meter. *The Journal of the Acoustical Society of America*, 27:269.

Bartlett, B. and Bartlett, J. (1999). *On-location recording techniques*. Focal Press.

Beranek, L. (1954). *Acoustics*, volume 6. McGraw-Hill New York.

Berkhout, A. (1988). A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, 36(12):977–995.

Bertet, S., Daniel, J., and Moreau, S. (2006). 3d sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone. In *Proc. 120th Conv. Audio Eng. Soc. Paris, France*, volume 120.

Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization*. The MIT press.

Blumlein, A. D. (1933). Improvements in and relating to sound-transmission, sound-recording and sound-reproduction systems. British Patent Specification 394. Issued June 14.

Bonsi, D. (1998). *Theoretical and experimental study of energetic properties of confined sound fields.* PhD thesis, Doctoral Dissertation in Physics, Co-supervisors: G. Schiffrer, D. Stanzial, University of Ferrara, academic year 1997-98.

Bonsi, D., Fanucci, L., Fontana, F., Gonzalez, D., L'Insalata, N., Schffrer, G., and Stanzial, D. (2005). Demonstration of measurement and recording for acoustic quadraphony. *IP-RACINE deliverable*, 7(2).

Bonsi, D. and Stanzial, D. (2001). Preliminary study of ambisonics technology as a method for sound intensity measurements. *Proc. International Symposium on Musical Acoustics, Perugia.*

Bonsi, D. and Stanzial, D. (2002). Energetic analysis of the quadraphonic synthesis of sound fields for sound recording and auralisation enhancing. In *Proc. EAA Forum Acusticum, Sevilla.*

Boone, M. (2004). Multi-actuator panels (maps) as loudspeaker arrays for wave field synthesis. *Journal of the Audio Engineering Society*, 52(7-8):712–723.

Bradley, M. and Lang, P. (2007). Emotion and motivation. *Handbook of psychophysiology*, 3:581–607.

Cengarle, G. and Mateos, T. (2011). Comparison of anemometric probe and tetrahedral microphones for sound intensity measurements. In *Proc. 130th Conv. Audio Eng. Soc. London, UK.*

Cengarle, G. and Mateos, T. (2012). A clipping detector for layout-independent multichannel audio production. In *Proc. 132th Conv. Audio Eng. Soc. Budapest, HU.*

Cengarle, G., Mateos, T., and Bonsi, D. (2011). A second-order ambisonics device using velocity transducers. *Journal of the Audio Engineering Society*, 59(9):656–668.

Cengarle, G., Mateos, T., Olaiz, N., and Arumí, P. (2010). A new technology for the assisted mixing of sport events: Application to live football broadcasting. In *Proc. 128th Conv. Audio Eng. Soc. London, UK.*

CLAM (2011). Clam framework main website. `http://clam-project.org/`. [Online; accessed 14-May-2012].

Clapp, C. and Firestone, F. (1941). The acoustic wattmeter, an instrument for measuring sound energy flow. *The Journal of the Acoustical Society of America*, 13:124.

Corteel, E., Horbach, U., and Pellegrini, R. (2002). Multichannel inverse filtering of multiexciter distributed mode loudspeakers for wave field synthesis. In *Proc. 112th Audio Eng. Soc. Munich, Germany.*

Cotterell, P. S. (2002). *On the Theory of the Second Order Soundfield Microphone*. PhD thesis, University of Reading.

Craven, P., Law, M., and Travis, C. (2009). Microphone arrays using tangential velocity sensors. In *Proceedings of the Ambisonics Symposium*.

Daniel, J. (2000). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, University of Paris VI, France, 2000.

Daniel, J. and Moreau, S. (2004). Further study of sound field coding with higher order ambisonics. In *Proc. 116th Conv. Audio Eng. Soc. Berlin, Germany*, volume 6017.

Daniel, J., Nicol, R., and Moreau, S. (2003). Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging. In *Proc. 114th Conv. Audio Eng. Soc. Amsterdam, the Netherlands*.

De Bree, H., Leussink, P., Korthorst, T., Jansen, H., Lammerink, T., and Elwenspoek, M. (1996). The μ-flown: a novel device for measuring acoustic flows. *Sensors and Actuators A: Physical*, 54(1):552–557.

De Bree, H. E. (2007). The microflown e-book. http://www.microflown.com/library/books/the-microflown-e-book.html. [Online; accessed 18-June-2012].

Dolby (2012). Atmos. http://www.dolby.com/us/en/professional/technology/cinema/dolby-atmos.html. [Online; accessed 18-June-2012].

Eigenmike (2012). The eigenmike microphone array. http://www.mhacoustics.com/mh_acoustics/Eigenmike_microphone_array.html. [Online; accessed 18-June-2012].

Elko, G. (2000). Microphone arrays. In Gay, S. and Benesty, J., editors, *Acoustic signal processing for telecommunication*. Springer.

Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proc. 108th Conv. Audio Eng. Soc. Paris, France*.

Farina, A. and Ayalon, R. (2003). Recording concert hall acoustics for posterity. In *24th AES Conference on Multichannel Audio, Banff, Canada*, pages 26–28.

Farina, A., Binelli, M., Capra, A., Campanini, S., and Amendola, A. (2011). Recording, simulation and reproduction of spatial sound fields by spatial pcm sampling (sps). In *Proceedings of the International Seminar on Virtual Acoustics, Valencia*.

Farina, A., Capra, A., Conti, L., Martignon, P., and Fazi, F. (2007). Measuring spatial impulse responses in concert halls and opera houses employing a spherical microphone array. In *19th International Congress on Acoustics (ICA), Madrid*.

Farina, A., Glasgal, R., Armelloni, E., and Torger, A. (2001). Ambiophonic principles for the recording and reproduction of surround sound for music. In *19th AES Conference on Surround Sound, Techniques, Technology and Perception*.

Farrar, K. (1979). Soundfield microphone. *Wireless World*, 85(1526):48–50.

Fascinate (2010). Format agnostic 3d audio. http://www.fascinate-project.eu/index.php/tech-section/audio/. [Online; accessed 18-June-2012].

Foster, D. H. and Bishof, W. F. (1997). Bootstrap estimates of the statistical accuracy of thresholds obtained from psychometric functions. *Spatial Vision*, 11(1):135–139.

Gardner, W. (1998). *3-D audio using loudspeakers*, volume 444. Kluwer Academic Publishers.

Gerzon, M. (1973). Periphony: With-height sound reproduction. *J. Audio Eng. Soc*, 21(1):2–10.

Gerzon, M. (1975a). The design of precisely coincident microphone arrays for stereo and surround sound. In *Proc. 50th Conv. Audio Eng. Soc. London, UK*.

Gerzon, M. (1975b). Recording concert hall acoustics for posterity. *The Journal of the Audio Engineering Society*, 23.

Gerzon, M. (1992). General metatheory of auditory localisation. *preprint*, 3306(92):24–27.

Glasgal, R. (1995). Ambiophonics: The synthesis of concert-hall sound fields in the home. In *Proc. 99th Conv. Audio Eng. Soc. New York, New York, USA*.

Grassi, M. and Soranzo, A. (2009). Mlp: a matlab toolbox for rapid and reliable auditory threshold estimations. *Behavior Research Methods*, 41:20–28.

Griesinger, D. (1990). Binaural techniques for music reproduction. In *Proceedings of the AES 8th International Conference*, pages 197–207.

Hamasaki, K., Hiyama, K., and Okumura, R. (2005). The 22.2 multichannel sound system and its application. In *Proc. 118th Conv. Audio Eng. Soc. Barcelona, Spain*.

Hawley, M., Litovsky, R., and Culling, J. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115:833.

Hoffmann, H., Dachselt, R., and Meissner, K. (2003). An independent declarative 3d audio format on the basis of xml. In *Proc. International Conference on Auditory Display*.

immsound (2012). imm sound. http://www.immsound.com. [Online; accessed 18-June-2012].

iosono (2012). iosono. http://www.iosono-sound.com/. [Online; accessed 18-June-2012].

Jacobsen, F. (1997). An overview of the sources of error in sound power determination using the intensity technique. *Applied Acoustics*, 50(2):155–166.

Jacobsen, F. and De Bree, H. (2005). A comparison of two different sound intensity measurement principles. *The Journal of the Acoustical Society of America*, 118:1510.

Katz, B. (2000). Integrated approach to metering, monitoring, and leveling practices, part 1: Two-channel metering. *Journal of the Audio Engineering Society*, 48(9):800–809.

Kidd, G., Mason, C., Brughera, A., and Hartmann, W. (2005). The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta acustica united with acustica*, 91(3):526–536.

Kirkpatrick, S., Gelatt Jr, C., and Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kolundzija, M., Faller, C., and Vetterli, M. (2010). Sound field recording by measuring gradients. In *Proc 128th Conv. Audio Eng. Soc. London, UK*.

Laborie, A., Bruno, R., and Montoya, S. (2004). Designing high spatial resolution microphones. In *Proc. 117th Conv. Audio Eng. Soc. San Francisco, CA, USA*.

Larsen, J., Norris, C., and Cacioppo, J. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, 40(5):776–785.

Malham, D. (2003). Higher order ambisonics systems. http://www.york.ac.uk/inst/mustech/3d_audio/higher_order_ambisonics.pdf. [Online; accessed 18-June-2012].

Mann, J. and Tichy, J. (1991). Acoustic intensity analysis: Distinguishing energy propagation and wave-front propagation. *The Journal of the Acoustical Society of America*, 90(1):20–25.

Mann, J., Tichy, J., and Romano, A. (1987). Instantaneous and time-averaged energy transfer in acoustic fields. *The Journal of the Acoustical Society of America*, 82:17.

Nicole, R. (1999). *Sound Spatialization over an Extensive Area: Application to Telepresence and Videoconferencing*. PhD thesis, University of Le Mans.

Olson, H. (1931). Mass controlled electrodynamic microphones: the ribbon microphone. *The Journal of the Acoustical Society of America*, 3:9.

Olson, H. (1946). Gradient microphones. *The Journal of the Acoustical Society of America*, 17:192.

Olson, H. (1974). Field-type acoustic wattmeter. *The Journal of the Acoustical Society of America*, 55:S70.

Park, M. and Rafaely, B. (2005). Sound-field analysis by plane-wave decomposition using spherical microphone array. *The Journal of the Acoustical Society of America*, 118:3094.

Poletti, M. (1996). The design of encoding functions for stereophonic and polyphonic sound systems. *Journal of the audio Engineering Society*, 44(11):948–963.

Poletti, M. (2005a). Effect of noise and transducer variability on the performance of circular microphone arrays. *Journal of the Audio Engineering Society*, 53(5):371–384.

Poletti, M. (2005b). Three-dimensional surround sound systems based on spherical harmonics. *Journal of the Audio Engineering Society*, 53(11):1004–1025.

Potard, G. (2006). 3d-audio object oriented coding. *University of Wollongong Thesis Collection*, page 539.

Potard, G. and Burnett, I. (2004). Decorrelation techniques for the rendering of apparent sound source width in 3d audio displays. In *Proc. of the 7th Int. Conf. on Digital Audio Effects*, pages 5–8.

Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466.

Pulkki, V. (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503.

Ravaja, N. (2004). Contributions of psychophysiology to media research: Review and recommendations. *Media Psychology*, 6(2):193–235.

Rayleigh, B. (1896). *The theory of sound*, volume 2. Macmillan.

Sanchez-Vives, M. and Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, 6(4):332–339.

Schiffrer, G. and Stanzial, D. (1994). Energetic properties of acoustic fields. *The Journal of the Acoustical Society of America*, 96:3645.

Schultz, T. J. (1955). Acoustic wattmeter. U.S. Patent 2836656. Application October 31.

SPS422B (2004). Soundfield sps422b. `http://www.soundfield.com/products/sps422b.php`. [Online; accessed 18-June-2012].

ssr (2012). Soundscape renderer. `http://www.tu-berlin.de/?ssr`. [Online; accessed 18-June-2012].

Stanzial, D., Bonsi, D., and Prodi, N. (2000). Measurement of new energetic parameters for the objective characterization of an opera house. *Journal of sound and vibration*, 232(1):193–211.

Stanzial, D., Bonsi, D., and Schiffrer, G. (2003). Four-dimensional treatment of linear acoustic fields and radiation pressure. *Acta Acustica united with Acustica*, 89(2):213–224.

Stanzial, D. and Prodi, N. (1997). Measurements of newly defined intensimetric quantities and their physical interpretation. *The Journal of the Acoustical Society of America*, 102:2033.

Tetramic (2007). Core sound tetramic. `http://www.core-sound.com/TetraMic/1.php`. [Online; accessed 18-June-2012].

Theile, G. and Wittek, H. (2011). Principles in surround recordings with height. In *130th Convention of the AES*.

Tjis, E. H. G., Nejade, A., and De Bree, H. (2009). Verification of pu intensity calculation.

Tsang, P. and Cheung, K. (2009). Development of a re-configurable ambisonic decoder for irregular loudspeaker configuration. *Circuits, Devices & Systems, IET*, 3(4):197–203.

Tsang, P., Cheung, W., and Leung, C. (2009). Decoding ambisonic signals to irregular loudspeaker configuration based on artificial neural networks. In *Neural Information Processing*, pages 273–280. Springer.

Watson, A. (1979). Probability summation over time. *Vision research*, 19(5):515–522.

Weibull, W. (1951). A statistical distribution of wide applicability. *Journal of applied mechanics*.

Wiggins, B., Paterson-Stephens, I., Lowndes, V., and Berry, S. (2003). The design and optimisation of surround sound decoders using heuristic methods. *Proceedings of UKSim*, 3:106–114.

Williams, M. (2012). Microphone array design for localization with elevation cues. In *132th Convention of the AES*.

Yntema, D. (2008). *An integrated three-dimensional sound-intensity-probe*. PhD thesis, Ph. D. Thesis, University of Twente, The Netherlands.