The development of a 'Virtual Studio' for monitoring Ambisonic based multichannel loudspeaker arrays through headphones

Fabio Wanderley Janhan Sousa

Dissertation presented as main submission for obtaining the MA (by research) in Music and Technology degree.

The University of York - Music Technology

September 2011

Abstract

After presenting some technological background related to the main subject, contemporary Binaural and Ambisonic tools are discussed. In spite of all the possible applications of both systems, user friendly tools for compositional applications are the main foci of the research.

The first chapter presents the historical and theoretical background behind Ambisonics, binaural systems and head tracking technologies, as well as some recent developments. The main objective of this chapter is to present an overview of previous researches and define how the main subject of the present research has been treated so far. The chapter that follows describe in detail some of the currently available tools for working with three dimensional sounds, their principles and possibilities as described by their developers. The third chapter discuss some experiments realized during the past year and how the tools presented in the second chapter can be put together to build a 'Virtual Studio', the difficulties faced during this process, what kind of compromises were required and what assumptions were made in order to make it work intuitively.

Since MAX/MSP is probably the most commonly used software within the academic electronic music composers community, most of the work described was based in this environment in order to allow composers to work in a more intuitive way and to focus on the music itself. MAX/MSP tools are described and analysed as well as recent VST plugins for working in the most popular environment in the music production community outside the academy - digital audio workstations (DAWs). Beyond this, analyses of some of the author`s experiences are reported and some reflections are discussed. Through these discussions, focused on the usage of the previously mentioned tools in practical applications, such as music productions and two case studies related to composition processes, a conclusion points to some future work on binaural reproduction of multichannel systems over headphones.

List of contents

List of tables

List of illustrations

List of accompanying material

DVD containing:

- AES article and MAX patch used in the listening tests
- MAX patch and files used in the compositions
- B-format samples of performed recordings
- Virtual Studio MAX patch, files and orientations
- Installers for needed softwares for Virtual Studio
- PDF version of the dissertation

Acknowledgments

First of all, I would like to thank my parents who have supported me in everything I have done until now, especially this past year, during my quest for an international master's degree. I would also like to thank my little sister, who I have missed a lot living far away from home, and my sweet fiancée for always being there for me, even if it was through skype.

I would like to extend a special thank you to my supervisor, Dave Malham, and the other lectures in the music and technology course, who shed some light on the many thoughts I had before coming here. Many thanks too to Margarida Borghoff, whose friendship went beyond the Viva Musica concerts and the piano classes and my supervisor Sergio Freire, a constant inspiration for my studies and intellectual development.

I also need to thank the friends I made in York, mainly Avnish, Laura and Galyna, and all the good moments we had together inside and outside our college. I would like to thank Catherine Duncan a lot for offering me the assistantship, which kept me busy and channelled my workaholic impulses, as well as for her friendship. I would also like to thank Ben Eyes for his great friendship, mixed with busy working and camping activities, which made my choice of working with music and technology especially beneficial.

Last but not least, I would like to thank all the music department students, who made the recording sessions possible, and the music and technology masters' students for their company during this journey, as well as for their opinions and time spent in participating in my crazy listening tests.

Thank you.

Author's declaration

The work contained in this dissertation is that of the author. Some of preliminary work produced during the course of working on the dissertation was published in the following paper:

Sousa, F. W. J. 'Subjective comparison between stereo and binaural processing from B-format Ambisonic raw audio material'. Convention paper presented at the 130th AES convention, London (UK), 13-16 May 2011.

I think that all music will became space music and that space becomes as important as pitch in traditional music, as duration and rhythm and meter and there is a very new development of harmony of space and I mean space chords, space melodies and that doesn`t mean pitches, it means movement on several levels around the listener: above, below, in all directions.

(Stockhausen, 1997; also quoted in Worrall, 1998: 93).

Introduction

Art workers have their inspiration in real world and all the possible sensations it can offer. A lot of research has been done last 20 year aiming making three dimensional spatial attributes of sound more reproducible to be included in artistic manifestations. Applications in domestic environment such as in high definition television (HDTV), games and films, as well as in spatial music compositions reproduction, virtual reality, auralization, etc. are just a few of the fields where distribution of three-dimensional audio can be applied and have been developed for more realistic experiences.

Since there is much good theoretical and experimental work involving human hearing perception (Blake and Sekuler, 2006; Blauert, 1997; Kendall, 1995; Malham, 1998a; Plack, 2005; Stern, Brown and Wang, 2006), as well as a lot of work describing the historical development of audio reproduction systems from mono to 3D audio (Malham and Myatt, 1995; Rumsey, 2001), the current work starts from a brief theoretical review of new developments of contemporary systems.

In order to understand the principles behind the concept of the ‘virtual home theatre’ (Rumsey, 2001: 75) and what the present author calls ‘virtual studio’, the principles as well as the most recent developments in Ambisonics, binaural and head tracking systems are the main concern of the chapter that follows.

Chapter 1 – Theoretical background

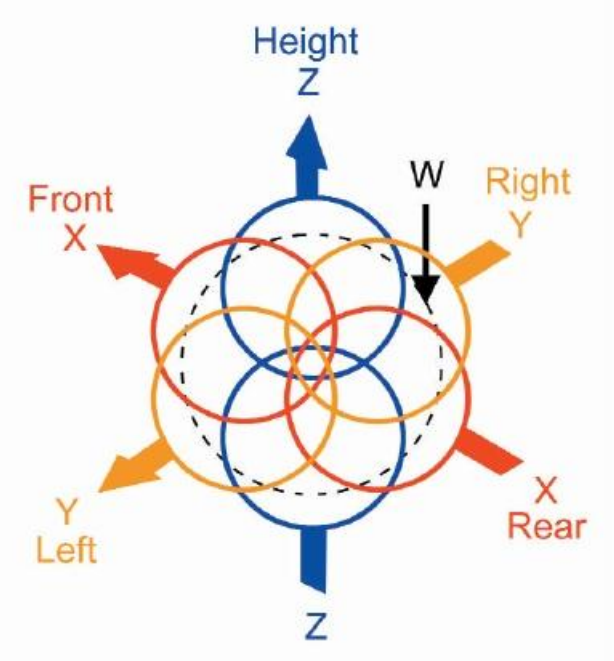
1.1 Ambisonics

1.1.1 Historical background and overview

The Ambisonic system was developed from the principles of Alan Blumlein's work in Britain in the late 1920s whilst experimenting on stereo systems using coincident pairs of directional microphones (Malham, 1990: 118). This further development by, amongst others, Peter Fellgett, Peter Craven and Michael Gerzon, was documented by Michael Gerzon in 1970s and happened at a time when multichannel audio started to become more available through the development of the so-called quadraphonic systems and the possibility of recording 4 channels in tapes.

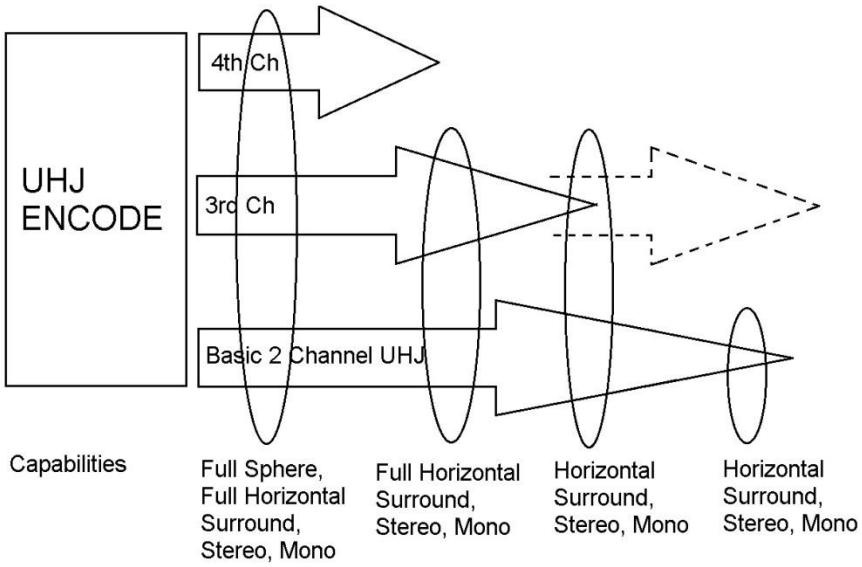
The basic principles are that the whole three dimensional soundfield observed at a single point in space can be captured by a suitable four-capsule microphone (named Soundfield microphone) that can capture all the directional information. In a later process this signal is decoded into multichannel loudspeaker signals that can vary depending on the number of loudspeakers and their positions.

> The important thing to note is that there is no need to consider the actual details of the reproduction system during the original recording or synthesis, provided the B-format specifications are followed and suitable loudspeaker / decoder setups are used. In all other respects the two parts of the system, encoding and decoding, are completely separate, giving considerable creative freedom to the composer, who no longer has to consider the performance space during composition.
>
> (Malham, 1990: 119)



Directional polar patterns for full-sphere B-format signals W, X, Y and Z. Extracted from <http://www.soundfield.com/downloads/b_format.pdf>.

Malham also highlights that one of the most interesting features of the Ambisonic signals is that they can be processed as a single entity, performing movements like rotating, tumbling or tilting, as well as adding some reverberation to give better distance cues. Compatibility with stereo and mono reproduction is also a very important feature since when the system started being developed, as well as nowadays, there is the need on doing so, mainly by broadcasters, due to the fact that not every user can afford the multichannel reproduction decoding system or even the stereo one. A method of encoding the Ambisonic signals into a stereo one, also containing all the three-dimensional information, known as UHJ format was also developed (Malham, 1990: 120).



UHJ hierarchy intercompatibility. Derived from Gerzon, 1985.

$$\text{Left} = (0.0928 + 0.255j)X + (0.4699 - 0.171j)W + (0.3225 - 0.00855j)Y$$

$$\text{Right} = (0.0928 - 0.255j)X + (0.4699 + 0.171j)W - (0.3225 + 0.00855j)Y$$

Formulas to obtain the UHJ stereo signal from B-format. Where W, X and Y are the signals from the horizontal B-format and j is phase shifting by 90 degrees. Derived from Malham, 1990.

In the first paper describing Ambisonic systems (published one year after its presentation at the 2nd AES convention in Munich, Germany), Gerzon defines microphone techniques for 19 periphonic loudspeaker arrays and establishes procedures for designing other systems. According to the author 'the reproduction of sound with height can in principle be achieved via any arrangement of loudspeakers that forms a solid, enclosing the listener' (Gerzon, 1973: 2) and

> although irregular speaker layouts are likely to be used for domestic with height sound reproduction when it arrives, it is c onvenient here to consider only fairly regular arrangements of speakers, e.g., at the vertices, face centers or edge centers of an Archimedean solid.
>
> (Gerzon, 1973: 2)

In the same paper, Gerzon argued that the obtainable directional resolution is proportional to the number of channels while decoding a sound and conclude that

> those who have had the opportunity of hearing periphony at its best can have no doubt that the

> height effect is important in the perception of sound and the enjoyment of music, and it thus seems worthwhile to ensure that current recording media have the possibility of adding the height effect at some future time.
>
> (Gerzon, 1973: 10)

Seven years later in a paper presented at the 65th AES convention, Gerzon shows himself aware of the impracticalities of Ambisonic systems such as the 'use of 12 speakers at the face-centres or of 20 speakers at the vertices of a regular dodecahedron' (Gerzon, 1980: 1) and report that optimum arrays would be the cuboid, the octahedron and the birectangle, since those follows the basic rules that are also the three main point of a 'diametric decoder theorem':

> 1-) all speakers are same distance from the centre of layout
>
> 2-) speakers are placed in diametrically opposite pairs
>
> 3-) the sum of the 2 signals fed to each diametric pair is the same for all diametric pairs
>
> (Gerzon, 1980: 4)

Looking for practical application of the Ambisonic system, Barton and Gerzon, in 1984, presented a detailed description of Ambisonic formats as well as of the equipment related to an implementation of an Ambisonic studio. In their concept, mixing in B-format has the main advantage that one can re-release the final mix in other formats (3 or 4 channels UHJ for example, instead of just the 2 channels UHJ). This idea is very interesting in practical application since a release made in stereo can then be released also in 5.1 surround or the other way round (very common practices nowadays, known as upmixing and downmixing) and any other multichannel format like 7.1, 10.2, 22.2, etc. without the need of redoing the whole mix again.

In their idealised Ambisonic studio, idealised in the sense that multi-track digital recorders were not yet sufficiently developed,

> if a Soundfield microphone is available, then the B-format output of its soundfield control unit can be used to feed one of the B-format inputs of the pan-rotate unit, and the converter can be used to feed the other B-format input of the pan rotate unit. This allows multitrack B-format mixed mono to be mixed with a Soundfield microphone signal. If the unmodified B-format of the soundfteld

> microphone is stored on 4 channels of the multitrack tape, the soundfield control unit can be used to modify the B-format sound, which can then be mixed with the B-format panned mono signals.
>
> (Barton and Gerzon, 1984: 9)

The possibility of B-format mixes offering 2-channel signals as well as full periphonic audio makes this format 'capable of being released both in current and future consumer surround-sound formats' (Barton and Gerzon, 1984: 9), and the authors concludes:

> When all is said and done, the engineering that allows studios to be converted to Ambisonic production is only half the story. The other half, what can be done creatively with the equipment, is a story that is only just starting to unfold.
>
> (Barton and Gerzon, 1984: 10)

One year later, Gerzon presents the Ambisonic system as a possibility for multichannel broadcasting video due to the fact that it 'provides for full upward compatibility to any number of loudspeakers in any reasonable configuration' (Gerzon, 1985: 859) and highlights the advantages of the system such the 'optimum compatibility with mono and stereo playback equipment' and its superiority to traditional stereo (Gerzon, 1985: 863). The possibility of the broadcaster not being responsible for the consumer loudspeakers position is also highlighted by the author since the B-format signal could be decoded to different arrays.

The possibility of rotating and zooming were also presented as a visionary interactivity between the listener and the audio material presented on TV, when one could be 'emphasizing the frontal stage and deemphasizing ambience and rear-stage sounds, whereas others might prefer listening with a more distant perspective' (Gerzon, 1985: 866).

> Ambisonics and UHJ should be thought of as the first systematic approach to handling and conveying to the listener a total sound field, rather than some arbitrarily chosen loudspeaker feeds. As such it allows both the broadcaster and the listener to make their own choices (in terms of convenience and cost) of how good an approximation to the original sound field is to be obtained, without creating unnecessary restrictions on either current or future possibilities.
>
> (Gerzon, 1985: 867)

In 1992, while presenting new decoders solutions for HDTV broadcasts, Gerzon

highlights another great advantage of Ambisonic systems that has becoming even more important for contemporary sound systems. The fact that in ‘making the maximum possible of auditory localisation cues’, the system provides robustness in cases of misadjustment of loudspeakers positioning or objects in front of the speakers, and ‘very low listening fatigue’, giving three-dimensional high quality reproduction in any circumstances (Gerzon, 1992: 1).

Continuing Gerzon’s work, Malham has written articles describing the advantages of Ambisonics systems, that ‘unfortunately due to financial and other factors unconnected with the technical merits of the system it has never caught on in the domestic market place’ (Malham, 1987: 1), as well as tools for Ambisonic manipulation. The author classifies the Ambisonic controls into two different classes: those that manipulate a single sound source within the soundfield and those that manipulate the whole soundfield. He also highlights the scarcity of tools to work with Ambisonic signals, particularly in the commercial domain.

According to the author, although Ambisonic systems were intended originally to be played back in small listening areas such as domestic environments, performances in large areas also proved to be well received. He argues that the sound image does not remain the same for all the listening positions but the distortions that occur are close to those that occur with real sound sources. Other important observation is that the image from outside the loudspeaker array can still be perceived. (Malham, 1987: 2)

The already quoted Malham’s work presents formulas to develop pan-pots, rotators and zoomers. At that time, the technical limitations of automated controls, lead to compromised solutions presenting granulation and lack of resolution. Nowadays these problems can be solved with full digital processing.

Five years later this approach of Ambisonics Systems in large areas was described as a very welcome solution for contemporary compositions reproductions, since the recording process and the playback can be processed independently. Experiences realised at the

University of York show that the phenomena of moving away from the centre of the loudspeaker array distorted the sound field reproduction in ways that were similar to natural situations and the fact that the system work also for listeners outside the array was considered surprising. Since all the other surround systems also present some problem, the simplicity of Ambisonics was perceived as a big advantage. (Malham, 1992).

In a retrospective report of spatialized sounds, Malham in 1996 argues that the stereo system developed by Alan Blumlein was very well accepted for domestic applications but people involved with electroacoustic music had been conducting intense research into the spatial characteristics of sound images. This research seems to have started around 1951, when the 'potentiomèter d'espace', developed by Jacques Poullin in Paris, 'was capable of feeding sounds to multiple speakers, including some above the audience, under the control of a performer using a form of hand held inductive loop transmitter' (Malham, 1996: 96). The way a sound channel was spread through the loudspeakers was controlled by changing the position of the transmitter coil. This same principle is used nowadays in head tracking systems for virtual reality.

According to the author, one of the most extravagant experiments related to spatialised sounds is Varèse's 'Poème Eléctronique', played in the Phillip's Pavillion at the Brussels World Fair in 1958 using 400 loudspeakers and 15 tape recorders, impractical for home systems reproduction. He also reports that in the 1960's serious work started in surround sound. However, the most common approach, the quadraphonic system, which was intended to be an extension of the stereo, ignored 'the very mechanism which enables two speaker stereo to work' (Malham, 1996: 97) while putting 90º of angle between the four loudspeakers, and was abandoned for consumer use in the end.

The author claims that the only two multichannel systems that survived into the digital era and that worked reasonably well with the compromise of matrixing the four channels into two to maintain stereo compatibility were the Dolby Stereo (or Dolby Surround) and the UHJ Ambisonics. The UHJ Ambisonics particularly survived as it providing good

stereo and mono compatibility. However, these compromises are likely to disappear with the new developments on multichannel media.

The full capabilities of Ambisonics systems like rotating, zooming, tilting and tumbling that are easily implemented in a four channel B-format file and were not possible with the two channel version (UHJ) could now be included in the remote control, when Ambisonic systems become available to home users, since multichannel consumer systems have already reached the market. Transmitting four B-format Ambisonic signal in the prominent DVD audio (5.1 + 2.0) is proposed by the author and only has to be added to existing systems.

1.1.2 Theory

Almost every paper or other work related to Ambisonic reproduction systems present some basic maths related to encoding and decoding (Malham, 1990: 119; Malham and Orton, 1991: 467; Malham, 1993: 62-64; Pulkki, 1997: 458; Avdelidis et al., 2009: 3-4, etc.). All of them are based on Gerzon's work developed in the 1970s (Gerzon 1973; also quoted in Malham, 1999a: 485) which expands the idea of using spherical harmonics to represent directional content of recorded sounds in such a way that it follows a hierarchy, presented by Cooper and Shiga (Cooper and Shiga, 1972; also quoted in Gerzon, 1973: 2), to a system with height information. In this work Gerzon also presented solutions for up to 3rd order Ambisonics decoding.

$X = \cos P.\cos Q$	(front-back)
$Y = \sin P.\cos Q$	(left-right)
$Z = \sin Q$	(up-down)
$W = 0.707$	(pressure signal)

Encoding formulas, where *P* is the anticlockwise angle from center front and *Q* is the elevation. Derived from Malham, 1990.

LFU = W + 0.707(X + Y + Z),
RFU = W + 0.707(X - Y + Z),
LBU = W + 0.707(-X + Y + Z),
RBU = W + 0.707(-X - Y + Z),
LFD = W + 0.707(X + Y - Z),
RFD = W + 0.707(X - Y - Z),
LBD = W + 0.707(-X + Y - Z),
RBD = W + 0.707(-X - Y - Z),

Decoding formulas for a cubic array, where L is left, R is right, U is up and D is down, corresponding to loudspeakers positions. Derived from Malham and Myatt, 1995.

Later, in 1977, Gerzon and Craven patented the mathematical formulas and physical layout of what would soon be known as the Soundfield microphone, able to capture the so called first order A-format Ambisonic signals, which were obtained from four cardioid capsules positioned close to each other. Formulas were presented to convert these signals into Ambisonic B-format that are more suitable for various manipulations. These signals consist of an omni-directional (W) and three bidirectional first order virtual microphones (X, Y and Z), the first oriented to front-back direction, the second to left-right and the third to up-down (Craven and Gerzon, 1977).

In 1992, Gerzon presented what he called a ‘General metatheory of auditory localization’, where he describes the mathematical theory behind the development of Ambisonic systems, based on directional sound localization and spherical harmonics. Some consideration is given to head movement influence (Gerzon, 1992: 12), pinnae colouration that has some influence in reproduction above 4 or 5 KHz (Gerzon, 1992: 43) and the dependence of the perceived spatialization due to the program material (Gerzon, 1992: 50).

For the system to work in perfect conditions according the models described, Gerzon present two main assumptions: the first one is that a monophonic sound is reproduced at a time through the loudspeakers, which does not normally happen in practice since different sounds are presented with different gains in different times during a music program reproduction, and the second is that the listener is exactly in the centre of the loudspeaker

array (Gerzon, 1992: 25). To improve the performance of the system, mainly in the region above 1KHz, the author consider adopting energy models, which will be described later in developments based on VBAP implementations (Gerzon, 1992: 42)

In 1998, Daniel extended Ambisonics principles and described a mathematical generalization for the encoding and decoding process that would be the basis of many of the following works and tools developments (Daniel, Rault and Polack 1998; also quoted in Monro, 2000) and fully explained in his Ph.D. thesis two years later (Daniel, 2000; also quoted in Frank, Sontacchi and Zotter, 2008), this being considered as 'one of the most comprehensive works about Ambisonics' (Frank, Sontacchi and Zotter, 2008: 1). A simultaneous development of the encoding and decoding processes is also presented by Monro that argues it is extending Malham's work started in 1992 on in-phase correction for loudspeakers symmetrically opposed (Monro, 2000: 1).

A review of the three-dimensional theory of sound recording and reproduction based on spherical harmonics can also be found in Poletti, 2005. His work focuses on horizontal only reproduction systems using 100 loudspeakers or more and all his tests were plotted in computer environment, not physical systems. It is interesting to observe how these developments are approximating High Order Ambisonics systems to Wave Field Synthesis systems and, since both are based on the same principles of Huygens (Malham, 2001b: 36) can present some generalizations as well as some practical solutions as reported by Poletti:

> The reproduction system requires a number of loudspeakers, which rises quadratically with the reproduction frequency. The cost may be reduced by using a small number of woofers and a large number of tweeters, and the tweeters could be flat panel transducers, which may be able to be incorporated into wall coverings.
>
> (Poletti, 2005: 1022)

Developments in VBAP (described later) allowed some researchers to implement simplified ways of decoding Ambisonics and developing tools such as the ICST tools that are going to be described in future sections of the present work. Neukom's developments for instance, describe panning functions that are equivalent to Ambisonic en- and decoding,

making the processes only dependent on the order, that also define the accuracy of the system, and with less cpu cost (Neukom, 2007: 5; Neukom and Schacher, 2008).

1.1.3 The development and use of the Soundfield microphone

The design of the tetrahedron microphone was first reported by Gerzon in 1975, but two years before, he had already highlighted that the nine channels 2nd order B-format signals, obtained with 12 cardioid capsules in a dodecahedron configuration with less than 5cm distant each other, would be quiet difficult to get, even more difficult would be a sixteen channels 3rd order B-format (Gerzon, 1973: 4). These considerations lead him to focus on the 1st order B-format microphone development that uses only four capsules.

4 capsules microphone arrangement. Extracted from <http://www.soundfield.com/downloads/b_format.pdf>.

In his first work about designing the tetrahedron microphone, he describe how to compensate the distance between the capsules to avoid the loss of accuracy in representing sound localizations that happen beyond 5KHz and argues that this permits the achievement of effective coincidence up to 7KHz which provides stability and lack of ambiguity

superior to any stereo reproduction even for non-centred listeners (Gerzon, 1975a: 4).

Intended to be used for stereo and surround recording, the tetrahedron microphone and the compensation described

> permits one to record the information from the microphone onto 4-channel tape, and to select any desired stereo coincident microphone technique (including adjustable angle of vertical tilt) at any later time. For the first time, this allows a full mixdown capability off coincident microphones. This, of course, is only practical because the microphones are, in effect, precisely coincident.
>
> (Gerzon, 1975a: 4)

Despite the description and design of the Soundfield microphone predates the patent, the patent itself, for the development of the tetrahedron microphone as well as the formulas for decoding A-format to B-format, was accomplished in 1977 (Craven and Gerzon, 1977).

The so called native Ambisonic B-format consists on obtaining the W, X, Y and Z signal without the encoding of the A-format from a tetrahedron microphone, by using ordinary microphones, one with omni-directional polar pattern and three bi-directional, oriented horizontally (front-back and left-right) and vertically (up-down). The problem of this approach is to mathematically make them coincident, task made easier with the previously mentioned tetrahedron microphone design that presents the four capsules mounted as close as possible, intended to be at the same point (Malham, 1990: 118).

After two prototypes (Mark I and Mark II) produced by Calrec, the first manufactured tetrahedron microphone, the Mark III, from 1979, was used by BBC in a few recordings (http://www.ambisonic.net/sfexp.html) but rapidly substituted by the Mark IV, that is described as ‘the nucleus of Ambisonics and surround sound technology’ (Bridge and Jagger, 1984:4). In their paper, the quoted authors also describe some of the controls available in the developed control unit, such as azimuth, elevation, dominance, stereophonic polar pattern and angle (for stereophonic compatibility), highlighted as the main advantages of B-format. Besides presenting the A-format to B-format formulas and circuits developed for the manipulations listed above, the authors refer to the Soundfield microphone as ‘essentially designed to capture accurately all the sounds that exists at a

point in space' (Bridge and Jagger, 1984: 4) they add that

> B-format signals are capable of accommodating a full 360 of directional sound in three channels, with W, X and Y or, by including Z, the vertical component, a full sphere of directions. By storing signals in B-format on four tracks, optimum recordings may be issued not only in mono and stereo but in surround sound and later, even periphonically (with height).
>
> (Bridge and Jagger, 1984: 4)

Some other interesting features of the tetrahedron microphone use are highlighted by Malham, such as the convenience of using it for surround recording instead of a conventional five microphone array and points out that in the same way some agree that the crossed stereo pair is the best approach to obtain stereo images, the extension of this technique for surround sound would be the use the tetrahedron microphone. Since 'appropriate blending of these signals will produce any number of different virtual microphones' (Malham, 1998b: 54), the author suggests that with B-format signals one could obtain not only the stereo signal for left and right loudspeakers but also a centre and two surround signals to be used into the cinema standard. These signals could be obtained with complete control of directivity which allows one to choose the amount of reverberation and diffuseness.

According to the author, the use of such a microphone for surround recordings also deliver 'a more coherent image than current multi-microphone systems', has the advantage of being 'smaller and more unobtrusive microphone array' (Malham, 1998b: 56) and the advantage of the four recorded channels being reproduced in any loudspeaker array, 5.1, 7.1, 10.2, 22.2, etc. For better directivity the use of higher orders microphones could be adopted but their developments were not yet a reality in 1998 and mention as the major challenge in his paper from 1999

> The major challenge is, however, the use of mixed second and first order source materials. This comes about because the only commercially available microphone suitable for directly capturing sound Ambisonically is the Soundfield microphone which is limited to first order components.
>
> (Malham, 1999a: 486)

Recently available, a 32 capsule microphone, 5th order Ambisonic microphone, named Eigenmike and released by mh acoustics can create more directional polar patterns and higher orders Ambisonics (HOA).

The Eigenmike microphone. Extracted from

<http://www.mhacoustics.com/mh_acoustics/Eigenmike_microphone_array.html>.

1.1.4 Limitation and new developments

A problem some authors report in their work is that the ability of the Ambisonic system to form images throughout 360° begins to fail at high frequencies. Bramford and Vanderkooy, 1995 while comparing Ambisonics to stereo and Dolby Surround systems of representing sound images, concluded that, although Ambisonic presents better imaging and increased effective listening area compared to stereo and Dolby Surround, these benefits start decreasing with increasing frequency.

This limitation of the Ambisonic system is the main topic of much research and while some researchers have tried to develop High Order Ambisonics microphones, others try to develop different techniques to overcome this limitation. Bruno et al. for instance, present an array composed of 24 omni directional microphones inside a ball of radius less than 20cm as a solution for 'the lack of good high order capsules and the mandatory position at the center' (Bruno et al., 2003: 17).

Interesting developments can also be observed in convolution reverbs since once made in Ambisonic, it can be applied to any loudspeaker system, such as 5.1, 7.1, 10.2, 22.2, etc. This property of Ambisonic manipulation is very useful in virtual reality and auralization applications since 'if all of the signals are mixed and panned ambisonically, then only four channels of convolution computation are required (W, X, Y and Z)' (Dalenbäck et al., 1995: 5), the author does not actually refer to making Ambisonic impulse response recordings but they can be calculated and encoded ambisonically, as in combination with the direct sound source and echoes, so the computational requirements are dependent on the number of sources but not on the number of loudspeakers used for reproduction.

Ambisonic systems can also be considered a future proof format as claimed by Cába, 2002. The author describes practical experiences in which recordings were made without knowing the final format and Ambisoic mixes were used for decoding to both stereo and multichannel downmix that included height to be played back into any scenario (e. g. loudspeaker array). He argues that Ambisonics allows sound engineers to work in the same way they have been working for many years, using the same traditional techniques but with the additional advantage that the 3D mix can be manipulated as a whole. For future works he points out the need of developing B-format processing tools, such as equalizers, meters, room effects and reverb simulators, as well as a software for calibration of homemade microphones to allow the engineer to set up a B-format native signals microphone and align them in time and level.

Other work that needs to be taken into account when trying to make Ambisonics reproduction more precise is Avdelidis et. al 2009, that, in spite of being performed in a virtual environment (e. g. simulated inside the computer), use two Soundfield microphones to improve sound localization.

Higher Order Ambisonics (HOA) have also been the topic of recent research, starting from Malham 1999a, who presented an hybrid Ambisonic 1.5 when 1st order signals are added to 2nd order horizontally only, decision that fits with our better resolution on the

horizontal plane and requires only two additional channel (U and V). For a full three-dimensional representation, a 2nd order would need also R, S and T channels. He also mentions hybrid systems:

> Recently, combined Ambisonic and Holophonic or Ambisonic and Wavefield synthesis approaches have been put forward as offering the best features of both, although still without the ease of implementation and control that Ambisonic can offer.
>
> (Malham, 1999a: 485)

The fact that HOA is not as widely used as first order Ambisonics is due to the limitation of recording systems. This fact is mention by Daniel, 2003, while defining the principles of Near Field Compensated High Order Ambisonic (NFC-HOA), necessary to make the Ambisonic system practicable. His description of near field compensation is an extension of the bass boost that Gerzon recommends to compensate in early Ambisonic systems, aiming to preserve the 'curvature of the encoded wave fronts, it is now suggested to introduce the loudspeaker near field modelling into the re-encoding equation' (Daniel, 2003: 5). He also presents an encoding compensation to reproduce better distance effect, mainly for sound sources inside the loudspeaker array, based on parametric IIR digital filters.

Other experiments involving HOA and the sweet spot of an Ambisonic horizontal only system is reported by Frank et al. 2008, which confirm that 'the localization improves at higher orders' (Frank et al., 2008: 9).

Research has also been undertaken on a hybrid system made from VBAP and Ambisonics based on the premise that both can be used for arbitrary loudspeaker setups 'where the same distance of the loudspeakers from the listening position is assumed' (Batke and Keiler, 2010: 2). Vector Based Amplitude Panning (VBAP) was first described by Pulkki in 1997 and is based on the same principles the traditional stereo panning is, but extended to three-dimensional reproduction in any loudspeaker array placed arbitrarily in space. If the system is horizontal only, various 2D VBAP are put into practice, where only

one pair is working each time. For three dimensional systems, with height, 'three gain factors defines the virtual source direction perceived by the listener' and 'the virtual source can thus be placed on a surface of the three-dimensional sphere, the radius of which is defined by the distance between the listener and the loudspeakers' (Pulkki, 1997: 459). While comparing the Ambisonic system to the VBAP the author states that:

> This method [Ambisonic] differs from the standard amplitude panning method in that the gain factors gx and gy may have negative values. The negative values imply that the signals are stored on the recorder in antiphase when compared with the monophonic mix in the W channel.
>
> (Pulkki, 1997: 458)

The author also highlights three main characteristics of VBAP systems:

1) If the virtual source is located in the same direction as any of the loudspeakers, the signal emanates only from that particular loudspeaker, which provides maximum sharpness of the virtual source.
2) If the virtual source is located in a line connecting two loudspeakers, the sound is applied only to that pair, following the tangent law. The gain factor of the third loudspeaker is zero.
3) If the virtual source is located at the center of the active triangle, the gain factor of the loudspeakers are equal.

> These properties imply that VBAP produces virtual sound sources that are as sharp as it is possible with present loudspeaker configurations.
>
> (Pulkki, 1997: 461)

As any other system, VBPA also has limitations and the most easily recognizable of them is that the virtual sound source, as well as in the traditional stereo system, cannot be placed outside the region the loudspeaker cover. Since Ambisonic systems are very good on reproducing whole soundfields, B-format signals can be used to provide 3D reference image and VBAP to place signals, originated from close mics for example, into this soundfield, similarly to common practices in stereo and 5.1 music production applications.

Since most of the music production and sound applications are based on 'in box' manipulations, using mainly personal computers, and a multichannel loudspeaker setup (5.1) has already been widespread, recent developments have also been looking for

practical implementations appropriating these recent events. Ways of improving Ambisonic decoders for reproduction over 5.1 loudspeaker arrays have been the focus of some of those researches (Polettti, 2007; Moore and Wakefield, 2010). Experiments comparing hardware and software B-format decoding systems were made by Clark and Horsburgh and points to the software supremacy upon hardware decoders. The need of blind tests was reported and the fact that software version of such decoders can put Ambisonic format in the market through VST plugins was highlighted.

> In the digital age where DAWs and plug-ins are commonplace in all aspects of music creation, Ambisonics may finally have a place in the commercial world through high quality decoders, easy to use software and inexpensive multichannel interfaces.
>
> (Clark and Horsburgh, 2010: 7)

Further research also need to be done considering sound sources as complex entities instead of point sources, since each one has a different pattern of radiation that is frequency dependent and varies with distance. The development of an 'O' format Ambisonics (Malham, 2001a) and spherical loudspeaker arrays is just a start point.

1.1.5 Applications

Due to the fact that the B-format signal can be decoded to different loudspeaker arrays some experiments have been performed aiming at the evaluation of the loudspeaker array itself. Subjective tests can be focused on the naturalness – 'how close to real-life experience the sound reproduction felt to be' (Fredriksson and Zacharov, 2002: 2) - provided by multichannel loudspeaker systems, as this is considered to play an important role in the spatial impression of the listener, and that can include systems based on 'virtual home theatre' that 'provide different spatial impression compared to discrete multichannel reproductions' (Fredriksson and Zacharov, 2002: 1). This kind of comparison, according to the author, is not very interesting for film applications but certainly will be very useful for

audio purposes. Some of the conclusions of this paper points to the 5.1+H system, an hybrid of the traditional 5.1 surround system with the addition of a loudspeaker on the ceiling to reproduce height, as the most natural between those that were tested and the 5.1 transaural reproduction as the most unnatural. He also reported no influence of the sound material for grading the systems and that 'multichannel systems are considered better than two-channel systems, when aiming for a natural sound reproduction. However, increasing the number of speakers from five to eight does not necessarily improve the naturalness of sound' (Fredriksson and Zacharov, 2002: 8), phenomena caused by the poor channel separation of the Ambisonic system, which imply difficulty in sound localization when the number of loudspeakers used for reproduction increases.

Apparently, contemporary composers are those who get most benefits for using Ambisonic systems and since some of them consider diffusion of stereo audio files of electroacoustic pieces over various loudspeakers a kind of performance, different from the concept of a fixed diffusion made for a specific loudspeaker array, where the composer adjust only the gains, a system where individual tracks are documented in how they would be distributed over the loudspeakers, and use Ambisonic B-format as an interchangeable format for spatial music reproduction, seems very advantageous (Lyon, 2008).

Commercial applications using Ambisonic technologies have also being resumed. BBC staff for instance, have been reporting experiments on recording and mixing in Ambisonic domain before decoding for 5.1 or stereo and they argue that 'the availability of fast and low-cost digital processing, combined with freely-available digital production tools, means that it is now much easier for broadcasters to produce Ambisonic content' (Baume and Churnside, 2010: 1). The solution Ambisonics can provide by dealing with different audio formats such as mono, stereo and 5.1, being an easy and cheap way of archiving future-proof with height information for three-dimensional reproduction, as well as the availability of free softwares that deal with it, make its adoption an improvement that is not new, but has been highlighted (Baume and Churnside, 2010; Musil et al., 2008;

Nettingsmeier, 2010).

In the Baume and Churnside quoted paper, a music production mixed in Ambisonics is detailed. While mixing a live concert, the Ambisonic mix was made in the same way engineers are already used to make the surround 5.1, where a room signal (from the Decca tree and a Hamasaki square) is the basis of the mix and close mics are added to reinforce the image created. In the Ambisonic mix the room signal is provided by the Soundfield microphone and the close mics are panned through VST encoders. Other example of production made by the above mentioned authors, a radio drama, involves the creation of environments instead of their recreation and a post-production stage, since the performances were recorded in a dead space instead of a real reverberant one. Later on, reverberation was added to some tracks through impulse response convolution using the same Soundfield microphone (Baume and Churnside, 2010: 4).

However, some observations on the limited production tools and the lack of suitable tools 'suited to creative workflow' (Baume and Churnside, 2010: 4) are made and referred as the 'most noticeable barrier to the adoption of Ambisonics in production' (Baume and Churnside, 2010: 10).

In Cába 2002, a similar experience is reported and the author describe that in 5.1 decoding, the subwoofer can be obtained by filtering, with a low pass filter, the W channel and having an individual control of it. Although this approach is not completely suitable for film productions, the material obtained from this filtering can be added to the extra FX channel, and can also work well in music applications.

In contemporary spatialised music applications, the use of the Ambisonic approach have also the advantages of the flexibility while supporting different compositions by creating virtual loudspeaker signals, allowing the playback of compositions made specifically for Ambisonics possible as well as compositions designed for other loudspeakers arrays, such as 5.1, quadraphonic, etc. without the need of changing loudspeaker positions or cables or any other equipment during the performance (Nettingsmeier, 2010). Nettingsmeier also

lists what is needed to build the system and start working with Ambisonics - a playback engine, a virtual mixer and an Ambisonic decoder. In his paper he mention the possibility of controlling Ambisonic parameters, such as azimuth, elevation, rotation, etc. through MIDI controllers by the composer during live performances, what can improve the diffuse of pre-recorded pieces.

Although the Ambisonic system can be decoded to any loudspeaker array, Nettingsmeier points out some practical recommendations for the number of loudspeakers when reports that six horizontal speakers would give a listening area of one third of the circle and using third order operations it would be extend the an area of one half to two thirds the radius, improving the sharpness (getting closer to discrete panning) and using a minimum of eight loudspeakers in a horizontal array.

1.2 Binaural systems

'Hearing is the only one of our five senses that is truly capable of providing us with fully three-dimensional information about remote (i.e., non-contact) events' (Malham, 2001b: 32). This sort of observation probably led to developments in the early days of recordings that attempted to reproduce the whole three-dimensional sound around us only through two channels, what initially sounds obvious since we have only two ears. 'The perception that binaural is the "correct" system was, and still is, supported by the concept that if we achieve the exact duplication of what the ear would hear in natural situation, we will produce the best reproduction' (Malham, 2001b: 35).

1.2.1 Concepts and historical background

Although the concept of a binaural system seems very common nowadays, this concept has only really been consolidated since the start of the 2000's. Even in the early 1990's a

degree of confusion was still being shown by many authors.

In 1989, Eickmeier borrowed the concepts of binaural and stereophonic systems from Sunier.

> Binaural is a closed-circuit type of sound reproducing system in which two microphones, used to pick up the original sound, are each connected to two independent corresponding transducing channels which in turn are coupled to two independent corresponding telephone receivers worn by the listener. A stereophonic sound reproducing system is a field type sound reproducing system in which two or more microphones, used to pick up the original sound, are each coupled to a corresponding number of independent transducing channels which in turn are each coupled to a corresponding number of loudspeakers arranged in substantial geometrical correspondence to that of the microphones.
>
> (Sunier, 1960: 17-18; also quoted in Eickmeier, 1989: 2)

The above quoted author points out that a binaural system attempts to isolate the listener from the outside acoustic by the use of headphones (telephone receivers, in the definition) and playing back exactly what the listener would hear in the best position in a concert hall. According to the author, stereophonic systems are, on the other hand, intended to contain less of the acoustics of the space and more of the close sound of the instruments, since microphones are placed near them. These ideas generated a lot of misunderstandings since we know it to be possible to obtain a stereophonic reproduction that contains a lot of spatial information.

The author, trying to clarify these concepts, also affirms that 'with binaural we are recording and reproducing ear signals, whereas with stereo we are reproducing the orchestra itself, and the soundstage surrounding it, on a macroscopic scale in the playback room' (Eickmeier, 1989: 3). It is interesting to note that the concept of a binaural reproduction system was closer to the concept we have nowadays than the current concept of stereophonic systems, what means that this last one was still being developed. The author mentions analogies of the stereophonic reproduction system to making two holes in the wall that separates the listener to the concert hall. He highlights that this analogy is not

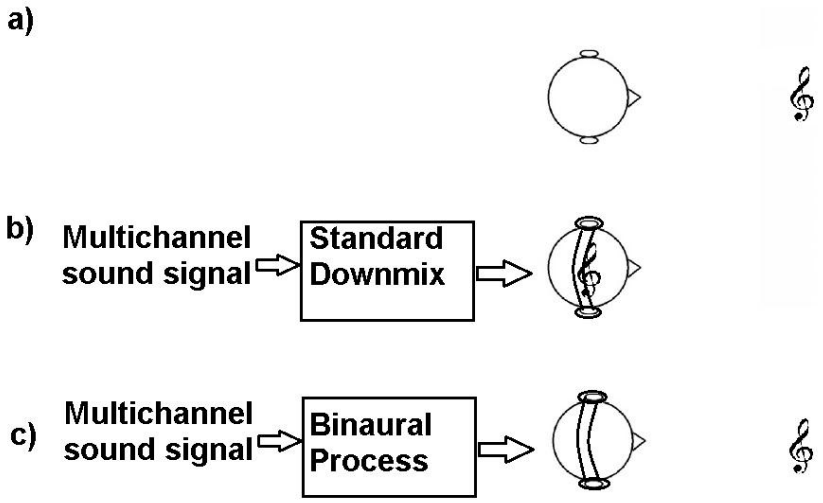
very well accepted by him since the reflections of the listening room are inevitable and not replaced by those that happen in the actual concert hall.

Two years after he publish this article, in a letter to the editor of the Journal of the Audio Engineering Society, the above mentioned author points out many mistakes made by other authors while describing their developments as a ‘solution for cross talk in stereo reproduction’ or ‘simulated binaural stereo system’. He highlights that ‘the difference between the two systems [binaural and stereo] is not that one is presented on headphones and one on loudspeakers. The difference is that one reproduces signals at the ears of the listener and the other reproduces a sound field in a room’ (Eickmeier, 1991: 261).

At this point new developments and poorly established concepts start to conflict. Concepts such as ‘artificial stereo’, ‘true stereo’ and ‘transaural’ were to be established, as can be noticed in the work of Bauck and Cooper (1989, 1996). Other authors argue that two channels reproduction through loudspeakers, also known as stereophonic, fails to meet Eickmeier’s definition of a reproduction of the original soundfield and Cooper argues that the only system to recreate a soundfield would be the Ambisonic system. Billingsley’s reply argues that a recording made with a Jeklin disk can be binaural due to the fact that it retains the space between the ears and uses a baffle to simulate head shadowing effect and that ‘the success or failure of the stereophonic reproduction will also be influenced by the design and placement of the original microphones’.

Observing this discussion about the adoption of the terms ‘binaural’, ‘transaural’ or ‘stereo’, the present author sees the need of a good definition of these terms and looked for the reason of these misunderstandings of concepts. In this dissertation I adopt the term ‘binaural’ for a two channel audio signal that contains all the spatial information needed to reproduce a three-dimensional soundfield over headphones, ‘transaural’ is adopted for a binaural signal adapted for loudspeaker reproduction, as described by Bauck and Cooper (1989, 1996), and ‘stereo’ is referred as a two-channel audio signal that can be reproduced over headphones or loudspeakers but is not capable of reproducing three-dimensional

soundfields and is the most common format adopted by music industry.



Different listening conditions: a) listening to natural sound sources, b) headphone listening of stereo material, c) headphone listening of binaural material. Derived from George, 2009.

Looking for the origin of the conflict of concepts described above it was found that one of its origins can be the fact that Alan Blumlein, one of the first engineers to realize the lack of realism in monophonic reproductions, defines his improvements as 'binaural sound' which nowadays we call 'stereo'. His observations are related to the fact that we hear through two ears and that the differences between the sounds at each of them receive are responsible for our spatial localization capabilities. In his experiments he used two microphones, recorded separately and reproduced over two loudspeakers, a practice that fits with our current concept of stereophonic reproduction (IEEEghn website).

In his patent called 'Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems', from 1933, he wrote:

> In a binaural transmission system, the transmission circuits comprise modifying arrangements whereby the relative loudness of a number of loud-speakers is made dependent on the direction from which sounds arrive at the transmitting microphones or; directionally sensitive microphones are used with or without such circuit arrangements. The system may comprise a recording and reproducing link which may be used in conjunction with motion pictures, or a wireless link in which

case modifying arrangements may be provided in the high frequency circuits.

(Blumlein, 1933)

The above discussed conflict of concepts is interestingly recent if we consider other authors works that defined them very well in early 50's. Snow, in an article written in 1953 (Snow 1955), perceiving a need for defining techniques for stereophonic recording and reproduction, presents these concepts not as a new development but due to technological means of making it possible for cinema and music applications in large scale. The author highlights the need of establishing definitions, since most people get confused about "binaural" and "stereophonic" meanings, and quotes another possible reason for such misconceptions: 'Alexander Graham Bell, writing in 1880, referred to the "stereophonic phenomena of binaural audition", in describing experiments on the directional sense of hearing conducted with his newly invented telephone' (Snow, 1955: 43). Following is presented his definitions of binaural and stereophonic system as well as some observation about their behaviour while reproducing systems:

Binaural

A system employing two microphones, preferably in an artificial head, two independent amplifying channels, and two independent headphones for each observer. This duplicates normal listening.

Stereophonic

A system employing two or more microphones spaced in front of a pickup area, connected by independent amplifying channels to two or more loudspeakers spaced in front of a listening area. This creates the illusion of sound having direction and depth in the area between the loudspeakers.

It is very important to distinguish between these systems. A binaural transmission system actually *duplicates* in the listener's ears the sound he would hear at the pickup point, and except that he cannot turn the dummy head, gives full normal directional sense in all directions. A stereophonic system produces an abnormal sound pattern at the listener's ears which his hearing sense *interprets* as indicating direction in the limited space between the loudspeakers. It has been said aptly the binaural system transport the listener to the original scene, whereas the stereophonic system transports the sound source to the listener's room.

(Snow, 1955: 43)

Snow also presents the concepts and a description of monoaural, diotic, monophonic and pseudo or bridge stereophonic systems, as well as symbolic schematics to represent them. As an extension of stereophonic reproduction systems, the principles of what we know nowadays as Wave Field Synthesis is presented as an 'ideal stereophonic system', where a curtain of microphones pick up the signals from a sound source in a stage and a curtain of respective loudspeakers in the other environment reproduce these signals. Snow's concept of stereophonic system can be confused with our current concept of multichannel reproduction but the truth is that it is the primary extension of stereophonic systems, presented as a practical implementation, being the adoption of three frontal channels instead of two, a standard adopted by the movie industry since then.

Snow's work was reviewed a year after by Postal (1954) and published in the Journal of the Audio Engineering Society following a historical development of stereophony dating from 1911 with the 'true binaural' reproduction presented by Augustus Rosenberg, whose 'patented a talking film system that used two separate sound recording channels to feed left and right sound to the loudspeakers at the left and right of the screen' (Hope, 1978: 97). Postal's review goes through multichannel sound of 1930's by Bell Telephone Labs and the Fantasound by RCA and Walt Disney, Kuchenmeister's patent on delay simulation of binaural and stereophonic effect from 1924, and Offenhauser's stereophonic demonstration from a mono source in 1939.

Postal reports discussed topics related to Snow's work presentation, when the presenter argues that the pan-pot method of generating stereo images would help generating stereo effect but should be eliminated or be used as a 'last resort' due to the fact that it will never be as good as the original. It is interesting to notice that panning mono sound sources instead of recording the real effect became the most common recent practice and the basis of music and cinema industry. This is probably due to the need for control of individual sound sources and the fact that the real effect can, in practical terms, only be manipulated by different stereo mic setups. The use of equalization or phase displacement instead of

amplitude panning is also discussed in the above mentioned session, as well as the combination of both, aiming at a better representation of sounds over a pair of loudspeakers.

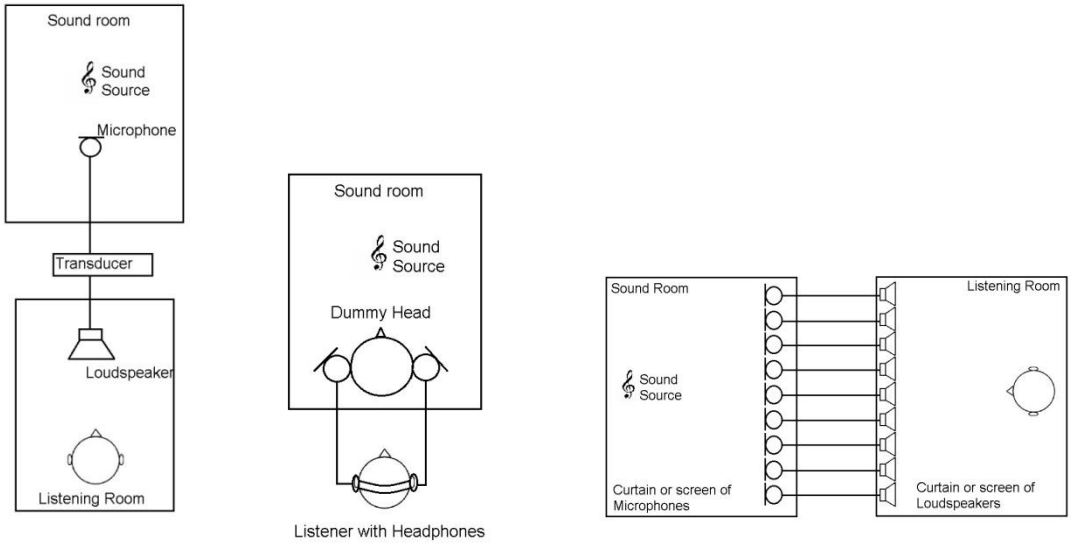
It is clear that the binaural concept was established well before the stereophonic one as is reported by Offenhauser when reviewing the history of stereo

> Soon after the telephone was invented, before the turn of the century, in 1881, an experiment was performed at Paris Opera in which two separate channels were used, each consisting of a telephone transmitter, an interconnecting wire and a telephone receiver. History records that it was quiet logical at the time [Caroll, L. 1897] to place one receiver at one ear and the other at the second ear.
>
> (Offenhauser, 1958: 67)

Offenhauser's review also mentions Rosenberg's patent for two channel recording and reproduction applications, as well as the work of G. W. Steward named 'Physical review' from 1912-1920, the work of Harvey Fletcher, Kuechenmeister's patent on delay effect from 1924, Otto Zobel's 'Theory and design of uniform and composite wave filters' from 1923 and 'Transmission characteristics of electric wave filters' from 1924, J. R. Carson's 'The building-up of sinusoidal currents in long periodically loaded lines' from 1924 and Kuechenmeister's patent applied to binaural recording from 1925.

At that time authors presented interesting analogies for the loudspeaker being a mouth and the microphone the ears, which led to what Offenhauser called the enigma of 'two ears and one mouth' and then 'three ears and three mouths' or 'two ears and two mouths', referring to multiple loudspeaker and microphones arrangements. He discusses works between 1931 and 1936 that suggests that 'it is possible to move a virtual source of sound produced by loudspeakers by means of circuitry and hardware when the real event occurred without such movement' (Offenhauser, 1958: 68) and the beginning of research on the understanding of the hearing mechanism, such as 'Hearing' from Stevens and Davis from 1937 and the work of Bekésy. This discussion presents similar aspects to the above mentioned Snow's discussion over pan-pot systems to allow mono sources be spatialized over a stereo pair of loudspeakers.

Belar and Olson defined in 1960 that an ideal system for home sound reproduction should reproduce sound indistinguishable from the original. They propose a monophonic system with a single loudspeaker reproducing material recorded from a single microphone and an artist singing in the same space as the original reproduction through the speaker. A second ideal system would be a binaural one where two microphones pick up sound from an environment and reproduce them through a pair of headphones by supplying the listener's ears with the same sound pressure supplied to the microphone. The third system, very similar to the ideal stereophonic system described by Snow, would be able to capture the soundfield by a curtain of microphones and reproduce it into the listening environment through a curtain of loudspeakers. According to the authors 'each additional channel helps, but the most striking improvement is gained when changing from a monophonic sound system to a two-channel stereophonic sound system' (Belar and Olson, 1960: 9).

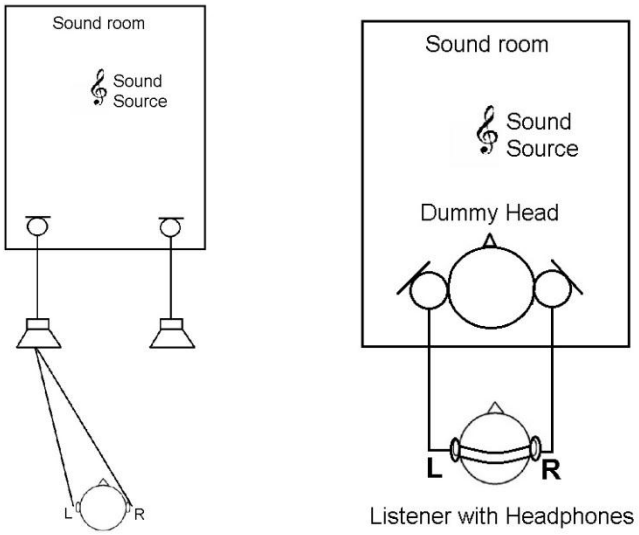


Ideal stereophonic systems described above. Derived from Belar and Olson, 1960, and Snow, 1955.

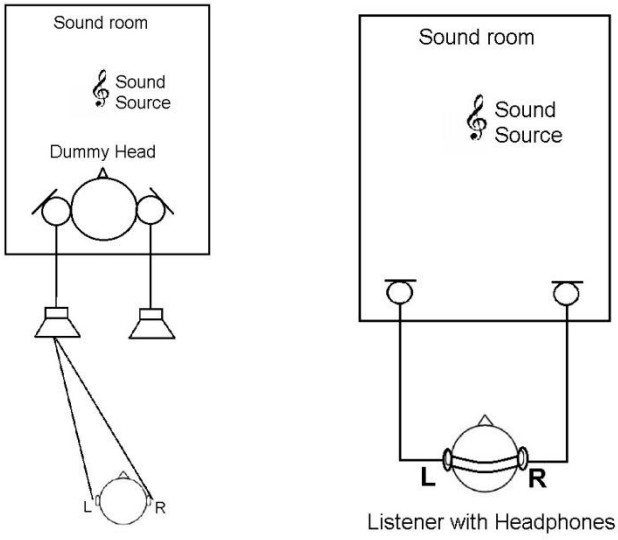
After making comparisons between monophonic system and stereophonic and exulting stereophonic reproduction qualities, also represented through binaural reproduction, the authors consider the 'binaural' concept as a hearing capability when they affirm that in the

same way people hear through two ears, a system must be binaural. They argue that any systems ‘which do not increase the amount of information transmitted are inherently deficient in performance’ (Belar and Olson, 1960: 11), an affirmation that lead us to think about subsequent developments in multichannel, Ambisonics and other 3D audio reproduction systems.

Other very interesting work that needs to be reviewed is Bauer’s. In his article from 1961, he describes the principles of crosstalk cancellation in binaural material reproduction over loudspeakers and establishes that ‘stereophonic sound is recorded for reproduction over spaced-apart loudspeakers. When earphones are substituted for loudspeakers the stereophonic space perspective is distorted’ (Bauer, 1961: 148). The author also presents the main differences between binaural and stereophonic reproduction when discussing that in binaural systems, two microphones are placed 8 inches apart on a dummy head or baffle, these two signals are reproduced through earphones. Disregarding head movements effects and front-back confusions, ITD and ILD are reproduced as if the listener was actually in the place of the microphones giving real fidelity to spatial localization of sound sources. In stereophonic reproduction through loudspeakers, spaced microphone pairs are placed in the recording room and suitably distributed between the loudspeakers, the listener will receive signals of each one of the loudspeakers in both ears but a sound source located in the extreme left will be heard predominantly coming from the left speaker. He argues that ‘an important observation here is that proper space perspective requires a precisely determined cross-feed between the two ears. With the natural arrangement this occurs at the microphones in the dummy head. With a stereophonic arrangement it occurs at the observer’s ears’ (Bauer, 1961: 148).



Binaural vs. Stereophonic systems. Derived from Bauer, 1961.



Incorrect perspective of binaural and stereophonic reproduction. Derived from Bauer, 1961.

According to the author, if a binaural recorded signal is played back over loudspeakers or a stereophonic signal through headphones, the previous well established situation is no longer performed, resulting in an incorrect perspective of the recorded space.

> As a result the observer will perceive a virtual image practically directly midway between the loudspeakers. Therefore, if loudspeakers are used for reproduction of a binaural program, much if not all of the directional information is lost. Stereophonic programs heard in this manner [through

> headphones] provide the sensation that the extreme left or right sounds originate directly outside the observer's ears, resulting in a gross distortion in space perspective. The reproduction is almost bizarre, with a sensation that the various instruments form a "musical hat" on the observer's head.
>
> (Bauer, 1961: 149).

He concludes describing the design of a system that, to reproduce stereophonic signals through headphones, can simulate ITD and ILD as well as an impulse response (HRTF) correspondent to 45 and -45 degrees. For reproducing binaural through loudspeakers a hypothetical crosstalk cancellation circuit is presented.

In the 1990's the above conflicted concepts seem to be solved and applications for binaural technology start to be the main subject of most researches. The developments in commercial products – Dummy Heads – for good binaural recordings are reviewed by Gierlich in 1992.

> … the first experiments with an artificial head having been carried out as long ago as 1886 in the Bell Laboratories. 1939 saw the development of a forerunner of the modern artificial head at Philips. In this head design, the microphones were located in the approximately simulated pinna of a female head. This was the first system used for electro-acoustic transmission. Further developments occurred in Berlin, Göttingen and Aachen.
>
> The first professionally used artificial head was introduced by Kurer, Plange and Wilkens and has been built since 1973 as the "Neumann" artificial head. In the interim this system has been improved several times.
>
> In 1975 the "Kemar" artificial head was introduced by Burkhard and Sachs. This head had been designed for acoustic research and found applications in the measurement and design of hearing aids. The Bruel and Kjaer artificial head HATS, also designed for measurements engineering purposes, is based on the geometrical data arrived at by Burkhard and Sachs and was introduced in 1985.
>
> (Gierlich, 1992: 220-221)

According to the above mentioned author, despite all the developments aiming at obtaining a good binaural recording, this reproduction system did not become popular and the main reason for that was its incompatibility to loudspeaker system reproduction. In 1973, as soon as artificial heads were introduced into a studio, the noticed enthusiasm for

binaural recordings dissipated due the three main factors: the lack of loudspeaker compatibility, the insufficient localization, especially frontal-rear, and the insufficient signal-noise ratio. He argues that modern artificial heads no longer have all these disadvantages.

An interesting development highlighted by the author and that makes binaural processing more interesting is the possibility of using impulse response convolution of previous measured signals from a Dummy Head with mono inputs to obtain a directional controlled binaural output. This led also to the development of binaural mixing consoles as well as the creation of a binaural space for locating these mono sources, i. e. a binaural reverberation processor, which also can be created by convolving binaural sound sources with binaural impulse responses from a room (Gierlich, 1992: 231, Jot et al. 1995). The main difference between a mixing console that can deal with binaural audio and others that deal with stereo material is the HRTF processor in the panning section and the binaural room simulator or binaural impulse response reverberator. A more interesting design, as reported by other authors, Gerzon for instance, would be an Ambisonic mixing console that deals with B-format signal and can perform a downmix for traditional stereo, as well as 5.1 and binaural formats. For reproduction of binaural recorded material over loudspeakers, Gierlich, as well as other above mentioned authors, affirms that a crosstalk cancellation process must be implemented and refers to three articles about this subject: Damasle and Mellert 1969, Bauck and Cooper 1989, and Moller 1989.

Moller in more recent work states that

> The idea behind the binaural recording techniques is as follows: the input to the hearing consists of two signals: sound pressures at each of the eardrums. If these are recorded in the ears of a listener and reproduced exactly as they were, then the complete auditive experience is assumed to be reproduced, including timbre and spatial aspects. The term *binaural* recording refers to the fact that the *two* inputs to the hearing are reproduced correctly.
>
> (Moller, 1992: 171-172)

According to the author, the recording can be made with a microphone inside the ear canal of the subject or with a dummy head (more common practice) and played back through headphones to ensure that the signal correspondent to each ear reach only that ear and not the other - phenomena that does not occur in case of loudspeaker reproduction. He holds that this kind of recording is not very common in broadcasting due to lack of mono compatibility but argues that the use of headphones and application of binaural technologies may increase due to the spread of portable players and the recent development of artificial environments projects.

Disregarding other senses and their interaction with sounds, the author argues that binaural techniques, if well implemented, are superior to any other recording technique for presenting to the listener exactly the same hearing experience he would have had in the place it was recorded. He claims that 'it gives the most valid representation of the original sound, not only with respect to timbre, but also in relation to spatial aspects' (Moller, 1992: 206).

Another important concept that grew up with the development of binaural recording and reproduction techniques is the concept of 'auralization' that, according to Hammershoi, has a common origin and the same intentions – reproduction of an authentic auditory experience by reproducing at the eardrums of the listener the same sound pressure he would receive in real life, including cues necessary for the perception of distance and direction of the sound sources.

> Auralization is the process of rendering audible, by physical or mathematical modelling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modelled space.
>
> (Dolenbäch, Kleiner and Svensson, 1993; also quoted in Hammershoi, 1996: 3)

Hammershoi also presents some experiments that, while comparing localization capabilities of listeners in natural environments and in binaural recordings or synthesized reproductions concludes that a true reproduction is possible. This affirmation leads to a

deep investigation on how to do it in its best and with currently available tools, which is one of the main challenges of the current work.

1.2.2 ILD, ITD, HRTFs and beyond

Human spatial perception of sound sources is based mainly on the differences between the signals our two ears receive. ILD (interaural level differences), ITD (interaural time differences) and HRTF (head related transfer functions) are the three main cues, but some more is to be investigated. One of the first works investigating how we perceive directionality of sound sources and applying the findings straight into the development of stereophonic systems was presented by Blumlein in the early 1930's when he affirms that

> Directional hearing sense is due to phase and intensity differences between sounds reaching the two ears, phase differences being more effective for the lower frequencies and intensity differences for higher frequencies. As phase differences in two loud-speakers (both heard by both ears) do not produce the required effect and normally reproduced intensity differences are not sufficiently marked, the modifying arrangements translate low frequency phase differences into intensity differences and amplify the higher frequency intensity differences.
>
> Directionally sensitive microphones may be of the light moving strip type, and two elements may be mounted vertically in line in a common casing and with a common magnetic system. For horizontal directional effects, there are no phase differences and the intensity differences may be sufficient without the previously described modification.
>
> (Blumlein, 1933)

Although Blumlein's work was very important for the developments that were to come, not everything was covered. In the 1950's Snow presented some of the parameters that would later be systematically described by Blauert in 1974, such as the already known time and intensity interaural differences, but also reverberation, dynamic localization and depth perception (Snow, 1955).

For binaural recording and reproduction systems, where these cues are extremely important for the final results, extensive research has been done the twenty past years.

Moller, for instance, argues that one of the most common problems of binaural systems, that is also an impediment to its success, is the lack of frontal localization, i.e. when sources that were to be heard in the frontal hemisphere are heard in the back, closer than originally presented or inside the listener's head. His explanation for that phenomenon is related to individual differences of human heads and pinnae, one of the main tools humans use to distinguish between front and back sound sources. To solve this problem, Moller, as well as many other authors, suggests the use of individualized HRTFs (personal set of recordings for each direction while performing synthesized binaural reproductions), referred below as 'coloration'. He also makes interesting associations between frequency regions and cues responsible for their respective localization that have also been reported by other authors. Moller reports that

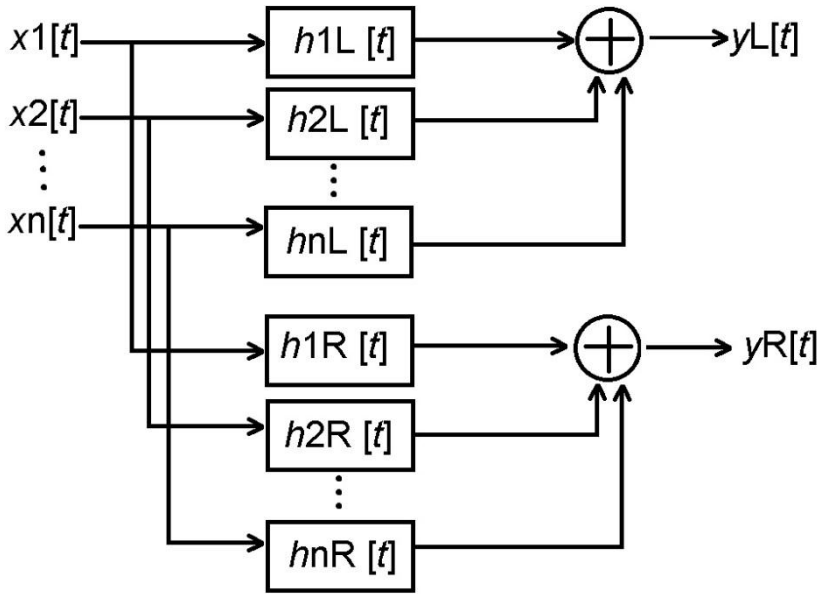
> It is traditionally said that the hearing uses a number of cues in the determination of direction and distance to a sound source. Among the cues are
>
> (1) coloration
>
> (2) interaural time differences
>
> (3) interaural phase differences
>
> (4) interaural level differences
>
> These cues are claimed to be responsible for the directional hearing in each of their "domain". For instance, in the horizontal plane low frequencies are said to be assessed by interaural phase differences, medium frequencies by interaural time differences and high frequencies by interaural level differences. Coloration is claimed to be responsible where no interaural differences exist, that is in the median plane.
>
> (Moller, 1992: 176)

According to Hammershoi, all the binaural cues can be represented by the HRTF and, when synthesizing binaural material by applying HRTFs in mono sound sources, the spatial cues can be completely fulfilled.

> Any descriptor of localization cues, such as the interaural differences in time (ITD), phase (IPD), level (ILD) or in intensity (IID), or monaural cues, group delay, etc. are *all maintained* in the HRTF. The HRTF thus represent completely and uniquely the sound transmission for the particular angle.

> In binaural synthesis the filtering which in the real life situation is carried out by ear, head and body, is done electronically. Filters representing the HRTFs are typically implemented by means of digital processing and the binaural signals are thereby generated artificially by a computer.
>
> (Hammershoi, 1996: 2)

Working with HRTFs can bring lots of problems during its implementation. One of these problems while dealing with the convolution of HRTFs during binaural synthesis is the high cpu cost and this is reported by many authors (Jot, et al. 1995, 2006). The authors point out that measuring HRTFs only in discrete positions leads to the need for interpolation to obtain a continuous set and that this process requires significant processing power. In the quest for lowering processing time and power, many possible solutions have been presented and this subject will be developed further in this dissertation (chapter 3).

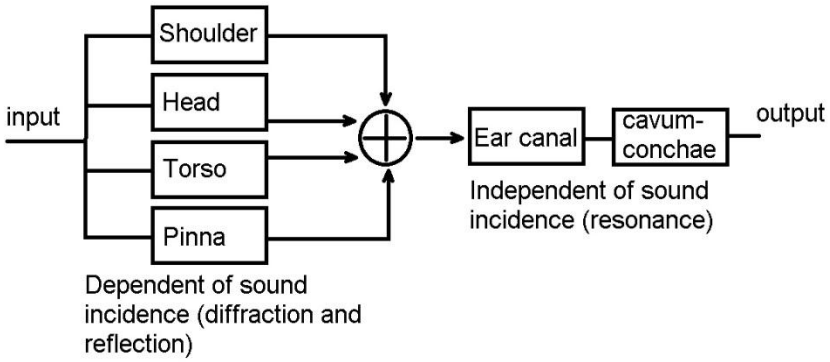


Conventional virtualization of sound sources or binaural synthesis. Derived from Goodwin and Jot, 2007.

The question of where, in the subject's external ear, the measurement microphone needs to be placed also needs to be answered. Moller (1992) states that there are three main methods to record binaural signals (at the eardrum, at the entrance to the open ear canal, at the entrance of the blocked ear canal), according to him, in these three situations, 'the

correct transfer function is obtained when the electronic circuit compensates for the microphone sensitivity and the head transfer function from the terminals to the sound pressure' at the position of the microphone in the ear of the dummy head (Moller, 1992: 189). The author also considers that the propagation of sound inside the ear canal is independent of direction and distance. Inside the ear canal what one can see is a resonance frequency dependent on the individual but independent of the direction of the sound source.

These same arguments are made by Gierlich (1992) when he argues that 'the physical effects which determine the head transfer functions are diffractions, resonances and reflections as caused by the acoustic relevant elements: head, torso, shoulder, pinna, cavum conchae, ear canal and ear drum, etc.' (Gierlich 1992: 223), and divides these elements into direction dependent and direction independent.



Basic elements of HRTFs. Derived from Gierlich, 1992.

Hammershoi (1996) also points out practical issues and compromises related to the position of the microphone for making the measurements, since it can be placed millimetres away from the eardrum or in the entrance of the ear canal. It is said that 'the transmission from the entrance of the ear canal to the ear drum is directional independent, and the entrance may also serve as recording point for binaural signals' (Hammershoi

1996: 7) and concludes that measurements at the entrance of the ear canal work fine and avoid completely the need for measurements made deep inside the ear. This kind of measurement is less influenced by individual variations and, although some differences in impedance relations are reported, they can be ignored.

Moller (1992) also points out problems in the use of miniature microphones for measurements since they are difficult to calibrate, have low sensitivity, non-flat frequency response and at high frequencies they measure sound pressure in only one point not on a surface as our eardrum does.

The third most discussed problem, briefly described a few lines above, is that a measured HRTF probably will work perfectly only for the subject in whom it was measured and a lot of discussion have been taken about the real need of individual HRTFs. Moller (1992), for instance, argues that the need of each individual for his own recording will restrict very much the application of binaural technologies. Some authors argue that it would be possible to find a set of HRTFs that would work for almost every one. Han (1992) is one of these authors that state that there are better HRTFs and worse ones. Gierlich (1992) describes demonstrations performed in 1975 with binaural reproduction of pre-recorded sounds, when it was proven that there are HRTFs better than others and that they would work for more than one subject. Some other authors (Jot et al., 1998a) affirms that individual HRTFs for frequencies below 5KHz are not needed since there are no significant differences between individuals up to this region. Begault (1991) is another author that argues that particular HRTFs can work better than others, especially for externalization effect.

Malham (1990, 1998a, 2001b), Jot et al. (1999, 2006), Hammershoi (1996), on the other hand, argue that there is the need of individual HRTFs. Hammershoi for instance has done extensive work on experimenting individual and non-individual HRTFs. Some of them comparing localization errors in real life and with different HRTFs conclude that ‘nonindividual binaural recordings results in localization errors such as front-back

confusions, elevations, and in-the-head localization'. An interesting observation is that in none of their tests they consider head movement (subject of the next section) or other sensory stimulus.

> When compared to real life, the localization performance was preserved with individual recordings. Non-individual recordings resulted in an increased number of errors for the sound sources in the median plane, where movements were seen not only to nearby directions but also to directions further away, such as confusion between sound sources in front and behind. The number of distance errors increased only slightly with non-individual recordings. Earlier suggestions that individuals might localize better with recordings from other individuals found no support.
>
> (Hammershoi et al., 1996a: 451)

In the same experiments they also report that with non-individual HRTFs, localization errors increased mainly for median-plane errors (front-back reversals). The idea that non-individual HRTFs bring sound sources closer to the subject, i. e. changing the perceived distance, was not supported but front-back reversal were perceived mainly by frontal sources being perceived in the back. It also seems that individual recording present better elevation cues. They also report that the suggestion that there could be some HRTFs which were better than the others was not supported (Hammershoi et al. 1996a: 464).

In other papers, trying to define a 'typical' subject head, able to perform better localization than others, was not successful. They report it to be possible to have a good HRTF that worked better for a lot of people but it will never be better than individual HRTFs (Hammershoi et al., 1996b).

A question that needs to be answered considering these tests is that, although they report subjects' comments about the fidelity to reality of the headphone reproduction, where some subjects, while wearing headphone had the sensation of the sound coming from the external loudspeakers, is how can one compare natural reproduction (from loudspeakers or real sources) to headphone reproduction, while there is a process and a change of subject status between the two reproductions?

Considering all these factors that influence binaural reproduction, Anderson et al. 2001

performed experiments to establish which parameter is more significant. In their conclusion it is clear that a factor that was being neglected by most authors was the reverberation and that this is the most significant one. Jot et al. (1995, 1998a) also noticed it and propose the possibility of recording HRTFs sets in live room which would include reverberation, named BIRs (Binaural Impulse Responses). The possibility of creating models to describe environments and their particular reflections is also referred to, but the quality of the sound would be extremely dependent on the complexity of the model. Clearly, some attention needs to be directed to the recreation of reflections models or binaural impulse response measurements considering their importance for a successful binaural reproduction.

The influence of the headphones on the binaural material which is going to be played back is also analysed as an important element to be considered. First of all, due to the fragile nature of spatial cues based on minor frequency variations, the headphone frequency response needs to be compensated and this is observed as an important variable for the success of all the above mentioned experiments.

Moller (1992) highlights the importance of headphone equalization since it also contributes to the total sound transmission and the correct reproduction of the sound pressure at the eardrum can only be guaranteed if the characteristics of the headphone are also known. He use the term ‘open headphone’ to define a headphone that ‘does not disturb the radiation impedance as seen from the ear could be a relatively small unity positioned some distance from the ear’ (Moller, 1992: 188). This concept is not the same as adopted by commercial use where it is generally taken to mean that sounds from outside can be heard whilst wearing the headphones. As recording from the entrance of a blocked ear canal also need to be compensated for ‘the transmission difference caused by different acoustic source impedances in the two listening situations’, when the open headphone is used this last compensation is not necessary.

Another characteristic of headphones that needs attention is that a headphone with free-

field equalization can be used for a traditional stereophonic reproduction ‘since this type of headphone simulates listening to the direct sound from a loudspeaker with an ideally flat free-field frequency response’ (Moller, 1992: 198). The use of diffuse-field equalized headphones on the other hand would represent some reflections and might be more suitable for compatibility with loudspeaker reproduction.

Jot et al. (1995) refer to the achievement of an ‘open headphone’ characteristic as another filter ‘for compensating the coupling of the earphone to the ears’ that the mono sound sources need to pass through after the HRTFs. Hammershoi et al. (1992, 1995) makes an analogy of the ‘open headphone’ reproduction to the Thevenin model and define it as a FEC (free air equivalent coupling) headphone that would be ideal for binaural reproduction since ‘the use of blocked ear canal in measurements of headphone transfer functions reduces the individual variations considerably’. They also observe that ‘only the “headphone” consisting of small loudspeakers mounted away from the ears proved to have FEC properties if a strict criterion is used’. With regard to the equalization of the headphone, the authors argue that this must be carefully done and is also dependent on the subject. An ‘equalization with an average curve may be accepted though, since the errors that occur for each individual are characterized by dips rather than peaks’ (Hammershoi et al. 1995: 216).

Azzali et al. (2005) refer to headphone equalization as

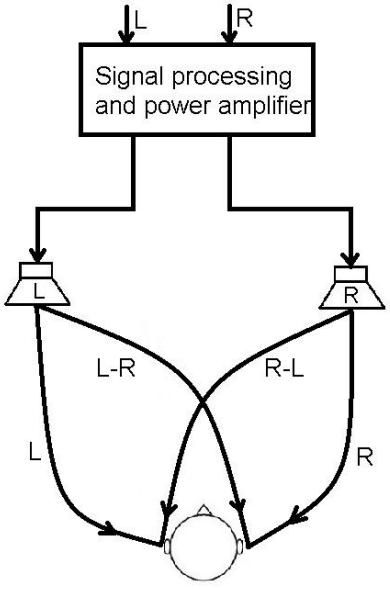
> probably the most subtle key aspect of using a non-individualized binaural headphone system: simply reproducing a dummy head recording over unequalized headphones means that the sound is subject to the manufacturer’s designed frequency response (which is unlikely to be optimized for binaural reproduction), and subject to effects of both the dummy head ear and listener’s ear effects.
>
> (Azzali et al., 2005: 2).

Binaural reproduction over loudspeakers, also known as transaural system, was first describe by Bauer in 1961 and has received big efforts the past few years due to the fact that it can be the key development for the adoption of binaural technologies, since it is extremely compatible with the already existed stereophonic systems through loudspeakers.

For binaural material to be reproduced over a spaced pair of loudspeakers some processing need to be done, known as cross talk cancellation and Moller presents a very interesting definition of this process:

> The good directional characteristics of an artificial-head recording are destroyed if it is reproduced through loudspeakers. This is due to the crosstalk, which is introduced in any free-field situation. Crosstalk means that the right speaker is heard not only by the right ear, but also with the left ear, and vice versa. However, it can be shown that it is possible to add an artificial crosstalk which cancels out the natural crosstalk. Systems that perform crosstalk cancellation on binaural systems are sometimes called *transaural systems* or – earlier – *TRADIS systems* (true reproduction of all directional information by stereophony).
>
> (Moller, 1992: 199)

Limitations of this system exist of course and are mainly related to the effects of head movements, mentioned by Jot et al. (1995) as constraints on position and orientation of the listener's head, also inherent of binaural reproduction over headphones. Despite this limitation, Jot et al. argue that the choice of reproducing binaural content over a pair of frontal loudspeakers can be interesting in the sense that frontal images can be better but lateral rear and elevated sources are compromised (Jot et al., 2006: 13).



Crosstalk cancellation scheme. Derived from Moller, 1992.

Azzali et al. (2005) while performing a comparison between stereo systems, including binaural over headphone, traditional stereo and transaural (single stereo dipole and double stereo dipole – with a pair of loudspeaker in front and an additional pair in the back for 360 degrees of horizontal reproduction), lists other limitations of the system, such as the requirement of an absorbent environment for reproduction, the critical head position for high frequencies and small or negligible ILD at low frequencies. They argue that solutions for crosstalk improvement are positioning the two loudspeakers closer to improve high frequencies reproduction in the cost of low frequencies that can be reproduced by different set up of speakers or just played back without crosstalk cancellation.

1.2.3 Limitations and new developments

Many authors have been describing the problems of binaural reproduction through headphones and possible solutions to make this kind of reproduction more realistic. The three main implementations most of them report in their works as solution for the problems this system present are also reported by Anderson et. al (2001):

- head tracking, to keep the sound source in a constant position in relation to the listener
- individualized HRTFs, for easy recognition of sound space by the subject
- realistic representation of diffused reverberant sound fields

In the above mentioned work, the author affirms that one of the biggest problems other researchers have in evaluating these different parameters and their influence in binaural reproduction is that they always evaluate those three parameters separately, whereas they propose a simultaneous evaluation. Another consideration made by the authors is related to the signals researchers have been used for testing these parameters and in their proposal they include ‘real world’ sounds instead of test sound (clicks or noise), choosing speech signals to perform the test. In their conclusions the main point we need to consider is that reverberation is proved more important than individualized HRTF or even head movement

simulations. Considering the three error concepts the authors introduce in their work: externalization (judgement of distance errors), localization (azimuth or elevation errors) and reversal (front back confusion), they conclude that

- azimuth and elevation perception are affected mainly by reverberation, no significant affect were perceived for head tracking or individualized HRTF
- head tracking reduces reversal (front-back) significantly, HRTF or reverberation are not that significant to correct this kind of error
- reverberation is the only one to affect significantly externalization perception

The above mentioned errors during binaural reproduction are reported in many other works. Begault for instance lists the three major challenges as follows:

> 1) Eliminating front-back reversals and intercranially heard sound, and minimizing localization error
>
> 2) Reducing the amount of data necessary to represent the most perceptually salient features of HRTF measurements
>
> 3) Resolving conflicts between desired frequency and phase response characteristics and measured HRTFs.
>
> (Begault, 1991: 865)

He also points out the dependence of the binaural reproduction on the material and the system it is being reproduced through. He highlights that 'broad-band, impulse sounds will be easier to localize to a specific position than low-frequency sounds with slow amplitude envelops' and that 'the nonlinearities in amplification, headphone frequency response, and donning of headphones by the listener are additional sources of error in any audio reproduction system' (Begault, 1991: 865). Despite the known problems of adopting generalized HRTFs the author points out that the solution of using personalized HRTF is definitely not a practical one, goes on to argue that since the objective of researches is to obtain a generalized HRTF that can perform good spatialization for the overall population, it can be achieved by manipulating synthesised HRTFs by average, structural modelling or component analysis.

The limitations of binaural reproduction through headphones is also reported by Toole

(1991) whose argues that timbral distortions, front-back confusion and difficulty of externalizing sound sources are generated mainly due to non-individualized HRTFs, faults in headphone coupling, conflict between visual and auditory cues, lack of head tracking systems, loss of tactile sensations and vibrations. He add that the use of headphones that can isolate the subject from the outside world, such as those that has a compressible foam plug, can generate better binaural reproduction (by 'transporting' the listener to other locations).

Another improvement, observed by Dalenbäck et al. (1995), would be the implementation of Doppler Effect to simulate moving sources, which can be obtained by modifying the delay between source and receiver sample by sample.

Travis (1996a, 1996b) points out the need for head tracking systems as a priority for virtual reality implementations and mentions work by Wallach (1939), where the quoted author conclude from experiments that individualized HRTFs are not as important as head movement simulation. This fact leads to a big problem that concerns virtual reality implementations and perhaps explains the failure of dummy head recordings by the music business. The inability of this two channel signal to be changed by head tracking processing generates all the problems already listed and the development of material that can be adapted for loudspeaker reproduction as well as for binaural reproduction can be the solution. He argues that, although most people prefer loudspeaker reproduction, headphones are much cheaper and much better in the sense that they reach their design goal better than loudspeakers. The non-adoption of binaural technology is restricted to the means by which it is presented and broadcasted. The author concludes that the same way surround cinema saved multichannel reproduction maybe virtual reality applications may save binaural reproduction.

Developments in headphone design have also being subject of many researches. Hammershoi's observations that the direct sound is just part of the sound captured by the microphone and that the reflections are the biggest part of it proves that a diffuse-field

calibrated headphone would perform better localization. The author argues that ‘it is easily verified that the sound transmission from the entrance of the ear canal to the eardrum is equal in the two situations, since the transmission only depends on the ear canal and its termination’ (Hammershoi, 1996: 221). The author’s experiments with commercial headphones that claim to be diffuse-field calibrated prove they are not, according to the author’s design of a headphone of that type.

A very interesting development is reported by Gan and Tan (2000) on a headphone prototype that avoids the need of individualized HRTFs. The observation that binaural reproduction over loudspeakers, despite the small sweet spot, presents good reproduction of frontal images and good externalization lead to a type of headphone that has a different position in relation to the ear. In their prototype headphone the position of the small loudspeakers is changed to simulate FEC (free air equivalent coupling) characteristics, to simulate the sound pressure on the ear as it was in a free field. The role of the external ear in the individual spatial perception is maintained, and so generalized HRTFs can be used. They argue that ‘the in-the-ear headphones preclude any kind of concha excitation. The circumaural and supraaural headphone types to include some ear interaction since the transducers are placed in parallel over the ears’ (Gan and Tan 2000: 644) and this motivated their work on a different kind of headphone that can still excite the subject’s concha.

They observe that headphone reproduction can present good lateral representation due to the fact that the concha is excited as if by lateral sounds, on the other hand frontal loudspeakers always represent good frontal images due to their position in front and directivity that excites the concha as if by frontal sounds. Their proposed headphone with improved FEC can improve naturalness and in their experiments they conclude that the concha headphone affect strongly frontal images but not rear ones, which is perfectly suitable for headphone binaural reproduction since rear images are well reproduced and the biggest problem is frontal sound sources representation. He also reported significant

changes when small repositions are performed.

Other interesting development is reported by Nicol et al. (2007) that focus on a 'transparification' of the headphone, mainly for comparing a binaural sound reproduced through headphone with a real sound, without getting the headphone out. Jin et al. (2009) establish that 'an ideal system for delivering SARA [Spatialized Augmented Reality Audio] would present a high-fidelity VAS [Virtual Auditory Space] without interfering with normal hearing'. By saying this, the authors mean that the earphone must be transparent to outside sound sources that reach the ear. This improvement can be used to achieve better externalization since the listener will have the background ambient sound as a reference to localize artificial reproductions over the headphones (considering the implementation of a well-developed head tracking system). Nicol et al., for instance, base their work on the observation made by Hartman and Wittenberg (1996) 'that the presence of headphones suppressed subjects' abilities to discriminate between front and rear directions'. Their main objective is the comparison between real and synthesized sound sources reproduced respectively through loudspeakers and headphones. This cannot be achieved if the headphone itself changes the sound coming from an outside loudspeaker.

The 'headphone transparification' is intended to compensate the effect of wearing a headphone and make those experiments more reliable. According to the authors it 'uses low level compensating signals presented over the headphones simultaneously with the loudspeaker source to correct for the interfering effect of the headphones on the real sound (Nicol et al., 2007: 2). A problem in the implementation is that the compensation needed for 'headphone transparification' is dependent on the listener position. Although the authors propose strict immobilisation for better performance of the system, one might suggest that maybe interpolations between impulse response measurements might compensate for the variations caused by the movement, at the cost of more processing power.

Although modelling HRTFs is a very cpu expensive processing task, some proposals

have been made aiming at the reduction of this cost. Algazi et al. (1999) for instance report a method of using contralateral HRTF to create the ipsilateral one. According to the authors, while a sound source is moving, the HRTF of the contralateral ear (in the opposite side of the head, where the sound source is moving from) became more complex than the HRTF from the ipsilateral ear, and that can be simulated. They argue that 'near the median plane, the contralateral and ipsilateral responses are quite similar. Away from the median plane, simple modifications of the ipsilateral response due to the head shadow may be adequate to approximate the contralateral response' (Algazi et al. 1999: 313).

Implementing this solution they achieved localization results with 5 degrees error compared to the measured HRTFs. Observations are made during the development of the model, highlighting that it is based on a spherical head model, since HRTFs are not similar for both ear not even in the median plane while changing elevation, due to small differences in the head shape that influences the shadowing effect differently for both ears. They argue that 'it is known that ILD varies with elevation. Because the ILD using the model is independent of elevation, spectral errors are inevitable' (Algazi et al. 1999: 318).

In order to make improvements in reducing the cpu cost of HRTFs processing, Braasch et al. reduce the number of non-zero coefficients of the HRTF filters. They report that 'reducing the number of non-zeros coefficients in HRTFs from 128 to 64, 32 or 18 did not significantly affect the perceived directions of the corresponding binaural cues when the simulated directions were at ear level with azimuth angles of 0º, 90º and 270º' (Braasch et al. 2006: 704). Their tests were performed over a KEMAR set of HRTFs and with the filters correspondent to the directions of a 5.1 loudspeaker array.

Among other authors, for instance Angus et al. (1998), Cheng and Wakefield (2001) and Carpentier et al. (2010) report a calculation method to interpolate HRTFs to simulate proximity effects and they focus on comparing 'spherical acoustics expansion method for the prediction of near-field HRTFs and measurements of a *HEAD acoustics* mannequin in an anechoic room' (Carpentier et al., 2010: 1). A similar work was also presented by

Pollow (2010) describing extrapolation and interpolation methods based on spherical harmonic decomposition to obtain near field HRTFs. All these works intend mainly to reduce the HRTF measurements and the size of the data base to be processed.

As there has already been much discussion of individualized HRTFs some developments need to be presented and discussed. Nicol presents the three main methods to achieve HRTF individualization:

> *optical* measurement: for instance anthropometric parameters (from direct measure or derived from pictures) or 3D mesh of the morphology (obtained from a laser scanner or derived from pictures),
>
> *acoustic* measurement: set of HRTFs measured in a few directions which gives approximate information to feed spatial interpolation,
>
> *perceptual* feedback: the parameters are somehow the listener's perception, i.e. his or her judgement of the Virtual Auditory Space (for instance by a localization test or a multi-criteria assessment), which is used to fit the model.
>
> (Nicol, 2010: 51)

Aarts and Schobben (2005) presented a system that fits with Nicol's *acoustic* measurement concept by making measurements on the individual ears and compensating differences between wearing headphones and real presentation over loudspeakers accordingly. They also perform headphone calibration using noise cancellation techniques. Their results are claimed to have high accuracy up to 7KHz and they argue that 'from informal listening tests it appeared that listening to the proposed system is undistinguishable from listening to a true multi-channel loudspeaker set-up' (Aarts and Schobben, 2005: 449)

A very interesting example and successful implementation of an optical measurement based on 2D photographs is described by Asselot et al. (2008). They argue that since computer graphics reached a high level of realism they can be used to calculate individual HRTFs from a head model obtained by personal photographs. The fact that the use of laser scanning is very expensive, lead to the need of cheaper techniques and theirs is based on five photographs and key points indicated on them. The whole process (3D model

reconstruction and HRTF calculation) after acquisition of the input data is completely automated, fast and reliable.

They agree with the fact that there is a need for individualized HRTFs and argue that it can solve front-back confusion and the 'inside the head' effect. One of the main authors' goals was to make the process simple for the user, so a guide of how to take the five pictures is given as well as orientation on how to mark the points in the photographs. Basically one of three primary models is chosen based on the external ear contour taken from the pictures to be adapted. A process of deforming the chosen dummy is divided in three main steps: single view head deformation, global head deformation and ear deformation, all based on the key points defined by the user in the pictures. A scale parameter, given by the user (such as the size of the nose) is used to scale the 3D model and, according to the authors, is extremely important for computing the HRTFs. They report a successful 3D model when compared to laser scanning measurements, and their results indicate that the system proved to be fast and to guarantee sufficient geometric accuracy, considering they can represent head features and influence the HRTF profile.

Some thoughts need to be discussed going beyond what the measurable auditory cues can determine in terms of spatial localization. The first of all is the influence of visual cues, first reported by Postal (1954) while describing a discussion in the SMPTE 74th convention. He highlights that the influence of the image for sound localization make applications of binaural audio for cinema different from those for music. Nicol et al. (2007) argue that

> The interaction between visual and auditory stimuli may also be important for externalization. Zahorik [Zahorik, 2001] showed that in a semi-reverberant environment vision could improve distance estimation compared with when subjects were blindfolded. It is, therefore, conceivable that, if a possible sound source is visible, externalized auditory images could be produced, despite deficient acoustical cues. On the other hand, Mershon et al. [Amerson et al., 1980] found that hidden sound sources nearer and further away than a visible but silent loudspeaker were localized at the position of the silent speaker. So, it is also possible that a listener who hears perfectly synthesized

> virtual acoustical cues for a source in front of him but can see that there is no viable sound source, may experience an auditory image at some other unseen location, i.e. behind or inside the head.
>
> (Nicol, 2007: 1-2)

Begault (1999) argues that normally we study the senses isolated from each other and that studies of their interaction have been growing with the increase attention to virtual reality systems, home theatre, game developments and teleconferencing. The author presents a review of studies involving audio-visual interaction focused mainly on home theatre systems applications, followed by studies on potential interaction between audio, visual and vibro-acoustic sensation. He quote studies which affirms that reversals can be reduced by the use of head movement cues, others that suggest exposing listeners to non-individualized HRTFs for long periods to improve sound localization or 'synthesizing "supernormal" cues with larger ranges of interaural differences than normal cues' to achieve better virtualization of the sound sources.

He also points out that these implementations can actually exceed our listening capabilities and that considering other factors may be more effective and affirms: 'the influence of cognitive cues, memory and associations must also be a controlled factor' (Begault, 1999: 15). The author considers influences that go beyond the audio-visual reproduction aspects such as the environment lights, reflections, loudspeaker directivity, room modes, early reflections, position of the listener and the reproduction system, etc. He quotes some researches into audio-visual reproduction, games and TV, where the authors conclude that a good sound can be observed as a good picture, but a good picture cannot be perceived as good sound, actually they make sound worse. Another experiments quoted by the author involve vibration. He wrote:

> Why then simulate vibration? The main reason is that by ignoring or not simulating its presence, the vibration in the real environment of the listener predominates and could potentially conflict with the intended audio-visual virtual experience.
>
> (Begault, 1999: 21)

Observations about the fact that an original sound source directivity pattern, the

loudspeaker directivity (when a sound source is played back, in any system) and the directivity pattern of a virtual sound source panned three-dimensionally and processed through HRTFs will never be the same, are made by Jot et al. (1995, 1998a, 2006), Dalenbäck et al. (1995), Begault (1999) and Moller (1992). This is a topic for extensive research and a possible solution was briefly described in the first section, named the Ambisonic 'O' format (Malham, 2001a). Considering the sound source a complex entity instead of a pointing source can solve the limitations of the binaural reproduction over headphones as well as the problems of many other reproduction systems.

Toole (1991) points out the difficulties of performing blind tests. He argues that 'subjects will inevitably have information about the acoustical setting before the test commences' and these psychological effects are also reported by Griesinger (1990) when writing that

> At this point my best guess is that the ability to achieve OHL [out of head localization] depends not only on the ears of listeners, but also on their ability or willingness to suspend the evidence of their other senses.
>
> (Griesinger, 1990: 202)

Begault add that 'auditory localization judgements are highly malleable as a function of expectation or memory'. Referring to a binaural demonstration he observe that

> When, on the company's demonstration tape, a person lights a cigarette and drinks a glass of water, it is probably difficulty to imagine the virtual source to the rear simply because we know our mouth is positioned in front of the head.
>
> (Begault, 1991: 866)

Experiments involving the memory effect were performed, for instance, by Han (1992) where the author tries to prove its influence. In the experiment two sound sources are placed in front and behind the listener with closed eyes. First a burst noise filtered between 800 and 1200 Hz is played in the rear; the subject moves his head and perceives the sound coming from his back. The same sound is played in front and the loudspeaker is moved from left to right while the listener keeps his head static. The listener still perceives the sound source in the back until he is allowed to move his head. He affirms that

> Using narrow-band noise, only one directional band is activated. If head movement is suppressed, the auditory system can be tricked to a false localization. This experiment proves that movements of the source alone cannot resolve the ambiguity. The subject continues to have a false localization as long as he holds his head still. This in itself is enough to falsify the contention that localization can be fully described by static HRTFs.
>
> (Han, 1992: 10)

He add that if one swaps the back stimuli to the front, the opposite results are obtained, since the first stimulus is in front, the listener will recognize the back loudspeaker in front if it is not permitted for him to move his head. The author highlights that any experiment made that does not consider memory effect, in these cases presenting stimulus randomly, is suspect.

Other senses, such as touch, also cannot be ignored. Since the nature of sound itself is movement, it already suggests this as having some influence on sound perception. In binaural reproduction the fact that the listener is wearing headphones, a 'piece of technology' is already an unnatural situation and 'can provide a counter-cue, diminishing the degree of reality equivalence' (Malham, 2001b: 35).

1.3 Head Tracking

Spatial sound perception can be assessed by static and dynamic cues. The static cues, such as ITL, ILD and HRTFs were already described in this work and analysed by different authors. Dynamic cues can only be assessed by head tracking.

Malham, while describing homogeneous and non-homogeneous systems, defines a homogeneous system 'one in which no direction is preferentially treated' and argues that a coherent system would be 'one in which the image remains stable, i.e. is subject to no significant discontinuities, if the listener changes position within it, though the image may change as, indeed, a natural soundfield does' (Malham, 1999b: 25). Such a system would require head tracking implementation to achieve 'high-fidelity' since it implies

homogeneity and coherence. According to him, a loudspeaker system that fits to his definitions is the Ambisonic system, and, binaural reproduction systems are normally homogeneous but can only be coherent if head tracking is used. He also argues that head tracking can improve even poor binaural presentations (such as when using generalized HRTF's).

Faure (2004) describes four main ways of achieving head tracking:

> *acoustic*: The system is composed of an ultrasonic fixed emitter and a receiver attached to the listener's head. The range is limited because of the absorption of ultrasonic by the atmosphere.
>
> *inertial*: An inertial sensor is based on a gyroscope (measurement of the head's direction by single integration) and an accelerometer (measurement of the head's location by double integration). These systems are disadvantageous by a weak accuracy for low speed, and a problem of measurement stability.
>
> *optical*: The principle relies on a camera in combination with algorithms of image analysis.
>
> *magnetic*: The system is composed of an emitter (fixed) and a receiver (attached to the listener) of a magnetic field. This technology gives the best accuracy, but its disadvantage is a limited range and a strong sensitivity to electro-magnetic disturbance.
>
> (Faure, 2004; also quoted in Nicol, 2010: 42)

In any of these systems implementations two main problems are reported as needing to be solved: the latency time and the interpolation between the HRTFs for moving sources or moving listeners.

1.3.1 Historical background and psychoacoustic hearing perception

Only a limited range of documentation on the historical development of head tracking systems is available since most of them are related to the whole body movement tracking or face recognition researches and not always dedicated specifically to head movements. However, a lot of articles report the use of some kind of head tracking technology while performing their tests. Following is a description of many of tests and systems which indicate a timeline of head tracking developments according to particular need of

experiments.

Probably the first article to describe the need of such implementation was presented in 1953 by Snow (Snow, 1955). He argues that servo connections between listener and dummy head would improve localization but it is not practicable for multiple users or recording purposes. Also in 1955, Held reports some experiments related to binaural perception of changes in sound source position and concludes that head movements is definitely used for acquisition of auditory localization, conclusions previously made by Wallach in 1939 through his experiments on head movement influences on sound localization.

Some other studies related to the influence of head tracking systems improving spatial perception of binaural reproduction present contradictory results. Algazi et al. (2002), for instance, argue in their conclusions that, compared to dry sound and reverberated sound without head tracking, head tracking implementation improved sound localization significantly. The opposite is presented by Anderson et al. (2001), who argue that reverberation can give the most important spatial cues compared to head tracking and individualized HRTFs, but they also affirm that the main reason for implementing a head tracking system is to reproduce our ability to resolve what some authors named the ‘cone of confusion’ or ‘front-back’ confusion through head movements.

Moller, while discussing about the lack of frontal localization in binaural systems, argues that individualized HRTFs could solve this problem but the use of head movements is probably the main tools humans use to distinguish between front and back sound sources. He explains that ‘a right turn of the head would cause sound from frontal sources to arrive earlier in the left ear and later in the right ear, the opposite happens for sound sources behind’ (Moller, 1992: 173). This phenomenon is learnt since we are born for localizing sound sources. Despite the advantages of using head movements to localize sound sources, he quotes some experiments that argue that humans can localize sound sources even without performing any movements, which is not really unexpected, since it also depends

on other changes in the soundfield, such as reverberation, as described previously.

Also in 1992, Han presented an article describing mainly what head movements imply e.g. what changes in spatial sound perception when we move our head. Intending mainly to present 'how ILD, ITD, head position vector and memory could interact to resolve front-back ambiguities', the author describes interesting relations between movement and changes in frequency domain observed on a Kemar dummy head. Based on previous research by Blauert, (1969, 1974) and Hebrank and Wright (1974) he highlights three main conclusions that oriented his work:

> 1. The pinna encodes azimuth and elevation of a sound source in the spectrum of the first millisecond of the arriving sounds
>
> 2. The most plausible way of extracting directional information from the internal spectrum is through feature detectors
>
> 3. There is a clear connection between pinna-based spectral features and some of Blauert's directional bands.
>
> (Han, 1992: 2)

The author quotes Hebrank and Wright's results in experiments filtering white noise reproduced through a pair of loudspeakers where:

> 3.9 – 8.0 KHz low pass cutoffs induce a front sensation. Increasing the cutoff frequency within this range is perceived as an increase in the elevation angle from 0º to 60º.
>
> 4.0 – 7.2 KHz bandpass filtering is perceived in front with an elevation of 60º.
>
> 7.4 – 10.8 KHz notches induce "frontness". Increasing the center frequency of the notch causes the perceived elevation to increase from 0º to 60º.
>
> (Han, 1992: 4)

In this quotation is interesting to notice that what is primarily observed as tonal coloration is associated to spatial cues, this is clear in the author conclusion when he argues that

> …there are basically two kinds of pinna cues. One cue is near-vertical spectral edge that moves along the frequency axis depending on elevation or azimuth of the source. Cues of the second kind, which work on signal level in frequency band, do not specify an angle, but indicate whether the

source is in front or behind, or fulfil a secondary function, such as resolving an ambiguity.

(Han, 1992:6)

Whilst analysing other authors' works where HRTFs are averaged, the author argues that it is allowable for frequencies below 3KHz, a region where the HRTFs do not depend on the pinna. In comparing two different situations, first a moving source from -10º to 10º (left to right) in relation to a static listener, and second a static source in front of a listener who moves his head from right to left over the same angle, the author argues that even if the ILD, ITD and HRTFs are similar for both cases, which means there are no differences between the signals that go to the listener's ears in both situations, we perceive these situation differently. The reason is, according to Han, that listeners know the position of their heads in relation to the environment. I must add that the listener also knows the relation of the reflections of the environment and maybe that the whole body 'hears' the sound coming from the front and not back.

Han also points out the angle of the pinna in relation to the head as a parameter for front-back discrimination. It can be observed that the position of the pinna varies depending on the headphone position and this can be considered another variable for an ideal binaural reproduction. This has been reported by several authors, for instance Fels and Masiero (2011). A possibility to be evaluated is that this variable can be avoided with the use of in-ear headphones which allow less freedom for personal adjustments. In Han's conclusions, however, the author argues that while up to 2KHz localization is dependent on level variations and that front-back ambiguities can be solved through head movements, associative memory is used when head is not moving.

Other authors that support the implementation of head tracking systems include Jot et al. (1995) who argue that it would be a decisive improvement to solve the undesired in-the-head effect and enhance out-of-head localization. However he highlights the problem that it 'imposes strong real-time constraints on the time-variant binaural synthesis process', to be discussed later.

A very well-known system that implemented head tracking in binaural reproduction is described by Wenzel (1996). This system, called the Convolvotron, uses head tracking to update the HRTF filters to be convolved with the sound source and minimum-phase filters to reduce comb-filtering. The author quotes the work of Wallach (1939, 1940), that suggests that dynamic cues may be a factor in externalization and tend to dominate pinna cues, e.g. HRTFs. Concluding his experiments the author argues that

> ... enabling head motion appears to reduce the perceptual errors that are observed when synthesizing virtual sources from non-individualized HRTFs. Front-back confusions are reduced to levels similar to those observed with individualized HRTFs and there is some indication that externalization is also improved. Thus, correctly synthesized pinna cues may be quite important for virtual acoustic displays if one is to gain maximum benefit from dynamic cues.
>
> (Wenzel, 1996: 9)

The author also highlights that while in static environments listeners tend to be 'insensitive' to random synthesized reverberation, this 'will change as soon as the simulation is allowed to become dynamic'. Two interesting tables are also presented, one of them listing 3D sound systems and their performance characteristics and the other a comparison between perceptual parameters and engineering specifications of such systems.

A study of reaction times for subjects to localize 3D sounds is presented by Chen (2002) and sheds some light onto localization adaptation. In his experiment aiming to analyse time people take to identify 3D virtual sound sources through headphones the author use the Huron Lake CP4 system to synthesize the virtual sound sources and the Flock of birds motion tracking system, from Ascension Technology (resolution of 100 samples per second), for dynamic cues acquisition. The use of generic HRTFs is pointed out as responsible for individual differences and sometimes for the slow localization, arguing that those whose HRTFs are closer to the one used for the experiments presented faster localizations. The author also argues that 'since the head tracking system was introduced, front-back confusion can be easily resolved by moving the head to enlarge the dynamic cue for sound localization' (Chen, 2002: 8). The localization adaptation reported happening

‘after a long time exposure to the same sound stimulation from the same sound location’ is questionable since it can be related to the memory effect and not a real HRTF adaptation.

Hess in 2004, while performing a comparison between a pair of loudspeakers and headphone with head tracking reproductions and after defining concepts such as ‘spatial impression’ and ‘reverberance’, refers to the concepts of ‘auditory source width’ and ‘listener envelopment’ to describe his experiment focused on the influence of head tracking on the perception of ‘auditory source width’. For the experiment he uses noise stimuli, already reported by other authors as not very trustworthy. Some references where other authors use head tracking to avoid in-head effects and front-back confusion even with non-individualized HRTFs are quoted. The author also reports some listener confusion due to a greater level of externalization and a ‘high degree of correspondence of the two systems’. He concludes that the rotation of the head is irrelevant for ‘listener envelopment’ perception, the ‘auditory source width’ is presented slightly higher in the headphone system than through the loudspeakers and ‘head tracking has a negligible influence on spatial perception’ (Hess, 2004: 4).

A commercial 3D audio system is evaluated by Minnaar and Pederson (2006) and aim mainly the evaluation of sound source localization. According to them, ‘it is generally accepted that the ability to utilize head movements in a real environment greatly improves sound localization’ and that ‘in general head movements are seen to reduce directional errors in the median plane and on cones-of-confusion and particularly aid to resolve front/back confusions’ (Minnaar and Pederson, 2006: 2). From this starting point the authors present some references to very successful head tracking binaural synthesized reproductions and conclude that ‘when head tracking is used the improvements in binaural synthesis are similar to those in real life and especially front/back confusions are almost completely eliminated’. Their experiments were performed using white noise bursts and the whole system used included a head tracker system, the convolution processor, a HRTF database, binaural reverberation and equalization of headphones. The head tracker they

used updated listener's position and orientation at 60Hz and presented a maximum latency of 35ms. They confirmed the angles of accuracy for azimuth and elevation from Blauert (1974), Bronkhorst (1995), Seeber (2003), Kistler and Wightman (1989) and show that even with real sound sources there is some front back reversals. With virtual sources these front back reversals are bigger but head tracking clearly makes these discriminations in binaural reproduction more accurate.

Another interesting article on the importance of head tracking is presented by Brookes et al. (2007). Their main objective was the evaluation of head movement not only for localization but also while the listener is evaluating other audio attributes such as timbre, envelopment and auditory source width. They argue that rotation is the most significant movement for localization, the head movement is individual for each person to perform different tasks and are used in different levels to evaluate different sound parameters. They conclude that 'subjects moved their heads in wider ranges when they were evaluating spatial impression, than when localising the sources or judging timbre' (Brookes et al., 2007: 15). This makes head movements important not only for sound localization but also for spatial impression.

Short sounds, for instance percussive sounds, are observed as not demanding as much movement as other sounds, a phenomena not desirable for the authors for evaluating head movement due to movement restrictions but I strongly believe that those short transient signals simply are, by their nature, easily recognizable and localizable, then there is no need of significant head movements. Differences between the first run of tests and the second can be related to memory effect applied not just for localization but also for timbre attributes.

1.3.2 Applications, limitations and new developments

Considering binaural reproduction allowing some interaction with Ambisonic systems, as described by Malham, a head tracking system must be implemented and the sound sources need to be delivered individually. He argues that 'all current head tracking binaural systems use either totally synthesized auralisations, or they auralise an underlying Ambisonically encoded soundfield' (Malham, 1996: 98).

One of the application that uses binaural reproduction with head tracking is known as 'auralization' and defined very well by Dalenbäck et al. in 1993 as 'a term introduced to be used in analogy with visualization to describe rendering audible (imaginary) sound fields'. This process is considered a recreation of an aural sensation a listener would have in a hall produced by real sources or loudspeakers and the first attempt to do so were made by Spändock in the 1930's using physical scale models (Dalenbäck et al., 1993: 2). Presentations are made in real time and must consider source and receiver directivity, sound absorption, wave phenomena, etc. It also needs to simulate different angles of incidence of direct sound in objects and surfaces as well as sound diffraction.

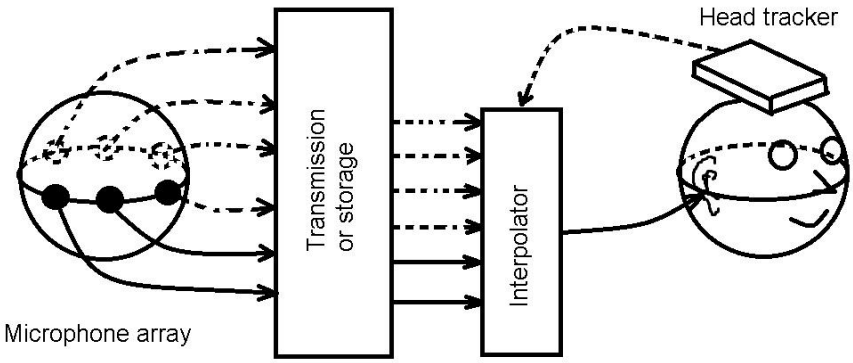
Since this first attempt was intended to be reproduced through a binaural system based on headphones, the author describes the three main problems of binaural reproduction: in-head localization, back-front ambiguity and lack of head tracking. The two first are related to non-individualized HRTFs and implementing head tracking can reduce the two first problems during reproduction. The need for accuracy raises the question of the differences mainly in the directivity pattern between a real sound source emitting sound in a concert hall and a loudspeaker doing so.

McGrath and Reilly, in 1995, described an auralization system using Lake's Huron convolution system, where the DSP system allows big impulse response to be convolved in real time. The reverb is divided in two parts, the early reflections and the tail and the changes due to head movements, captured by InsideTrak Polhemus head tracking system

with limited range but high accuracy, processed only the early reflections, saving cpu cost and making the whole process a bit faster. Although they used only 128 head positions (32 azimuths and 4 elevations angles) for the interpolation process of the reverb, they conclude that the tail part can be stationary and the experiment is reported successful for four simultaneous sources with no perceived latency.

An interesting development of the binaural system that also implemented head tracking is presented by Algazi et al. This differs from the others in the sense that it does not rely on HRTFs convolutions. A spherical microphone array is controlled by the head tracking system to simulate the listener's head translation and

> The basic idea is to distribute a number of microphones over a surface that approximates the listener's head, and to use a head tracker to determine the location of each of the listener's ears relatively to the microphones. If one of the listener's ears happens to coincide with a particular microphone, the signal from that microphone is sent directly to the listener's headphone. If the ear is between two microphones, the signal is interpolated and send to the headphones.
>
> (Algazi et al., 2004: 1142)



Basic components of motion-tracked binaural system. Derived from Algazi et al., 2004.

The main objective is to capture the dynamic cues of the sound field and the great advantage it has is that it can be used by many listeners simultaneously. The authors consider this recording a binaural one liable to crosstalk when presented over loudspeakers and argue that due to the fact that there is no pinna or HRTF involved in the process, 'the

signals lack the listener-dependent spectral cues for elevation'. They also affirm that front-back confusions 'can be solved completely if head motion is taken into account'. They based their work on de Boer and van Yrk dated from 1941 that 'showed that front/back confusion could be eliminated with a spherical dummy head by rotating the head back and forth, provided that the listener turned his or her head back and forth in synchrony with it'.

By quoting important works such as Horbach et al. 1999, Wenzel's work with the Convolvotron in 1988 and Begault's book '3D sound for virtual reality and multimedia' from 2004, they highlight that, although the use of head tracking to simulate changing HRTFs has been common practice, dynamic cues are strong enough to dominate pinna cues, which is one of the reasons they adopt this spherical model of dummy head with multiple microphones.

In their implementation they interpolate between microphones to generate the sound that would come from any regions where there are none. Errors occur while doing this, mainly phase interference, producing comb filtering. They also describe the problems of an eight microphone implementation and argue that it would work fine for speech but not for music applications. To do so the number of mics needs to be increased. They report that low frequencies spectral cues are presented correctly but there is a lack of resolution at high frequencies and suggest adding an omnidirectional microphone to make response in this region better. They also describe three main types of applications: panoramic – only horizontal microphones; frontal – horizontal but limited to frontal microphones covering 45 degrees; and omnidirectional – with microphones equally spaced through the whole sphere. At this point their approach touches previously discussed Ambisonic systems as well as their problems and the possibility of working for multiple listeners.

Developments in head tracking also allow some researchers to acquire listener position and orientation. Härmä et al. 2004 reported using a binaural system for such purposes and then using this data to process something else, i.e. images. In this application an anchor sound source is used and positioning is done by detecting this known sound and calculating

time and level difference between the signals of the two ears, obtained by a microphone positioned inside the listener's ears canal. The main problem in this application is reported as being the background noise but a successful experiment was made comparing this technique of head tracking to an electromagnetic system.

Horbach et al. 1999 describe binaural with head tracking systems applied to auralization applications. Their main objective was to offer the sound engineer a virtual environment similar to that which he was used to work in through auralization techniques. The authors argue that it would be impossible to get such a system without the simulation of the head movements and the implementation they propose is based in a mechanical system to move the dummy head according to horizontal movements of the listener's head. They named this system 'binaural room scanning'.

While establishing the importance of head movements in sound localization the authors argues that 'the movement of the head leads to a different sound source position relative to the head's orientation, and hence results in different interaural time and level differences, that is altered spectra of the ear signals induced by the HRTFs' (Horbach et al. 1999: 3). The authors recognize the importance of visual cues and quote some research done on the subject. Their experiment points in favour of the use of head tracking since 'ambiguities vanish almost completely' and even if non-individualized HRTFs were used, they report very good results. Four main application of such a system are described by the authors

> - the opportunity for a sound engineer to bring along; e.g. into a broadcasting van, a high quality listening room as well as his familiar loudspeakers
> - to switch over and compare different listening situations that are stored in the database, e.g. a professional sound studio, film sound in a big cinema environment, reproduction equipment at a consumer home etc.
> - to compare different loudspeakers and – stereo formats, like new surround-sound standards
> - to allow music production under standardized conditions, e.g. a reference listening room with idealized loudspeakers which has been generated synthetically
>
> (Horbach, 1999: 9-10)

This implementation will be the main topic of discussion in the third chapter of the present work, which focuses on practical tools.

Head tracking systems can also be implemented for traditional stereo and multichannel loudspeaker reproduction improvements. Holman and Kyriakakis (1998) described a video based head tracking system for such purpose. They argue that 'without head tracking and dynamic adjustment of filters parameters in response to head movement, the functionality of such systems [3D binaural] is extremely limited'. They list the main problems found on 5.1 reproduction in home environments, such as the placement of the loudspeakers and the position of the listener in relation to the system. Reproduction of 5.1 material over 2.1 systems requires binaural processing and crosstalk cancellation which still present problems related to the positioning of the listener in relation to the loudspeakers and they argue that head tracking can be used to solve these limitations.

They quote the developments reported at the Laboratory of Computational and Biological Vision at the USC Integrated Media Systems Center (IMSC), in 'a vision architecture that is capable of recognizing the identity, spatial position (pose), facial expression, gesture identification, and movement of a human subject, in real time' and argue that this structure integrates various visual cues that allows researchers to identify the location and orientation of a listener's head (Holman and Kyriakakis, 1998: 5).

In the description of how the system picks up images from a webcam and extracts orientation and positioning data, starting by finding the listener's silhouette through the performance of motion detection (assuming the camera is fixed). An algorithm detects pixel disparity inside the region defined by the silhouette and other silhouettes are established by analysing the regions that are moving into the main silhouette. Two different algorithms detect skin tones colours and convex regions that are clustered and processed to define a centre of the head that is then converted to trajectories. A data base of possible trajectories is accessed to estimate possible trajectories and make the algorithm practical for desktop audio applications that consider periods when the head is not moving

and the appearance of a new head. The biggest advantage of a video based head tracking system is that it does not need additional hardware and can offer other parameters such as face and expression recognition and ear classification that can be used for developing a pinna pictures database with corresponding HRTFs that allow listeners to use a HRTF similar to their own by using image analysis. In their experiment the head tracker designed operated at 30 frames per second with an error of ±0.5cm and was experimented with an Ambiophonic system, trying to place a virtual loudspeaker between the two frontal ones by ITD manipulation (varying between 0μs to 340μs).

Another description to how to process webcam images for head tracking is presented by Murphy-Chutorian and Trivedi (2009) that starts by defining the concept of 'head pose estimation' as the 'process of inferring the orientation of a human head from digital imagery. They affirm that human head is limited to three degrees of freedom named yaw, roll and pitch.

Although one may argue that, for example, to know the direction a person is looking at, one can measure the white area sclera that surrounds the eye, the authors report that observations based on gaze estimation analysis is extremely dependent on head orientation and argue that 'an eye tracker should be supplemented with a head pose estimation system' (Murphy-Chutorian and Trivedi, 2009: 608). They also describe the various possibilities of analysing head pose in communications and list the variety of information one can have from this process.

According to the authors, head pose estimation in computer science can fill in a huge gap of communication between humans and computers, and describe eight particular methods to access head pose arranged by 'the fundamental approach that underlies its implementation':

> Appearance template methods: compare a new image of a head to a set of exemplars (each labelled with a discrete pose) in order to find the most similar view.
>
> Detector array methods: train a series of head detectors each attuned to a specific pose and assign a

discrete pose to the detector with the greatest support.

Nonlinear regression methods: use nonlinear regression tools to develop a functional mapping from the image or feature data to a head pose measurement.

Manifold embedding methods: seek low-dimensional manifolds that model the continuous variation in head pose. New images can be embedded into these manifolds and then used for embedded template matching or regression.

Flexible models: fit a non-rigid model to the facial structure of each individual in the image plane. Head pose is estimated from feature-level comparisons or from the instantiation of the model parameters.

Geometric methods: use the location of features such as the eyes, mouth, and nose tip to determine pose from their relative configuration.

Tracking methods: recover the global pose change of the head from the observed movement between video frames.

Hybrid methods: combine one or more of these aforementioned methods to overcome the limitations inherent in any single approach.

(Murphy-Chutorian and Trivedi, 2009: 609)

After describing each one of these methods, the authors present how to get the data from real head pose to computer and how to create a data base. They argue that visual tracking present fewer errors than other methods of head tracking and conclude that

> Head pose estimation is a natural step for bridging the information gap between people and computers. This fundamental human ability provides rich information about the intent, motivation, and attention of people in the world. By simulating this skill, systems can be created that can better interact with people.
>
> (Murphy-Chutorian and Trivedi, 2009: 620)

They also argue that in the future head pose estimation systems need to be accurate (with 5 degree errors or less), monocular (using a single camera), autonomous (no manual activation or configuration), multiperson (estimate similarly for different people), identity and lighting invariant, resolution dependent, allow full range of head motion and work in real time.

1.4 Some other past performed listening tests

Lots of tests were already performed aiming understanding our hearing perception as well as the working of complex systems through binaural or Ambisonic technologies. An important research for future chapters' development is presented by Nettingsmeier (2010), who presented some considerations on using B-format Ambisonic signals as intermediary to obtain 5.1 material both for music program and film. In his listening tests he argues that the use of Ambisonic systems is possible but affect sound engineers decisions and reports that the unstable virtual loudspeaker image generated mainly for the central channel in 5.1 film material was unacceptable being 'reported localisation ambiguities and coloration'. On the other hand, for electroacoustic composers and students and their purpose of reproducing quadriphonic compositions, the system worked quite well, what means they did not perceive large differences between the two reproductions, quadraphonic and virtual loudspeakers through Ambisonics. He highlights that opinions vary a lot.

Another example of binaural reproduction to evaluate an Ambisonic system is presented by Boland et al. 2010. Aiming to evaluate the coherence between Ambisonic depth reproduction and real reproduction, their tests were performed using virtual loudspeakers reproduction over headphones. 'Results demonstrate that first order soundfields are sufficient in representing distance cues for virtual loudspeakers reproduction' (Boland et al., 2010: 1).

They also highlights that 'as the order of sound field reproduction gets higher, the localization accuracy increases due to greater directional resolution' and that the perception of distance in such systems are not only due to reverberant to direct sound ratio but also due to directional reproduction of early reflections. They believe that the higher the order the better is the perception of depth.

For binaural reproduction they implemented a head tracking system with Intertia Cube sensors, HRTFs from IRCAM 'LISTEN' database (subject 1021) and manipulation of the

soundfield in the Ambisonic domain. The author reports ‘stable virtual images with head movements’ even with the use of non-individualized HRTFs and an ‘extremely effective’ externalization with no front-back confusion. A lack of accuracy was perceived when determining distance beyond 4m. They conclude that ‘Ambisonic reproduction matches the perceived real world source distances well at each order. No significant statistical difference was exhibited by increasing the Ambisonic order in this regard’ (Boland et al., 2010: 5).

Demonstrations have also been performed using binaural technologies, for instance Karamustafaoglu and Spikofski (2001) at the 19th AES convention in Bavaria (Germany) made a demonstration of a surround control room through headphones with head tracking system incorporated that offered the listeners the possibility of one to one comparison between real and measured environments.

Some other authors use binaural representation of complex loudspeaker systems to perform their experiments and an example of that is reported by Lindau et al. 2007. They first describe an automated head and torso simulator for binaural reproduction with head tracking impulse response measurement system. A binaural representation of High Order Ambisonics and Wave Field Synthesis is used to perform the tests in a virtual recreated environment, allowing comparison between environments that are very different between each other.

A point needs to be observed in such experiment that is related to hearing perception studies is that, although they consider the influence of the torso, it is not clear if they move the whole torso and head or just the head in relation to the torso, a difference that can influence the final results of measurements or of any kind of experiments that uses dummy heads.

The huge amount of variables that our hearing ability involves can make any kind of experiments less than wholly trustworthy since none of them consider all these variables. For instance, implementations of head tracking can vary a lot as can HRTF databases or

even measured individualized ones. Reverberation, being extreme influential on sound localization perception, can present a whole set of differences related to sound modification due to head movements that are still largely unknown and hardly measured. Despite all the uncertainty of achieving a high fidelity system, a lot can already be done to get closer to the ideal. The next chapters will describe tools to do so as well as the development of a ‘virtual studio’ based on reachable tools already available.

Chapter 2 – Tools

In an informal talk with Hugh Robjohns, presenter of the tutorial about surround sound formats, at the 130th AES convention in London (13-16 May 2011), on the question of why Ambisonic encoding and decoding was not adopted 30 years ago, he answered that possibly it was because of the lack of tools to work with it and that nowadays it would be perfectly possible to adopt such a system if others (i.e. Dolby) was not so very well established in the industry.

Solutions of the problem of tools to work with Ambisonic systems were foreseen by Malham in 1987 while presenting ways of developing pan pots, rotators and dominance effect (zooming) for such systems. At that time, the nature of automated controls led to compromised solutions that exhibited granulation and a lack of resolution with the equipment available. These problems, according to the author, would be solved with full digital processing. In 1991, Malham and Orton presented a description of a digital system to perform Ambisonic manipulation using MIDI controllers. Nowadays all the processes in Ambisonic domain can easily be achieved by digital processing and many tools have been developed in the intervening years.

2.1 Hardware tools

Some hardware tools, like the Lake Huron Digital Audio Convolution Workstation, have already been used for spatial audio manipulation, for instance those described by McGrath and Reilly (1996). This system, based on DSP chips, is designed for applications in virtual reality, animated auralization and multi-speaker presentation, amongst others (Huron technical manual). Among the other tasks, the above mentioned system can encode and decode 1st order B-format signals, decode for binaural through headphones reproduction by HRTFs processing, simulate Doppler Effect for moving sources and provide delay

compensation for irregular loudspeaker arrays.

In the description presented by McGrath and Reilly, they use the Lake Huron system to record and playback soundfield recordings, manipulate and monitor them through loudspeakers or headphones, these with a head tracking system performed in the Ambisonic domain which rotates the soundfield. They highlight the four main reasons why B-format was adopted: simpler real time manipulation of movements, playback to various loudspeaker arrays, 'stable acoustic image' while playing through headphones with head-tracking implementation, and also the fact that it is 'useful as an intermediate format for recording'.

2.2 Sets of tools

More recent packages of tools have been developed aiming allowing composers, sound designers and sound engineers to accesses the most recent developments in spatializing sounds through VST plugins, MAX externals or PD objects, the most common platforms used for music production and composition.

Braasch et al. (2008) for instance, describe a whole toolbox for MAX/MSP and PD, named ViMiC (unfortunately available only for MAC users) that simulates different microphone positions in a virtual three dimensional environment and supports reproduction using up to 24 discrete loudspeakers. The toolbox 'follows somehow this Stockhausen's traditions by using the concept of spatially displaced microphones for the purposes of sound spatialization' (Braasch et al., 2008: 1). According to the authors,

> ViMiC is a computer-generated virtual environment, where gains and delays between a virtual source and virtual microphones are calculated according to their distances, the axis orientations of their microphone directivity patterns. Besides the direct sound component, a virtual microphone signal can also include early reflections and an adequate reverberation tail. Upon both the sound absorbing and reflecting properties of the virtual surfaces.
>
> (Braasch et al., 2008: 2)

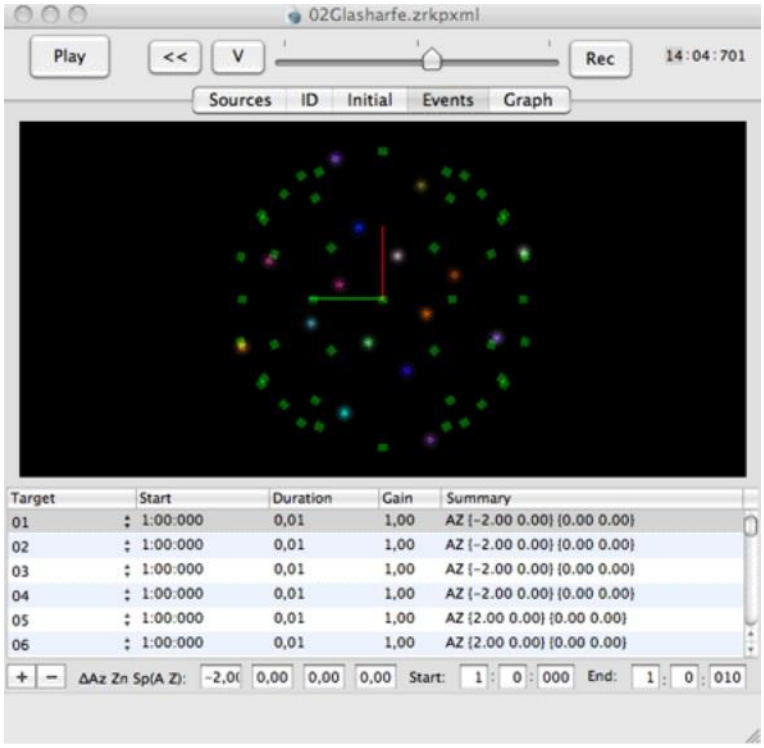
Directional properties of the sound source can also be simulated with frequency independence, as well as Doppler Effect for moving sources. In this implementation each microphone signal is associated to a loudspeaker output.

Another interesting package of tools is presented by Ramakrishnan and named Zirkonium (2009 - unfortunately available only for MAC users). This is described as a software that defines ‘a technique for routing sounds to the speakers, and a controller for the defining the routing’. From the above mentioned author’s previous experiences in electroacoustic music, there are two main approaches to spatialization: the first one, labelled the ‘acousmatic approach’ is characterized by the practice of playing composer’s two channel pieces in a bigger array of loudspeakers by taking advantages of, and creating new effects from, room acoustics, speaker placements and their characteristics; the second approach, labelled the ‘simulation approach’, is characterized by a reproduction that mimics sound sources movements using the whole system of loudspeakers. Parallel to this distinction between the ‘acousmatic approach’ and the ‘simulation approach’ there are cases where the composer wants the sounds exactly like they were when played back through the loudspeaker itself, other may tolerate some processing in order to generate virtual sound sources between the speakers and get a better illusion of movement.

The Zirkonium implementation, aiming to satisfy both approaches, is based on VBAP for panning sound sources, using Delaunay triangulation (Bern and Eppstein, 1992; Shewchuk, 1996) instead of Pulkki's proposal, but can also use Ambisonics B-format signals. This allows definition of positioned sound sources as well as the use of spreader tools, that ‘samples the specific range and creates extra virtual sound sources to distribute the sound over an arc, line or area’ (Ramakrishnan, 2009: 272), and simulation of different loudspeaker arrays to be listened through headphones (by the use of HRTFs).

The system itself stores sound sources positions as OSC messages which allows it to be recreated after. It can be used as a device and accessed through Jack Audio router or through any software like a DAW, MAX or PD. When reproducing quadraphonic pieces,

the system can be used to spatialize different copies of the quad material in four different squares that can be rotated, elevated and give more envelopment. It can also be controlled live.



User interface of the control software Zirkonium. Extracted from <http://www.zkm.de/zirkonium>.

Bresson and Schumacher (2010) also present a set of tools for spatialization developed for the Open Music environment. It allows the user to 'describe spatialization processes from an abstract level and render it into a multichannel audio format using one of the available spatialization techniques'. According to the authors it is easy to change between the techniques with no effect on the trajectory of the sound sources defined in the composition. The techniques available are stereo, quadraphonic, VBAP, HOA, reverberated VBAP and a mixed order Ambisonics with optional room simulation. The Ambisonics implementation includes a 'module to increase the impression of distance and motion of a sound source' (Bresson and Schumacher, 2010: 1), air absorption, Doppler Effect and decrease of amplitude due to distance increase. It is interesting to note that the

flexibility of Ambisonic encoding and its suitability to their approach, having the encoding process separated from the decoding, allow the composer to focus on spatialization without thinking about the loudspeaker configuration. They recommend users to take care about metadata information (since there is no standard format) in the encoding and decoding processes.

The system was developed to interact with OMPrisma and uses ICST tools to decode Ambisonics up to 32 channels. A binaural implementation based on spat~ from IRCAM is used in the absence of loudspeakers. They point out the need of implementing a webcam based head tracking system due to their availability on laptops without the need of dedicated hardware.

All these tools used for sound spatialization are actually derived from more basic ones that are going to be described in succeeding sections and that have been allowing composers to work with multichannel reproduction out of the standards defined by music and film industry.

2.3 Encoders and decoders

One of the first tools aiming to work with Ambisonic systems were designed and described by Gerzon in 1975, in his unpublished technology reports number 3. In this report he presents solutions for implementing 5 different pan pots for Ambisonic manipulation:

1-) horizontal only pan pot, that includes manipulation of the interior of the sound field.

2-) with height pan pot, also including interior manipulation.

3-) horizontal only pan pot with rotation and width control.

4-) horizontal and with height pan pot with rotation and width controls.

5-) hemisphere pan pot with three operational modes, one for horizontal panning and interior effects, other with height for an upper hemisphere and other for low hemisphere.

The author argues that

> Such an “encoding specification” consists of assigning each position in space to a set of gains on the channels used on the master mixdown. While the encoded material may finally reach the consumer 2-channel, 3-channel or 4-channel form, the master encoding specification is 4 channels in most cases.
>
> One method of correct encoding is to use a precisely coincident array of accurately defined directional microphones. Another method of correct encoding is to pan pot mono sources in to the sound field encoding by means of pan pots accurately following the chosen encoding specification.
>
> (Gerzon, 1975b: 3)

Gerzon also quotes some research where it was proved that a pairwise pan-pot, normally used for stereo reproduction, is not able to produce stable ‘phantom image’ between 2 loudspeakers 90 degrees separated from each other, so the need of an Ambisonic pan pot to perform positioning a sound source in a multichannel system such as the quadriphonic one or any other surround configuration. A very interesting characteristic of his work is that, while describing Ambisonic systems, he is aware of the fact that the system was not fully commercial compatible yet:

> Since the fourth channel is available in studio equipment [see Barton and Gerzon, 1984], the best use that can be made of it is to store height information that would help to avoid obsolescence of recordings when height reproduction becomes commercial.
>
> (Gerzon, 1975b: 3)

For MAX/MSP platform, a certain variety of encoders have been developed through the last ten years and available for musicians and composers as MAX externals. One of the first set of them was ported from Dave Malham’s VST plugins by Matthew Paradis. The encoding and decoding can be done only in first order and the decoder can be set for different loudspeakers arrays such as cubic, square, rectangular, hexagonal, two different octagonal setups, stereo and mono. It has a version for both MAC and Windows and can be downloaded from the link <http://www.york.ac.uk/music/mrc/software/objects/>. A version for PD can also be found in the same link. The whole set consist of the following objects:

ambimic~ - ambisonic decoder for stereo reproduction

amicctrl - offers controls of elevation, azimuth, angle and pattern of the virtual stereo pair

ambi2s~ - decoder of 1st order ambisonic to 6.1 surround system array

a2sctrl - offers controls of front and back angles and pattern, as well as loudspeakers gain levels

ambidec~ - ambisonic decoder for different loudspeaker arrays configurations

ambipan~ - panner for mono sources to B-format

apanctrl - offers controls of azimuth, distance, elevation, centre, zeroth, first, distance factor and volume for both left and right inputs

Another set of Ambisonics externals for MAX was developed in 2006 by Graham Wakefield and are detailed described in Wakefield (2006). They were designed primarily to be used in compositions for the Allosphere in the UCSB (University of California Santa Barbara). Based on Jerome Daniel's PhD work and developed by Graham Wakefield, Jorge Castellanos and Florian Hollerweger using the CREATE Signal Library (http://fastlabinc.com/CSL/index.html), the signals are labelled according to Daniel's convention (Daniel, 2000) and the user has the option of using an object to encode a single mono source or another one that is able to encode up to 16 signals. The encoding and decoding can be done in two or three dimensions and up to third order.

The author suggests that decoding for 5.1 can be obtained by decoding the Ambisonic signal to five loudspeakers in a horizontal plane and the LFE can be obtained by filtering the W channel. In his conclusion the author argues that 'A HRTF decoder for headphone listening would be a welcome addition' (Wakefield, 2006: 126). The whole set of externals can be downloaded at <http://www.grahamwakefield.net/soft/ambi~/index.htm> from a file named ambi_win.zip (windows version, also available for MAC), and consist of the following:

ambi.encode~ - encodes a monophonic source to a specified azimuth and elevation (optionally sample accurate or interpolated).

ambi.encoden~ - encodes up to 16 sources to distinct azimuth and elevation orientations

(specified individually or as a list).

ambi.decode~ - decodes an ambisonic encoded sound field to a user-defined speaker array of up to 16 channels (more can be added by using more than one ambi.decode~ object). Messages control the speaker layout, global gain, mono/spatialized balance, and decoding order weights.

The ICST tools for Ambisonic spatialization in MAX/MSP were described in Kocher and Schacher (2006), where the authors highlighted that

> To play back the encoded format a recomposition is made taking into account the exact location of each speaker. In theory the number of speakers and their position can be freely configured, but practical experience shows that symmetrical setups using at least as many speakers as there are components in the B-format are preferable.
>
> (Kocher and Schacher, 2006:1)

In another paper with the historical background of the development of such tools, Schacher acknowledges the help of Dave Malham and reported that the first objects (first order encoder and decoder) were written in C between 2002 and 2003. The implementation of distance attenuation follows descriptions made by Malham, where closer to the origin the closer to a monophonic signal (W) the sound is. In 2007, an attempt to connect the MAX/MSP externals to a DAW through Pluggo was made and in the same year Neukom developed the ambipanning object that approximates the Ambisonic panning to VBAP and DBAP, making a panning without a conversion to B-format in higher orders possible. The author also reports the expansion of the usage of Ambisonic tools into commercial music production.

The ICST Ambisonic tools, in its version 2.0 beta 9, both for windows and MAC, can be downloaded from the link <http://www.icst.net/research/downloads/ambisonics-externals-for-maxmsp>. It is constituted of the following objects:

ambidecode~ - decode Ambisonics between 1 to 11 order for different loudspeakers array and variable number of outputs (limited to 250 by MAX)

ambiencode~ - encode a mono source to Ambisonics B-format between 1st to 11th order,

(inputs are limited to 250)

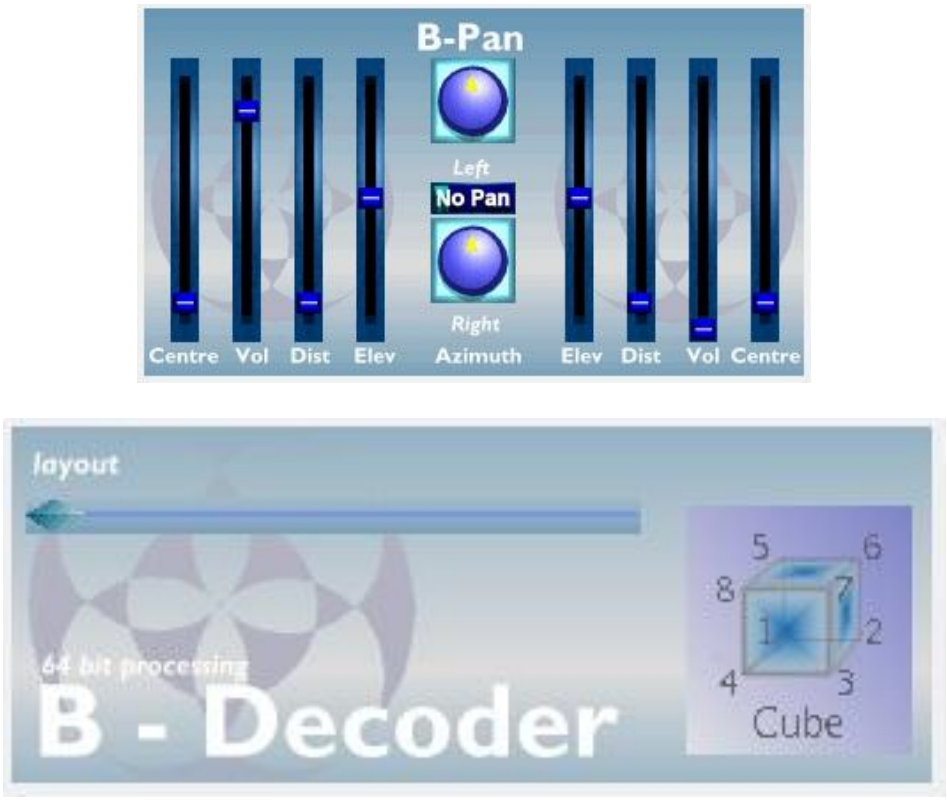
ambipanning~ - Ambisonic equivalent panning developed by Neukon

Although Malham's external allows a big variety of processes, as will be described next, a noticeable advantage of Wakefield's externals as well as the ICST tools is that the loudspeaker array is defined by the user and can even be asymmetrical. A major advantage of the ICST set over the other objects is the high order it can reach and the easy way of configuring the loudspeaker layout through messages. It is worth saying that choosing a set of externals to work with is dependent on the tasks to be performed and if one uses an encoder of one of these sets, it is not necessarily true that using the decoder of another set will be straight forward. Some differences can be perceived between these tools mainly related to the order of the encoder's outputs (W, X, Y, Z, …) as well as in the inputs of the decoders and processors. The messages to be sent defining elevation, azimuth, etc. also need some attention while patching in MAX.

Tools for MAX/MSP that uses VBAP panning have also been developed and, as described in the first chapter, can be used in association with Ambisonic panning. Pulkki described a MAX object in 2000 that allows the user to pan mono sound sources in a loudspeaker array. Sound source spreading can also be performed. The loudspeakers' positions are defined by the user but the distances from the listening point should be the same or be delayed to adjust for time of arrival. This object can be downloaded from the link <http://www.acoustics.hut.fi/software/vbap/MAX_MSP/>.

Since most of people in music production are using Digital Audio Workstations that take advantage, among other characteristics, of the use of various plugins, some developers have been working on Ambisonic encoders and decoders to be made available as VST plugins. One of the first of many was developed by Dave Malham and Ambrose Field with sponsorship of the Hochschule fur Musik Winthertur Zurich and the Swiss Center for Computer Music. It can be downloaded at <http://www.dmalham.freeserve.co.uk/vst_ambisonics.html> and offer the user to pan

mono or stereo sources in first order Ambisonics in the encoding process and to decode it to different loudspeaker layouts, for instance cubic, square, rectangular, hexagonal, two different octagonal setups, stereo and mono.



Ambrose and Malham's first VST encoder and decoder.

To be officially released but already available for Music Research Centre students, a third order B-format panner (or encoder) has been developed by Dave Malham, as well as a third order decoder to operate in the loudspeakers arrays mounted on Trevor Jones Studio and in the Rymer Auditorium, both in the Music Research Centre in the Univerity of York.

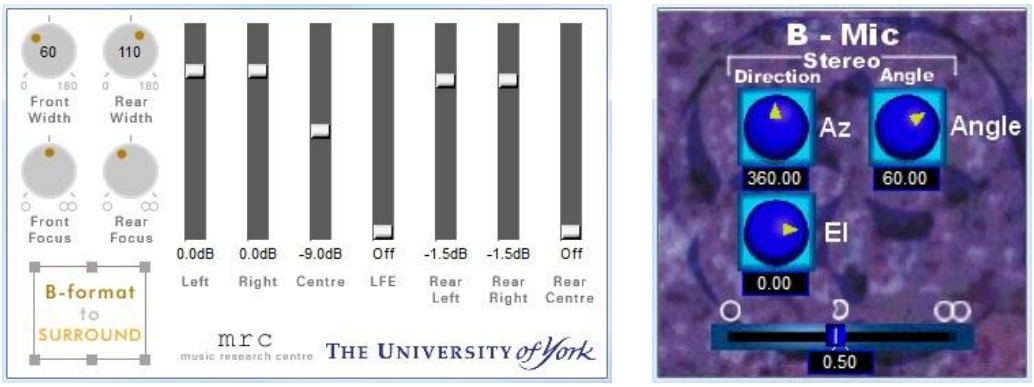
Malham's third order VST plugin encoder.

Malham's Trevor Jones decoder and the FlexDec full horizontal with height decoder.

The above mentioned encoders are supposed to receive a single mono or stereo signal that will be panned into four (1st order) or eight (2nd and 3rd mixed order) B-format signals. Users have access to manipulate azimuth and elevation of the sound source, as well as a distance parameter from the centre (also changeable). The decoders receive those signals, normally mixed in the DAW, and decode them to the specific layout of loudspeakers, already pre-defined in the case of the Trevor Jones decoder, or to be defined by the user from a list of most common regular arrays. The loudspeaker configurations are supposed to be symmetrical and equally distanced from the origin of a sphere (listening point). In the late decoders a test signal is available for calibration of the loudspeakers.

Other decoders for specific arrays such as surround 6.1 and stereo have also been developed by Padraig Kitterick at the MRC in 2004. The B-format to Surround plugin, BF2SGui, for instance, offers the user individual gain control for each loudspeaker as well

as full control of the angles between the front and rear loudspeaker positions and their focus, e. g. how much spread the sound of the front and rear is presented (variable between omnidirectional and figure of eight). The B-mic VST plugin decodes B-format signals by simulating a coincident stereo pair which the user controls both the azimuth and the elevation the pair is pointing to, as well as the angle between the capsules and the polar pattern they present.



Malham's B-format to Surround and B-mic decoders VST plugins.

Other available VST encoders and decoders were developed by Bruce Wiggins, from the University of Derby. From his webpage one can download encoders, decoders and an Ambisonic freeverb (<http://www.brucewiggins.co.uk/?page_id=78>). One of the encoders can perform 1st order Ambisonic panning through XY coordinates (two dimensional only) and has a NFC (near filed compensation) control. The second, also encoding to 1st order B-format, is based on polar coordinates with height and NFC as well, the central control perform distance from the centre adjustments.

Wigware 1st order Ambisonic encoders VST plugins.

Bruce Wiggins 1st order decoders are distributed in two versions, the first one for symmetrical loudspeaker configurations (square, hexagon, octagon, 8 speaker cube and octahedron) with or without CF (centre front loudspeaker). The second one is based on ITU surround sound 5.0 decoding with seven different algorithms (Berry et al.; Wiggins, 2007). Both of them present compensations for microphone distance and speakers distance, as well as low and high frequencies polar pattern adjusts of the simulated virtual microphones.

Bruce Wiggins 1st order decoders VST plugins.

Other tools for encoding and decoding Ambisonic signals are described in gerzonic.net webpage. The Gerzonic Panorama plugin is a 2nd order encoder that allows user to configure azimuth, elevation, distance and a distance scale (similar to that presented in Malham's plugins).

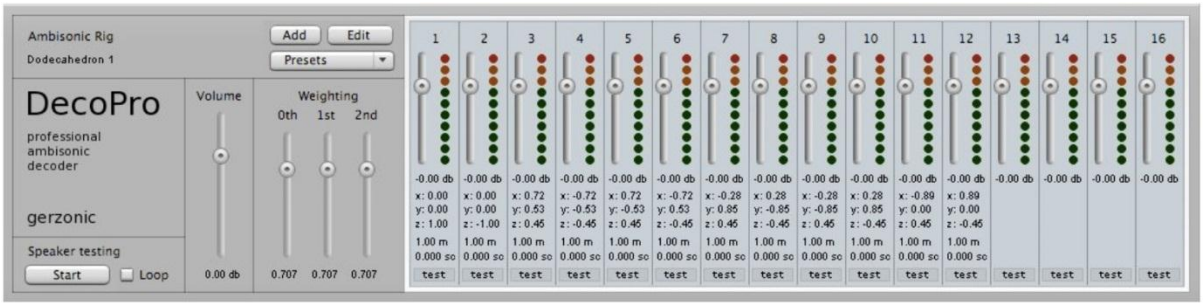
Gerzonic Panorama VST plugin.

The Gerzonic Emigrator is a first order decoder for stereo, square, pentagon, hexagon, octagon, surround 5.0, cubic and dodecahedron loudspeaker layouts. It also presents a strict soundfield vs. controlled opposites parameter (that control the in-phase effect), a master volume control, a parameter to force 1st order decoding and a speaker test signal.

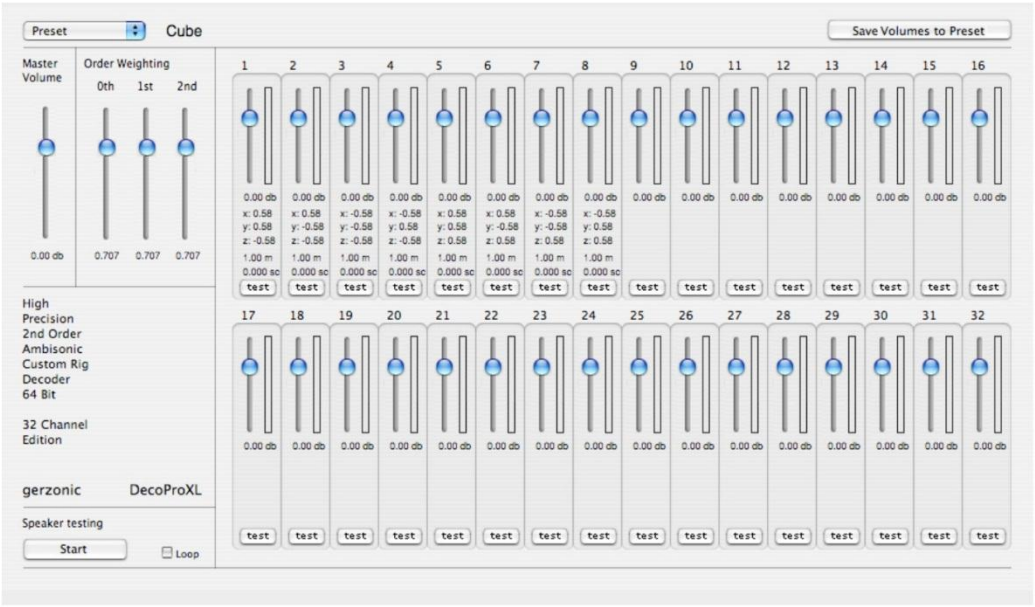
Gerzonic Emigrator decoder VST plugin.

Other decoders are described at the gerzonic.net webpage and named DecoPro, DecoPro XL and DecoPro XXL. These decoders can be configured by the user to decode Ambisonic B-format signal up to 2nd order for different loudspeaker arrays, defined through individual

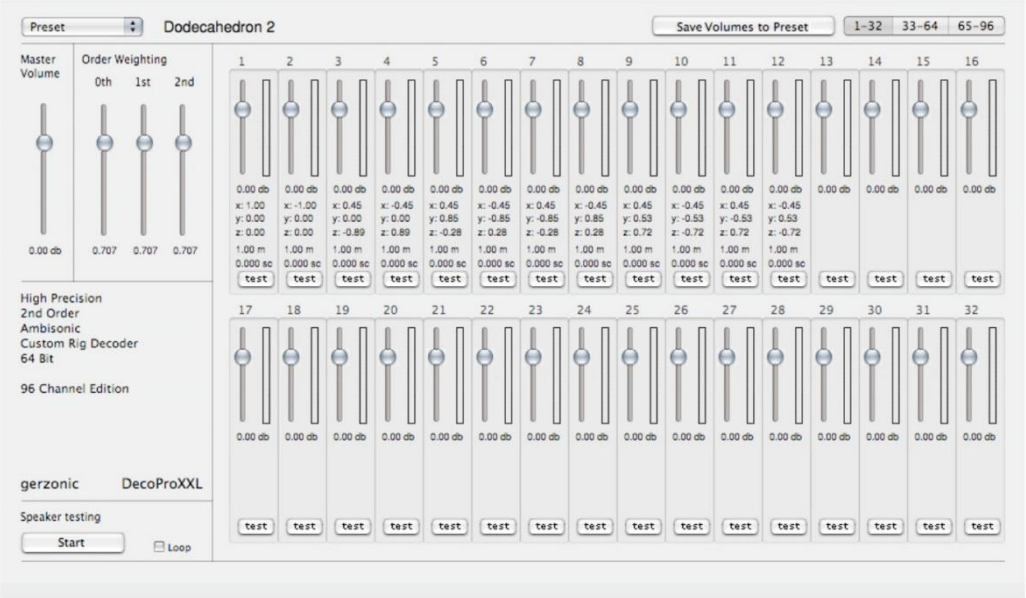
X, Y, Z and distance coordinates. The first version can decode up to 16 loudspeakers, the second up to 32 and the third up to 96 loudspeakers.



Gerzonic DecoPro decoder VST plugin. Extracted from gerzonic.net website.



Gerzonic DecoProXL decoder VST plugin. Extracted from gerzonic.net website.



Gerzonic DecoProXXL decoder VST plugin. Extracted from gerzonic.net website.

All these VST plugins described so far have versions for both Windows and MAC, which makes them very useful for any project since it is independent on the operating system. Another very well-known set of plugins were developed by Daniel Courville between 2007 and 2010, and provides 1st, 2nd, 3th, and 5th order mono and stereo encoders, quad, 5.1 surround, double MS and zoom signals transcoders, A-format to B-format converter, decoders for particular loudspeaker positions (up to 12 speakers), 5.1 surround and UHJ stereo. However, these are available only for MAC users.

Some of Daniel Courville's VST plugins. Derived from <http://www.radio.uqam.ca/ambisonic/>

One of the most recent releases on Ambisonic manipulation was developed from the work of Barrett and Berge (2010) and has as outcome a VST plugin (released on 31 March 2011) that can decode 1st order Ambisonic for multiple loudspeakers arrays (up to eight) as well as for 5.1 surround with better channel separation than previous attempts, and emulate non-coincident stereo techniques (ORTF and AB). The Harpex-B plugin can also perform some processing such as zooming, rotating, tilting and tumbling, adjust of direct and reverberant sound, and simulate binaural reproduction over headphones.

Harpex-B VST Plugin. Extracted from <http://www.harpex.net/index.html>.

The Soundfield Company ships a decoder with the SPS200 A-format Soundfield microphone. It allows the user to decode from A-format to B-format, stereo, octagonal array, 5.1, 6.1 and 7.1 configurations. Among other controls the user has access to controls for individual loudspeaker volume, as well as rotation, tilt and tumble manipulations, front and rear width angles and the polar pattern of the rear channels.

SPS200 Surround Zone decoder VST plugin. Extracted from

< http://www.soundfield.com/products/sps200_s_zone.php >.

2.4 Ambisonic processors

Following the development of encoders and decoders, due to the easy way the Ambisonic audio material can be manipulated, most of the developers make also available some processors that can rotate, tilt or tumble the whole soundfield, mirror it, zoom it or spread it in a particular direction or region of the soundfield.

The first processors seem to have been designed by Gerzon in 1975. These are described in detail in his unpublished Ambisonic technology report n°4, where he also describes artificial reverberation and 'spreaders' for Ambisonic systems. According to him, 'besides the ability to handle sharply localised direct sounds, any well-designed system of surround sound must also be able to handle diffuse sounds, such as reverberation' (Gerzon, 1975c: 1).

For MAX/MSP platform, processors have been developed and distributed together with the encoders and decoders. The set of externals described in the last section, ported from Dave Malham's VST plugins by Matthew Paradis also offers first order processing such as rotating, tumbling, tilting, mirroring and zooming, important compositional tools that go beyond just encoding and decoding spatialized sound sources.

ambiplane~ - process the B-format input in horizontal plane outputs B-format processed

aplanectrl - offers controls of azimuth, elevation and mirror (to be used with ambiplane~)

ambiproc~ - process the B-format three dimensionally and outputs B-format processed

aprocctrl - offer controls of rotate, tilt and tumble (to be used with ambiproc~)

ambizoom~ - process the B-format three dimensionally and outputs B-format processed

azoomctrl - offer controls of azimuth, elevation and zoom (to be used with ambizoom~)

The set of externals developed by Graham Wakefield's also offers specific objects to perform granulation, rotation, mirror and weighting, in a first order B-format signal.

ambi.granulate~ - (beta version) granulates an incoming signal into an Ambisonic sound field.

ambi.rotate~ - transforms an encoded sound field by rotating around axes.

ambi.mirror~ - transforms an encoded sound field by mirroring around axes.

ambi.weight~ - balances the components of an Ambisonic encoded soundfield per order, using a set of pre-defined or user-defined weights.

The ICST tools described by Kocher and Schacher brought, in 2006, the development of ambimonitor and ambicontrol objects, which allow the composer to perform panning in an intuitive way as well as visualize the panned sound sources. These developments were performed with the help of the composer Philippe Kocher and is also when they decided to use the same coordinate system adopted by IRCAM's spat~ object (to be described later).

ambicontrol - to perform trajectories and send it to ambimonitor or to ambiencode as aed (azimuth, elevation and distance) coordinates

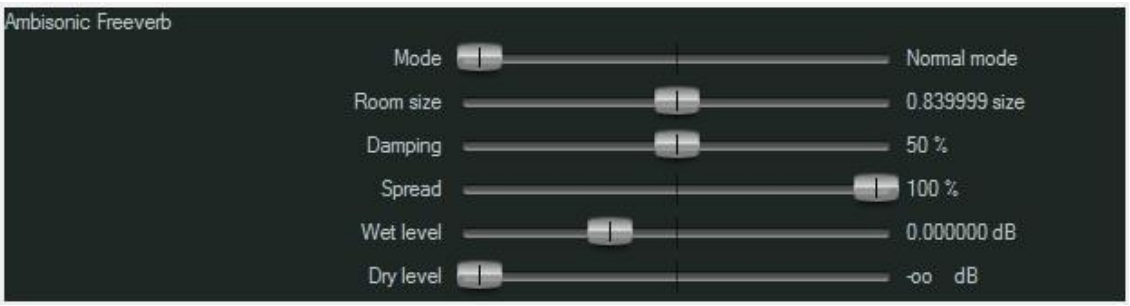
ambimonitor - for monitoring inputs (sound sources) or outputs (loudspeakers) positions

Processors are also available as VST plugins. Dave Malham for instance developed three of these processors to act on first order B-format signals. These are B-zoom, that allows approximation and estrangement from a sound source, B-proc, that allows rotating, tilting and tumbling manipulations on the soundfield, and B-plane mirror, that allows the user obtain a mirrored image of the panned sound source at another point in space, controlled by azimuth and elevation as well.

Dave Malham's B-zoom, B-proc and B-plane mirror VST plugins.

Another very important tool for those who work with music and film production is a

reverb processor. For now, the option available (as far as the present author's knowledge goes) as VST plugin was developed by Bruce Wiggins and is based on the Freeverb processor, allowing the user to switch between a normal mode and a freeze mode, manipulate dry and wet parameters as well as damping, room size and spread. Other reverber processors can be achieved by convolution, as mentioned earlier, with any two units of stereo convolvers, such as SIR (to be described later), where four impulse responses obtained from a Soundfield microphone signal are convolved with the B-format signal coming from an Ambisonic panner.



Bruce Wiggin's Ambisonic Freeverb VST plugin.

Another set of very useful plugins for manipulating Ambisonic files is presented at gerzonic.net. An AB switcher as well as a B-format player (bPlayer) and recorders (bRec) are available. The recorders can generate first or second order B-format, 16 or 32 bit files.

AB and bPlayer VST plugins.

bRec VST plugins (1st and 2nd order, 16 and 32 bits).

Daniel Courville also present a set of processors, unfortunately only for MAC users, as already mentioned in the encoders and decoders section. He developed a level and delay 1st order B-format corrector, B-format processors (rotation, elevation, tilt and axis flip), B-format reverberator and both 1st and 3th order B-format eight channel mixers. Interest in developing other processors such as delay, convolution reverb and granulators has already been mentioned on the Sursound mail list by several members. Some of them are already working on it, since it seems very simple to work with spatial effects in Ambisonic domain where the panning system is basically defined by numerical coordinates.

2.5 Binaural processors

Binaural technology is much less diffused than other since it is mainly developed to perform well over headphones and is not suitable for large area performances to mention one of the limitations discussed in the previous chapter. Besides the disadvantages, one of the reasons it may be not very diffused is the fact that it is very cpu intensive when done in real time and only a few tools are available for working with it.

As binaural processors here we deal mainly with algorithms that pick up a mono source and convolve it with a particular HRTF for a particular point in space. They also have been

presented as objects for both MAX and PD but only a few as VST plugins.

An objects to perform such task in MAX/MSP is the ep.binSpat~, available both for MAC and PC, that allows the user to position a mono source in horizontal plane (no elevation implemented) by the convolution of the mono signal source with a set of interpolated HRTFs. Optionally there is a patch developed at the University of California and named binaural panner (<http://www.ece.ucdavis.edu/binaural/binaural_tools.html?>), where the user can try out four different HRTFs, open mono files, pan them binaurally controlling azimuth and elevation, and record the output.

A more complex binaural processor described by Jot and Warusfel (1995), spat~, was designed specifically for musicians and sound designers users and is the main outcome of the Spatialisateur project. Basically 'the spat~ processor receives sounds from instrumental or synthetic sources (assumed to be devoid of reverberation), adds spatialization effects in real time, and outputs signals for reproduction on an electroacoustic system (loudspeakers or headphones)'. Available for MAC and PC users the set of objects have to be paid and consists of:

source~ - intends to reproduce pre delay and doppler effect to a sound source

room~ - add artificial reverberation

pan~ - 'directional distribution of primary signals and reverberation signals'

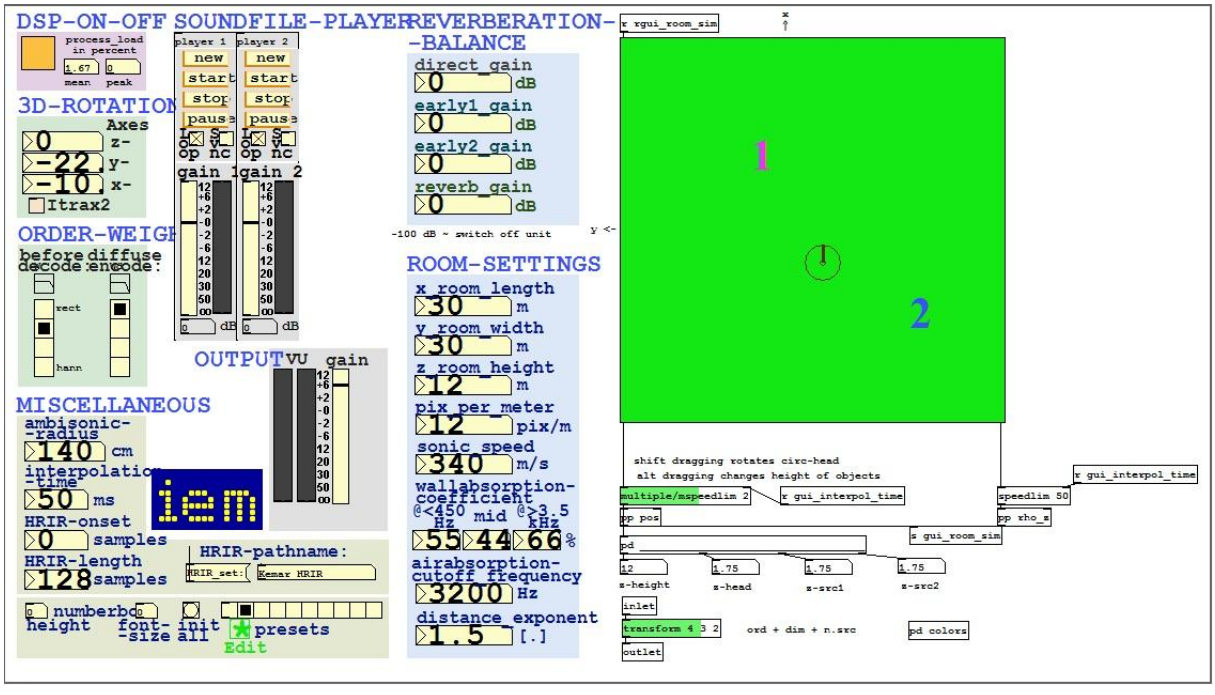
out~ - 'equalization of the output signals' to adapt to a particular situation headphones equalization or loudspeakers equalization

An interesting point on this set of tools is that the developers adopted a terminology that is closer to composers and musicians instead of using the technological vocabulary for spatial parameters. Another property to be noticed is that the binaural panner included in this set can perform the interpolation of the HRTFs for moving sources in a very simple way and will be very useful for the future development of the 'virtual studio'. In another work (Jot et al. 1995) the spat~ object is also used to perform transaural stereophony.

Another set of MAX object available to perform binaural rendering from Ambionic B-

format or 5.1 (Ambi2Bin~ and 51toBin~ - MAC only) is presented at the ZLB webpage and requires the user to decode the B-format signal to 7 or 5 channels before performing the rendering. It uses the concept of virtual loudspeakers (to be described in chapter 3) and the author suggests the use of Jack Audio router to link audio applications (topic of future section).

A Binaural Ambisonic Room Simulator is presented as a PD object, for both MAC and Windows, by Thomas Musil at the Institut für Elektronische Musik und Akustik and can be downloaded at <http://iem.at/Members/noisternig/bin_ambi>. The patch simulates a simple room with a listener and 2 sound sources with 24 early reflections each. The early reflections are encoded in 4th order Ambisonic. The listener can rotate the soundfield which is reproduced by 32 virtual loudspeakers filtered by corresponding HRTFs. In the patch, the two sound sources as well as the listener's position can be moved in X, Y and Z (height) dimensions.



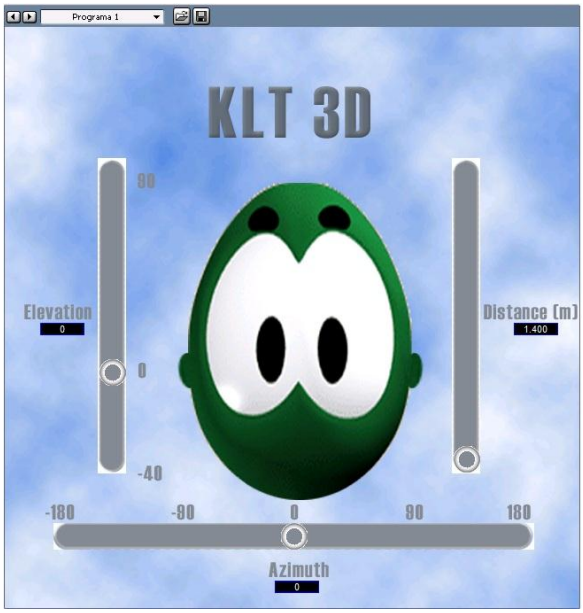
Ambisonic binaural room simulator screenshot.

Some HRTFs databases, needed for convolving the sound source signal and obtaining a

binaural reproduction, are freely downloadable, such as the 'LISTEN' HRTF database from IRCAM and the set measured at the Institut für Technische Akustik at Technisch Universität Berlin. This last one was recently described in Geier et al. 2011 and offers 1 degree steps horizontal only with measurements for four distances (0.5, 1, 2 and 3 meters) as well as IR for a few headphone compensation (including AKG K271mkII, very important for the present work).

Commercial processors are also becoming more available. These are mainly for the games and film industries. An example of such processors is the Rapture 3D software, manufactured by Blue ripple sound and designed for virtual reality applications, mainly games and surround sound playback. It is based on HOA up to fourth order. Binaural rendering is available with 5 different HRTFs. Rotation and other operations, controlled by visual content, are performed in the Ambisonic domain.

VST binaural panners can also be found. A free one, KLT3D, is distributed by the UFRJ Audio Processing Group. It offers the users the control of azimuth, elevation and distance of a mono sound source to be reproduced through headphones. A VST to convert 1st order B-format signals to binaural reproduction over headphones (Bformat2Binaural) was developed by Digenis and can be downloaded at <http://www.kvraudio.com/get/1685.html>. It decodes the four B-format signals to a horizontal square loudspeaker configuration and convolve the loudspeaker feed signals with correspondent HRTFs to 0, 90, 180 and 270 degrees. According to the author, this method requires less cpu since 0 degrees and 180 degrees HRTF usage avoid two convolutions with the disadvantage that the soundfield reproduced is not as clear as if the loudspeakers were positioned at 45, 135, 225 and 315 degrees. The two above mentioned plugins are unfortunately available only for Windows users.



KLT3D VST plugins interface. Extracted from <http://gpa.lps.ufrj.br/index.php/pt_BR/KLT3D-VST>.

The already described Harpex-B VST plugin is another of these few VST plugins that provides a multichannel reproduction over headphones using binaural technology. In addition to the possibilities of manipulation of the B-format signal described in the previous section, it can render a binaural signal for monitoring with the possibility of selecting the set of HRTF to be used plus a choice of four different equalization methods (diffuse, horizontal, front and none).

2.6 Head-tracking systems

The head tracking system has been described as an important tool to simulate the effect of the listener's head movement. It can perform manipulation in two different domains: the Ambisonic domain, controlling parameters that rotate, tilt or tumble the whole soundfield; and in the binaural domain, interpolating a set of HRTFs to be convolved with the sound source according to the listener position.

One of the first tools to be used is mentioned on McGrath and Reilly's work from 1995. In their auralization processing with head tracking, they used a PC-card based magnetic

device with a transmitter and a receiver, named ‘InsideTRAK Polhemus’, that is reported to perform the task well in a limited range but with high accuracy. In their work from 1996 they reported that such a tool was used to perform manipulations on rotation parameters of the B-format Ambisonic signals. They also reported that their system, based on the Lake Huron processor uses calculate HRTFs, to be convolved to the sound source signal, as an IR. They also used binaural impulse responses that include room reverberation. The head-tracking system is attached to the headphone and updates the orientation of the listener (McGrath and Reilly, 1996). It seems that this device is not being marketed any more or has been replaced by the Polhemus Fastrack, as have been noted on the manufacturer website. This system has been used, for instance, in the experiments by Algazi et al. (2002) and Lindau et al (2007).

The above mentioned Polhemus Fastrack is an electromagnetic based motion tracking system and claim to be the ‘industry standard in motion tracking since 1994’. The device tracks X, Y and Z Cartesian coordinates as well as azimuth, elevation and roll orientations with small sensors, and presents near zero latency (actually 4ms). The update rate is 120Hz that is divided by the number of sensors (up to four), and the resolution orientation is claimed to be 0.25 degrees. Its range is limited to 1.5 meters but can be extended up to 4.6 meters and the whole system can be linked to a PC by USB interface (Polhemus website).

Polhemus Fastrack system. Extracted from <http://www.polhemus.com/?page=Motion_Fastrak>.

Another tracking system developed by the same company (Polhemus) was used by Brookes et al. (2007) in their experiments. The Polhemus Patriot is also an expandable electromagnetic based system that allows multiple sensors and interfaces with host computer through USB or RS-232 (serial standard). Provides dynamic and real-time measurements of position and orientation of its sensors with continuous, discretely or incrementally data update (60Hz). It comes already calibrated and does not require adjustment, presents low latency (less than 18.5ms), high stability and an orientation resolution of 0.00381 degrees. An interesting advantage of these electromagnetic head tracking systems is that they are not affected by signal blocking or interferences, as would be visual based systems, allowing free movement in their active region range.

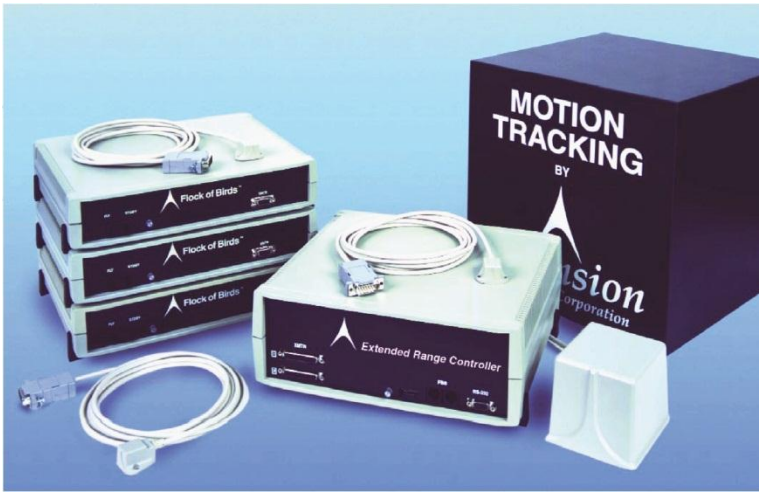
Polhemus Motion Patriot system. Extracted from <http://www.polhemus.com/?page=Motion_Patriot>.

Jin et al. (2009) describe the use of a head tracker named InertiaCube3, manufactured by InterSense. This electromagnetic sensor, also offered in wireless version, claims to be their smallest and most accurate orientation sensor. It presents an update rate of 180Hz, 4ms of latency and an accuracy of 1 degree yaw, 0.25 degrees roll and pitch. The sensor outputs accelerometer, gyro and magnetometer data and optionally can be connected to a host through USB. The same company also offers other two different sensors, the InertiaCube2+ that is just an orientation sensor with USB interface and presents almost the same specifications as the above mentioned sensor, and the InertiaCube BT that presents

similar specifications but is wireless, uses an integrated battery and interfaces through standard Bluetooth.

InertiaCube2+, InertiaCubeBT and InertiaCube3 electromagnetic sensors by InterSense. Extracted from <http://www.intersense.com/categories/18/>.

Murphy-Chutorian and Trivedi (2009) add to this list of electromagnetic sensors, the Ascension Flock of Birds, which is one of many sensors the Ascension Company developed. This sensor has a tracking range of almost 1m and 3m with the extended range transmitter, and an angular range of 180 degrees azimuth and 90 degrees elevation. The update rate is 144Hz and its interface with the host is made through RS-232. Other sensors are also distributed by the same company, such as the MotionStar (a whole body set of sensors dedicated mainly to 3D animation film) or the Hy-Bird, a hybrid optical/inertial sensor.



Ascension Flock of Birds sensors and transmitter. Extracted from <http://www.ascension-tech.com/docs/Flock_of_Birds.pdf>.

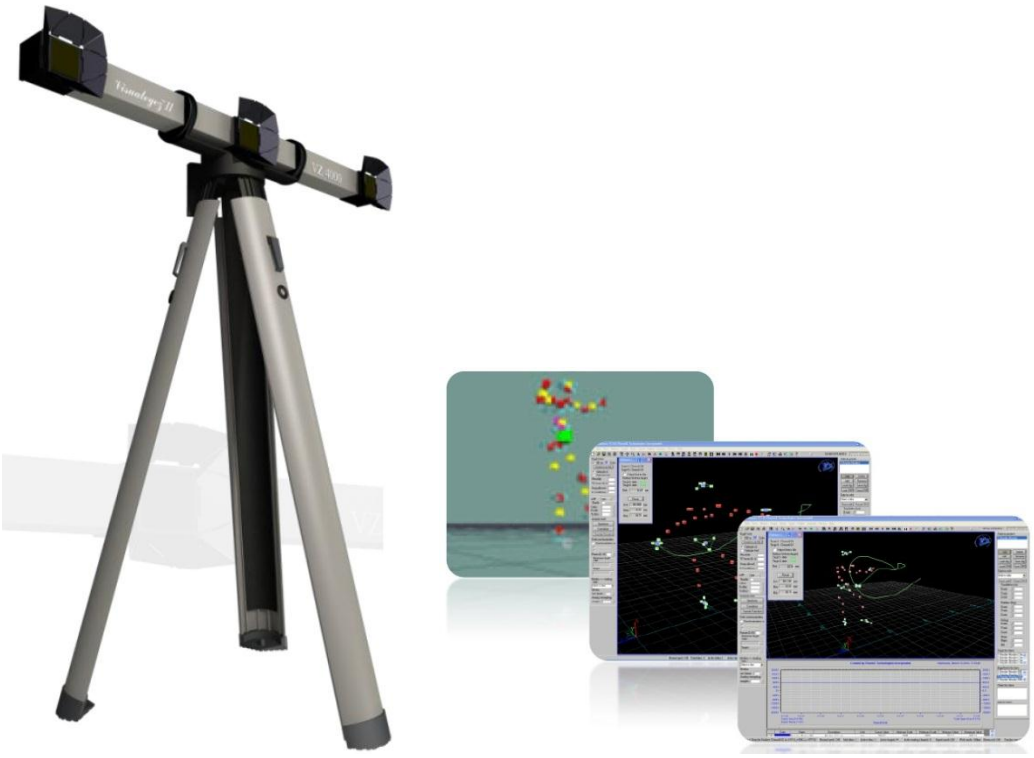
Other Ascension's sensors and transmitter. Extracted from <http://www.ascension-tech.com/medical/pdf/TrakStarSpecSheet.pdf>.

Ascention's MotionStar set of electromagnetic sensors. Extracted from <http://www.ascension-tech.com/docs/products/motionstar_10_04.pdf>.

Hybrid optical / inertial sensor. Extracted from <http://www.ascension-tech.com/docs/Hy-BIRD.pdf>.

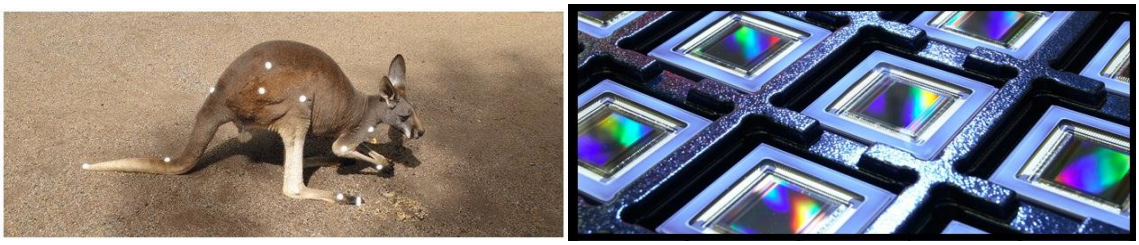
Two optical based head tracking systems are quoted by Murphy-Chutorian and Trivedi (2009): the Vicon systems and the Phoenix Technologies Visualeyez. Both these real time active-optical motion capture systems operate with specific cameras and tracking points attached to the body to be tracked. Dedicated hardware and software is needed.

Phoenix Visualeyez's markers. Extracted from <http://www.ptiphoenix.com/Accessories.php>.



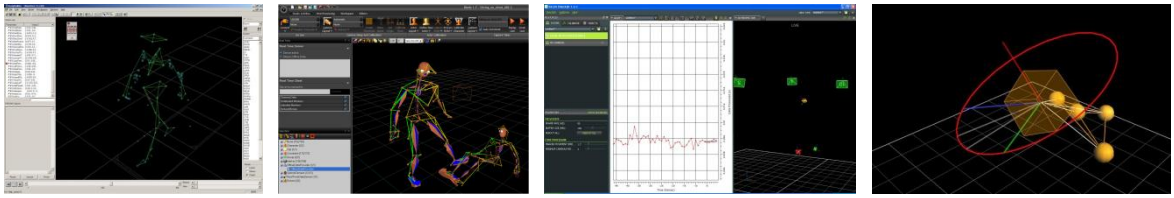
Phoenix Visualeyez's tracker and software. Extracted from

<http://www.ptiphoenix.com/images/Phoenix%20Technologies%20Brochure.pdf>.



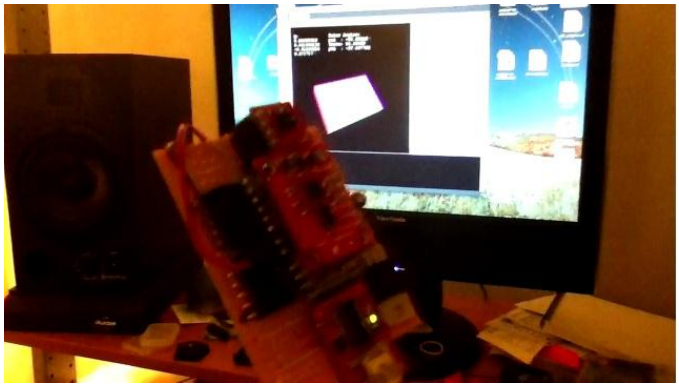
Vicon's markers on a Kangaroo and set of sensors. Extracted from <http://www.vicon.com/products/>.

Vicon cameras. Extracted from <http://www.vicon.com/products>.



Vicon softwares screenshots. Extracted from <http://www.vicon.com/products>.

Other interesting implementations mentioned on the SurSound email list use the wiimote or homemade sensors, which unfortunately also present the dependence of a physical device. Hector Centeno for instance presents a movement sensor device based on Arduino Pro Mini, IMU3000 Gyroscope, ADXL345 accelerometer and HMC5843 magnetometer. The software is a modified version of the FreeIMU library.



Centeno's head tracker device based on Arduino, to be mounted upon a headphone. Derived from <http://vimeo.com/22727528>.

Implementations based on the wiimote are reported by Ryan Pavlik, Johnny Chung Lee and Wouter Wognum.

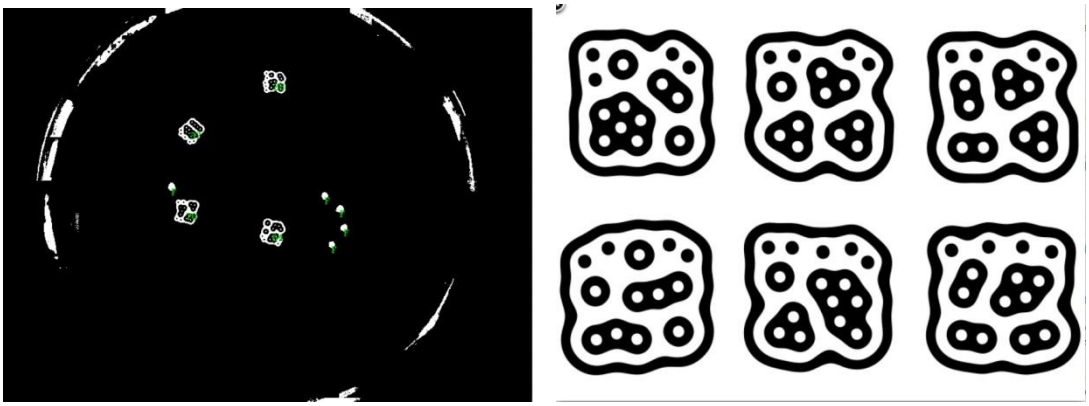
Implementations of head tracking using the wiimote by Ryan Pavlik, Johnny Lee and Wouter Wognum respectively. Derived from <http://rpavlik.github.com/wiimote-head-tracker-gui/>, <http://johnnylee.net/projects/wii/> and <http://wognum.home.xs4all.nl/wii/>.

The above mentioned developments in head tracking can be considered a bit unpractical for domestic use and the purpose of the present research in the sense that they need specific hardware and software to operate. As mention in the conclusion of Bresson and Shumaker`s work, a great development would be if one could use webcams as inputs to head tracking systems (Bresson and Schumaker 2010: 5).

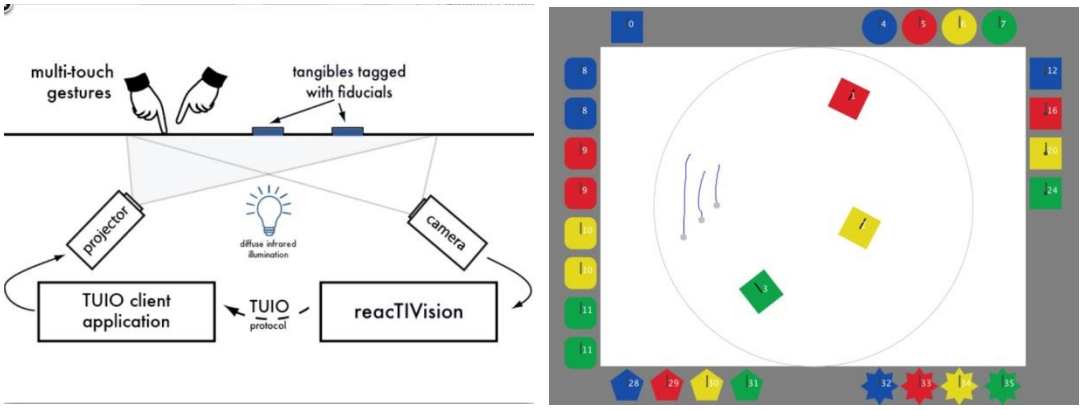
An approach that takes advantage of webcam systems would be the one based on fiducial marks, the same approach of the ReacTable project based on Reactivision sensor

components and FutureStories, including the Toronto Museum Project and the Magic Tunnel Popup Book.

According to the developers, the reacTIVision is an open source, cross-platform computer vision framework that allows fiducial markers tracking, normally attached onto physical objects, and multi-touch finger tracking. 'It was mainly designed as a toolkit for the rapid development of table-based tangible user interfaces (TUI) and multi-touch interactive surfaces' (ReacTIVision website). Developed by Martin Kaltenbrunner and Ross Bencina at the Music Technology Group of the Universitat Pompeu Fabra in Barcelona (Spain), the 'reacTIVision was designed as the underlying sensor component of the Reactable, a tangible modular synthesizer that has set the standards for tangible multi-touch applications'.



ReacTIVision screeshot and fiducial markers examples. Extracted from <http://reactivision.sourceforge.net/>.

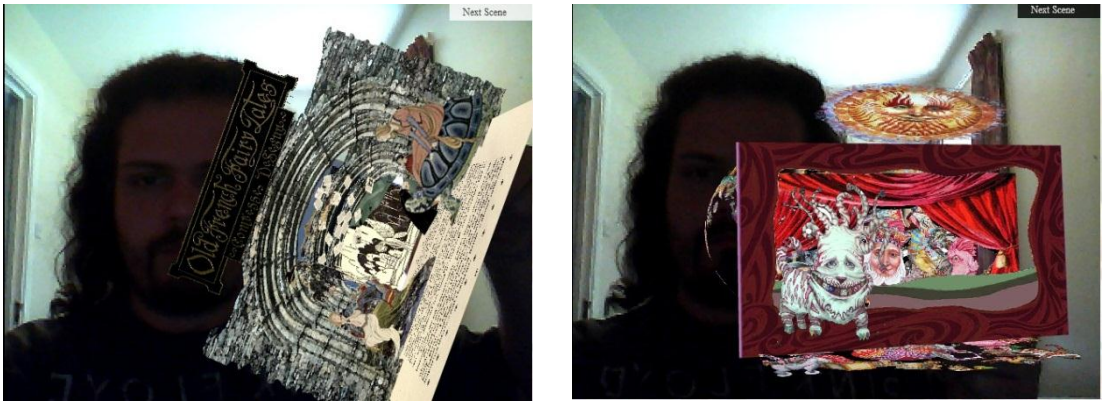


Framework diagram and TUIO simulator. Extracted from <http://reactivision.sourceforge.net/>

The ReacTable is sold as an instrument inspired by modular analogue synthesizer from the 1960's enhanced with digital effects and the ReacTIVision visual based webcam. It is also distributed in a mobile version for iPod touch, iPad and iPhone, and a version for public spaces and institutions, presented as a collaborative, didactic and entertaining tool.

ReacTable Live instrument. Extracted from <http://www.reactable.com/products/live/design/>.

The Magic Tunnel Popup Book and the Toronto Museum Project are part of an on-going work of the Augmented Reality Lab at York University in Canada. Both of them present to the public a handheld virtual landscape originated from a fiducial mark positioned in front of the camera. Gestures such as rotating, tilting and tumbling performed on the fiducial mark are reproduced by the virtual landscape.



Magic Tunnel Popup Book screenshots. Derived from <http://www.futurestories.ca/osc/index.html>.

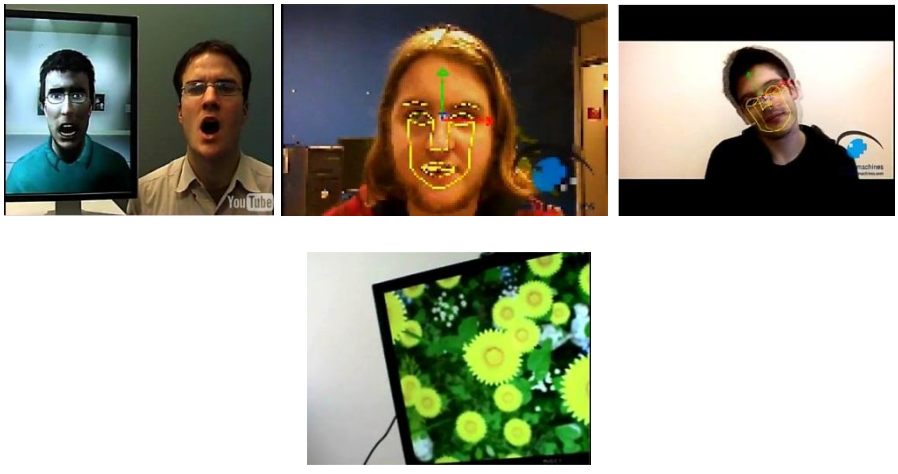
Toronto Museum Project. Derived from <http://www.futurestories.ca/toronto/>.

Recently, Dario Pizzamiglio's developments in this field provided a very interesting, useful and practical outcome. His MAX external 'headtracker', based on 'seeing machines API' application for face recognition, take the signal of a common webcam and gives the user three signals from x, y and z axis of head movements. Unfortunately until now (april 2011) there is just a version for PC but further developments will make this important tool available for all operational systems.

In his paper (Ludovico et al., 2010), the author describes this MAX object, used in the patch Head in Space, and argues that virtual reality 'does not involve only a virtual environment but also an immersive experience' and that 'instead of perception based on reality, virtual reality is an alternate reality based on perception'. To do so, the use of binaural techniques allows one to position a source outside the listeners' head, but to keep it the orientation of the listeners head must be known. Here the need of a head tracking

system is presented. Devices based on electromagnetic sensors currently in use demand dedicated hardware that make them suitable only for experiments and research. Since most personal computers are coming with a built-in webcam, their solution, developed in MAX/MSP, supports the dissemination of virtual reality applications.

The ‘seeing machines API’ algorithm, on which the external is based, takes frames from the webcam, processes them and delivers positional information in relation to the camera. The FaceAPI engine is accessed by MAX as a black box where bangs are used as inlet to activate the webcam and send floating point numbers corresponding to the position of the listener’s head. Their spatializer, application they use for demonstration, is based on the buffir~ object for HRTFs convolution and can store up to 256 coefficients. Distance simulation is implemented by amplitude compensation and low pass filtering for absorption simulation. The author suggests the use of more cameras to improve the implementation, since the angle of a single camera is very limited. A free for non-commercial use version can be downloaded at <http://www.lim.dico.unimi.it/HiS>. For the external to work, installing the faceAPI by Seeing Machine is required and can be downloaded at <http://www.seeingmachines.com/product/faceapi/downloads/>.



FaceAPI applications videos screenshots (controlling virtual character face expression, lip and eyebrow tracking, face tracking and 3D view dependent rendering). Derived from <http://www.seeingmachines.com/product/faceapi/faceapi-videos/>.

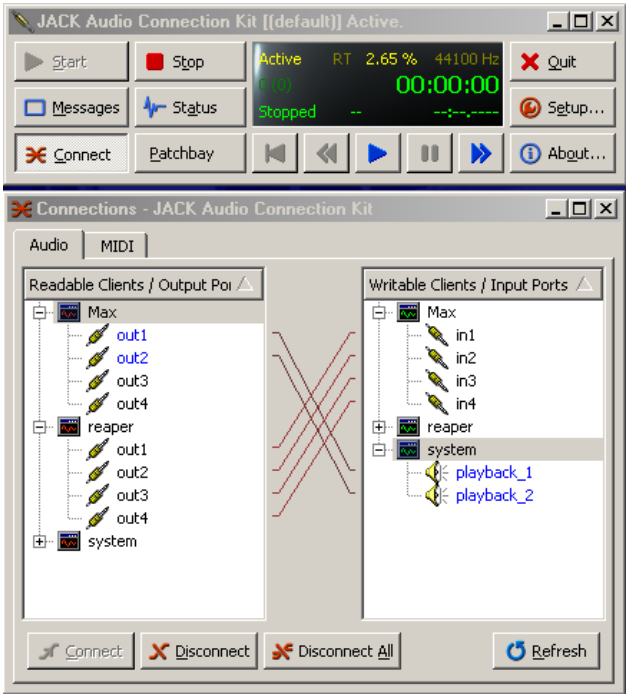
2.7 Linking a DAW to MAX/MSP

In the process of putting all the pieces together a noticeable difficulty on how to send audio information from the DAW, where sound engineers and composers can work as they usually do, to MAX/MSP, to be processed before going to the headphones, arose. The Rewire system was the first attempt but, although it can receive both audio and midi data, unfortunately it is not implemented to send audio data, making this choice inappropriate.

In previous experience using MAC operational systems (Sousa, 2010) the Soundflower virtual driver was the solution for such challenge. Unfortunately this is available only for MAC users. Attempting to develop tools suitable for Windows users that can take advantages of the 'headtracker' object, another tool needed to be found to perform the same task.

For Reaper DAW users, Rearoute, a virtual ASIO driver that allows routing audio to and from any other ASIO enabled application (similar to Soudflower in MAC), is a very handy solution and was the main reason the present author chose this DAW for the development of his research. In practice, an extra two channel track needs to be created allowing routing audio from MAX back to Reaper either for monitoring purposes or recording the binaural signal, as will be described in the third chapter.

Another tool that could perform such task, mainly for those that use any other DAW, is Jack Audio router, a virtual driver that allows the user to route any audio signal through any software or analogue inputs and outputs. With a few changes in the program register data, the number of inputs and outputs to be virtually generated by the software can be set up (<http://www.thuneau.com/forum/viewtopic.php?t=168>), if the user wants, for example, send a higher order B-format signal or the already decoded loudspeakers signals to MAX to perform the head tracking on the binaural domain.



Jack Audio router connections: sending B-format 1st order signals from Reaper to MAX and the binaural output from Max to the system outputs to be heard through headphones.

Nettingsmeier (2010) is one of the authors that are using a similar setup, with the Jack Audio router for connecting the DAW, in his case Ardour, to the Ambisonic decoder. He argues that this is a very interesting solution for surround mixing purposes since it presents an open structure of channels. Unfortunately restrictions always exist.

Chapter 3 – Tools performance into practice

Representing surround signals or even a whole three-dimensional sound field, complete with height, usually reproduced via many loudspeakers, through a binaural system over a pair of loudspeakers or headphones is not a completely new idea (Malham, 1998a; Manning, 2004; Rumsey, 2001). The concept of a 'virtual home theatre' first observed by the present author in (Rumsey, 2001: 75) is here extended to a 'virtual studio' where composers or sound engineers can work on any loudspeaker array they intend to reproduce their work, using only a pair of headphones. To do so, some tools to build it were chosen and the viability of the system put into practice.

3.1 Experiments and results

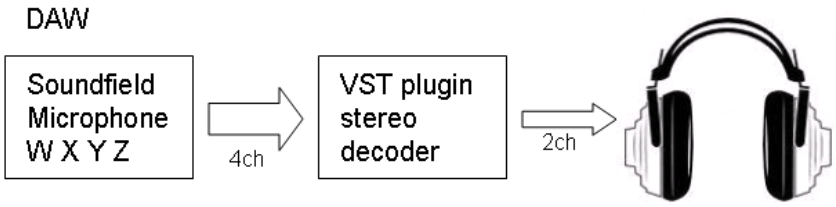
3.1.1 Binaural vs. stereo reproduction over headphones

A preliminary experiment was performed aiming at identify the strengths and weaknesses of binaural reproduction compared to stereo, both over headphones. Results were presented by the present author at the 130th AES convention in London (13-16 May 2011).

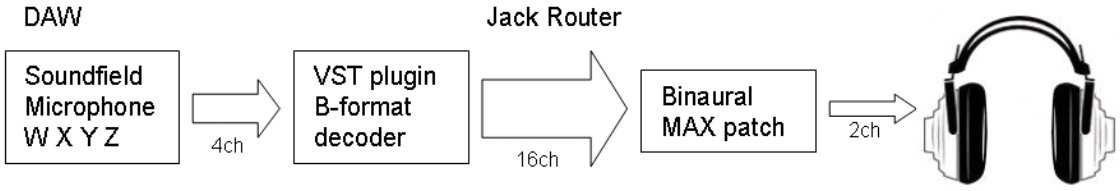
Two main objectives guided this experiment: to analyse the viability of transcribing B-format signals both to stereo and binaural formats in order to be heard through headphones and to recognize the perceptual advantages and disadvantages this transcription brings. To identify strengths and weaknesses of transcriptions from Ambisonic B-format to headphone based binaural systems is an important step for the development of the 'virtual studio'. The possibility of establishing binaural signals as a new format for distribution of multichannel musical material using Ambisonic B-format as a global exchangeable format between different already adopted formats is also analysed.

The 'virtual home theatre' concept applied to 5.1 multichannel signals reproduced over headphones has resulted in several interesting binaural reproductions (Sousa, 2010) and the application of this concept to B-format signals have been subject of some recent research (Barrett and Berge, 2010). Bearing in mind the known problems of binaural reproduction such as the in-head effect when there is a lack of a head-tracking system, as at this first stage, the use of generalized HRTFs and the influence of reverberation (Anderson, Begault and Wenzel, 2001), the experiment performed was intended to evaluate mainly the influence of the musical material and the spatial attributes that can be reproduced through headphones.

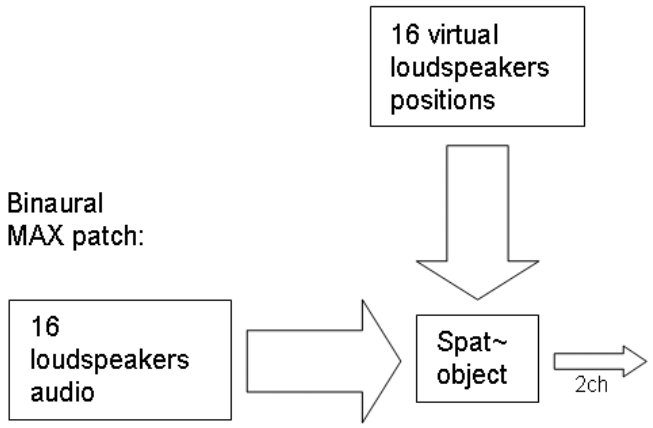
For the listening tests performed, a 1st order Ambisonic B-format signal was processed through a VST decoder plugin to obtain the traditional stereo file (Malham's B-mic VST plugin – set with an angle of 90 degrees between the capsules as a XY microphone array would be set), and a MAX patch developed, based on IRCAM's spat~ object (Jot and Warusfel, 1995), to deliver a binaural two channel signal, set so the signal would come from a 16 loudspeaker array (8 horizontal with front centre, 4 up and 4 down). The attributes evaluated were defined according to Rumsey's terminology (Rumsey, 2002) and divided into two different groups: preference and naturalness / spatial attributes.



Process chain to obtain the two channel file, from DAW to headphone reproduction, through VST plugin. Extracted from Sousa, 2011.



Process chain to obtain the two channel file, from DAW to headphone reproduction, through MAX patch. Extracted from Sousa, 2011.

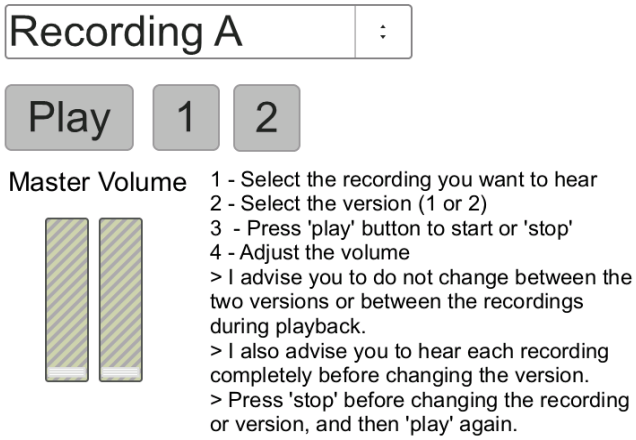


Process inside the MAX patch. Extracted from Sousa, 2011.

In the MAX patch using the spat~ processor, any special process were made, but only the positioning of the sound sources (virtual loudspeakers) into the binaural virtual environment, reverberation was obtained from the raw B-format recording of the natural environment.

The individuals that participated in the tests were music students since the familiarity with concert performance was necessary to fulfil the requirements for evaluating naturalness. Four recordings with different materials were evaluated. The first one was an orchestral performance in the Sir Jack Lyons Concert Hall (Music Department of the University of York) with a big choir behind it, the second was a solo piano concert performed in the same concert hall, the third was a field recording where a duck can be heard flying over the head of the listener and the fourth, a field recording as well, having some footsteps behind the listener as main characteristic.

A simplified graphical interface was developed in MAX/MSP for the students to change between the four recordings and the two versions, randomised between the binaural and the stereo to perform a blind test. A questionnaire was delivered before the listening test for the subjects to tick the version of the recording they preferred and which one best represented other attributes such as naturalness, wider environment, deeper environment, bigger group or elements in the recording, and presence. The possibility of not perceiving any difference between the files could be ticked in the (?) column, also available for the listeners, as well as some space for general comments.



Graphical interface for the selection of recordings and versions as well as the main controls for the hearing test. Extracted from Sousa, 2011.

Which recording did you prefer?
( ) 1 ( ) 2 ( ) ?
Which recording seems more natural?
( ) 1 ( ) 2 ( ) ?
In which recording the environment is wider?
( ) 1 ( ) 2 ( ) ?
In which recording the environment is deeper?
( ) 1 ( ) 2 ( ) ?
In which recording the group seems bigger?
( ) 1 ( ) 2 ( ) ?
Which recording gives you the greater sense of being there?
( ) 1 ( ) 2 ( ) ?

Questions asked for each one of the four recordings. Extracted from Sousa, 2011.

In contrast to what some other researchers have reported (Fredriksson and Zacharov, 2002) namely that 'using different sound material (classic music, rock, jazz, environmental sounds) did not seem to effect the grading' (Fredriksson and Zacharov, 2002: 8), while analysing reproduction through different arrays of loudspeakers, the high variance noticed in the present experiment between the four recordings for almost all parameters shows some dependence of the decisions on the musical material.

Similar results were found by Baume and Churnside in relation to loudspeaker reproduction and testing an Ambisonic system. They reported that, although Ambisonic is not preferred while playing back dialogues, 'the preference for Ambisonics seemed to show some correlation with the type of material' (Baume and Churnside, 2010: 10). They also reported that 'there was a clear preference for height when using atmospheric, non-directional content. For items containing sources in-front and in the horizontal plane, such as music, there was no clear preference for the use of height' (Baume and Churnside, 2010: 7). Music material dependence of localization reproduction in binaural systems can also be found in Lindau et al. (2007), where they reported in their conclusion that guitar and voice samples presented being easier to localize.

Although, in the present experiment, an overall analysis of preference points to the stereo recording versions as also does the naturalness parameter, a significant incoherence can be perceived between these two parameters. This incoherence may mean that people are very influenced by pop music that they usually hear through headphones, or that the participants of the present research are not as used to classical music performances in concert halls as expected, what could make the decision for naturalness more difficult. Similar incoherencies are also reported by Baume and Churnside, in their experiments between presence and preference parameters.

The spatial attributes seems to be harder to evaluate than the others, with parameters related to the size of the sound sources definitely being the hardest. The parameters 'environment wider' and 'environment deeper' presented a tendency in favour of the

binaural processing as expected, but in those parameters the variance between the different musical materials are the highest, which may represent some dependence of a good binaural reproduction on the musical material and to reverberation level and naturalness. The parameter 'presence', points to the stereo processing but again can be related to preference and the influence of pop music. An interesting incoherence is also perceived between the parameters 'presence' and 'naturalness' that should definitely not occur, but can be related to the 'presence' concept from pop music productions, when a source is more present when it is spatially closer to the listener.

In the general comments, some listeners reported the binaural processing as presenting some kind of 'phasing effect' that can be associated with the large number of loudspeakers represented virtually as a result of the interaction of their signals. In these cases, this effect strongly influenced the subject's decision on 'preference' and 'naturalness'.

Preference	**Binaural**	**Stereo**	**?**
Recording A	30%	60%	10%
Recording B	10%	80%	10%
Field recording A	40%	50%	10%
Field recording B	30%	60%	10%
Average	27,5%	62,5%	10%
Variance	15,8%	15,8%	0%
Naturalness	**Binaural**	**Stereo**	**?**
Recording A	30%	70%	0%
Recording B	10%	80%	10%
Field recording A	20%	70%	10%
Field recording B	30%	60%	10%
Average	22,5%	70%	7,5%
Variance	9,1%	6,6%	2,5%
INC preference/naturalness			
Recording A	30%		
Recording B	20%		
Field recording A	40%		
Field recording B	40%		
Average	32,5%		
Variance	9,1%		

Preference and naturalness data. Extracted from Sousa, 2011.

Environment Wider	Binaural	Stereo	?
Recording A	40%	30%	30%
Recording B	70%	20%	10%
Field recording A	30%	40%	30%
Field recording B	30%	60%	10%
Average	42,5%	37,5%	20%
Variance	35,8%	29,1%	13,3%

Environment Deeper	Binaural	Stereo	?
Recording A	60%	20%	20%
Recording B	40%	30%	30%
Field recording A	40%	50%	0%
Field recording B	40%	40%	20%
Average	45%	35%	17,5%
Variance	10%	16,6%	15,8%

Group Bigger	Binaural	Stereo	?
Recording A	40%	40%	20%
Recording B	40%	20%	40%
Field recording A	30%	30%	40%
Field recording B	20%	70%	10%
Average	32,5%	40%	27,5%
Variance	9,1%	46,6%	22,5%

Presence	Binaural	Stereo	?
Recording A	20%	60%	20%
Recording B	30%	60%	10%
Field recording A	50%	40%	10%
Field recording B	30%	60%	10%
Average	32,5%	55%	12,5%
Variance	15,8%	10%	2,5%

INC presence/naturalness	
Recording A	10%
Recording B	20%
Field recording A	30%
Field recording B	40%
Average	25%
Variance	16,6%

Spatial attributes data. Extracted from Sousa, 2011.

If one aim to distribute recordings with surround material, especially Ambisonics recordings, through two channel files for headphone listening, the effect of the interaction of the loudspeakers signals in a virtual environment needs to be analysed. Pulkki, in his studies related to VBAP, also performed similar hearing tests to evaluate multichannel systems reproduction through a binaural model. In his tests, aiming to evaluate directivity accuracy of Ambisonics and VBAP sound representations, the author reports that, in general, VBAP reproductions are more stable than Ambisonics. This directional quality can be improved by increasing the Ambisonics order but ‘when the number of loudspeakers is

increased, the directional quality is degraded' (Pulkki, 2001: 4). For VBAP systems, increasing the number of loudspeakers improve directional quality. These results points to the use of high order Ambisonics or VBAP for binaural systems in an attempt to reduce the reported 'phasing' effect.

Jot et al. also refer to this 'phasing' effect when reproducing multiple virtual loudspeakers through a binaural system. They argue that a possible cause of such audio artefact in binaural reproduction is that 'the same interpolation method applied to mixed-phase impulse responses yields unsatisfying interpolations of the spectra, due to "comb-filtering" effects caused by mixing responses with different delays' (Jot et al., 1995: 17). In another article Jot et al. reported that this comb filtering is due to the different delays of the virtual loudspeakers signals and added that one can reconstruct the HRTFs better by increasing the number of encoded channels through the use of HOA for example, and that increasing the order but not the number of loudspeakers may be effective (Jot et al., 1999).

The above mentioned 'phase effect' is also reported by Nettingsmeier, while playing back film sound over an Ambisonic system, but not for playing back music material. He argues that the dry acoustics of cinemas can make the 'phasing effect' more evident (Nettingsmeier, 2010: 4). These observations point to the discussion that direct transcriptions to binaural can be improved by simulating a virtual environment with some imperfections instead of working with the signals where there was no interference from loudspeakers response or room influence, as has been done so far.

Other considerations must be made in respect to the frequency range of good spatial representation of Ambisonic systems compared to binaural. According to Bridge and Jagger, the Soundfield microphone has some limitations:

> Ideally the sphere would be a point but the practical spacing of the capsules is fully compensated so that the resulting B-format signals appear to originate from a point up to frequencies of about 10KHz and are truly coincident to this frequency.
>
> (Bridge and Jagger, 1984: 6)

Other authors that observe the same limitations of Ambisonic signals are Jot et al. that argue that ‘as frequency increases, growing errors can be expected in the reconstruction of both interaural time differences and HRTF spectral cues’ (Jot et al., 1998b: 4) since the main differences in HRTFs are in the region from 8KHz and 10Khz (Jot et al., 1998a: 6). This known limitation of the Soundfield microphone conflicts with the resolution of HRTFs for a good reproduction of directional content in an Ambisonic based binaural system, since some important details of the HRTFs, directly related to our capabilities to localize sound sources, can be observed above 10KHz.

From this preliminary experiment we can conclude that a significant potential for binaural reproductions, mainly regarding the reproduction of spatial attributes, can be noticed. Future work needs to focus on avoiding the artefacts binaural reproduction can generate and user friendly head-tracking systems, bearing in mind that the need for specific software and hardware that can make such systems less acceptable for domestic applications.

Other considerations that also need to be reported are the possibilities for distributing binaural audio as a final product. The experiments of George et al. (2009), for instance, focused on subjective comparisons between a stereo downmix, a mono downmix and two different binaural algorithms reproductions, report that listeners used to binaural reproduction tend to prefer this kind of system. Some users reported that changes in timbre compromise binaural reproduction, as do the high amount of reverberation used to obtain good externalization, but the author highlights that the 38% of listeners preferring the binaural algorithm, as reported by the experiment, is a large number.

Some people in Sursound email list also reported interest on having their downmix from music originally mixed in 5.1 to be binaural instead of traditional stereo and, in the music industry, some music productions are already being distributed with two different versions, a binaural and a stereo one, as done in the 3D60 sound production process (<http://www.3d60.co.uk/index.php>).

3.1.2 Putting pieces together: the 'Virtual Studio'

What is described here as the ‘virtual studio’ consists of a MAX patch that takes advantage of the B-format Ambisonic signal, which can be decoded to different loudspeakers arrays, and, after decoded, applying HRTFs, corresponding to each one of the loudspeakers positions in the chosen array, by convolution, to their output signal. The HRTF set can be chosen between several different subjects’ measurements from the IRCAM ‘LISTEN’ database, as well as between the compensated (to eliminate amplifier, microphones and ears canal influence) and raw sets. Head tracking based on a webcam signal is also implemented as well as headphone compensation (only for the AKG 271MKII headphone model) and an optional high shelf filter boost, in an attempt to improve localization. The main objective of this supporting tool is to allow sound engineers or composers to work in a binaural based virtual environment, reproduced through headphones, before going to the studio to finalize their production by monitoring it through the loudspeakers. Following, a historical background on the same topic is reported.

The first reference to a similar system was found in Moller (1992), while describing applications of binaural systems. He argues that, among others, the ‘simulated loudspeaker reproduction’ application is ‘especially attractive for control room simulation’, and this approach ‘may compete with the control rooms or be a supplement to them’ (Moller, 1992: 210) since it is cheaper and smaller than the control rooms normally used in music and film production.

In 1995, Jot et al. mention the possibility of using binaural techniques to improve the reproduction of conventional stereo over headphones, aiming to avoid the in-head effect. Such process is also reported by Hammershoi in 1996, who says that among other applications, such as room simulation, binaural mixing, teleconferencing, 3D auditory display and virtual reality, binaural processing

> may also be used to transform *Dolby Surround sound to headphone* reproduction, or any other

> multichannel sound standard, for which it is not trivial to obtain a natural two channel representation for headphones. This goes for stereo as well, in fact, where the stereo signals intended for the loudspeakers gives the listener the quite unusual sensation of the sound sources being in the head.
>
> (Hammershoi, 1996: 3)

In the work of McGrath and McKeag, also from 1996, they include a head tracking system in the process and argue that their 'new "Binaural Decoder"', implemented using the Lake Huron hardware, 'solves the traditional problems of binaural sound and provides a way of listening to B-format recordings over headphones' (McGrath and McKeag, 1996: 1). In their approach a set of static binaural filters are used and the user can choose between the sets of HRTFs available. Distance cues are obtained by reverberation processed by the CATT Acoustic (http://www.catt.se/) and delivered in B-format. He highlights the possibility of monitoring loudspeaker arrays reproduction through the headphones:

> Until now, there has been no way to monitor spatial recordings during the production process without resorting to inconvenient and space consuming loudspeaker arrays. The Binaural Decoder solves this problem also. It can be used to monitor spatial sound at every stage in the production process.
>
> (McGrath and McKeag, 1996: 5)

Jot et al. (1998b) establish the concept of binaural B-format after a discussion about the cost of implementing interpolation between HRTFs to perform head tracking or sound source movements and especially when multiple sources came into play. According to them, this concept can be exploited in two different ways: a 'post filtering' approach, where the source signals are panned and mixed in an intermediate format and then transcoded to a binaural signal using time-invariant filters, and a 'pre-filtering' approach, where the sources are previously processed by the filters and then panned in real-time (Jot et al., 1998b: 2-3).

The concept of a binaural downmix is also presented and consists of 'exploiting a known multichannel panning technique for reproduction over loudspeakers, and using binaural

filters to produce a binaural downmix for reproduction over headphones' (Jot et al., 1998b: 3). It is important to notice that in such process the final result is a two channel file that does not allow changes while the listener performs head movements. To perform such a task, the Ambisonic B-format is presented as an intermediate multichannel encoding format. 'The B-format offers the possibility of mixing with an ambience recorded with a Soundfield microphone, and compensating for the rotations of the listener's head after mixing by applying a rotation matrix before transcoding to binaural format' (Jot et al., 1998b: 3). Jot et al., in 1999, add that this ' "virtual loudspeaker" paradigm is a general approach to designing decoders for adapting any encoding format to playback over headphones or any multichannel loudspeaker layout'.

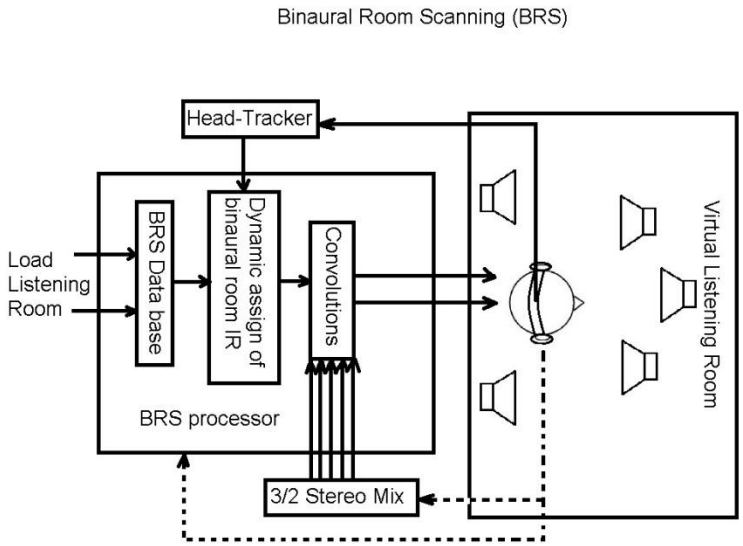
Felderhoff et al. (1999), in their experiments to evaluate the influence of head tracking processing, present a method of auralization based on reproducing a set of loudspeakers through headphones and named it 'binaural room scanning', or BRS, this process is described by them as a monitoring system of different multiple loudspeaker setups. They quote some other works where front-back confusions vanish due to head tracking systems implementations, as well as experiments with a step motor turning the dummy head according to head movements of the listener. They argue that

> Considering head movement as localization cues, not only front-back-inversions vanish, but the localization performance nearly equals that of natural hearing. Even, if the typical spectral cues of the pinnae are absent the localization in the horizontal plane remains good, as long as the head-tracking is enabled, though elevations are reported. Thus, an auralization system has to allow for head movements to enable sound localization that is comparable to normal hearing.
>
> (Felderhoff et al., 1999: 2)

They present a problem of this kind of reproduction that has to be considered in any listening procedure and is related to the fact that 'using only headphones for the display, low-frequency vibrations are not transmitted'. They also refer to studies that argue that height information of sound sources can be better represented with head tracking. In their application purposes they state that:

It is well known that the acoustics of an OB van is less convenient than that of a studio control room. However, if the sound engineer's control room is *binaurally scanned* and loaded into the BRS-processor, he will have a listening experience via headphones, comparable to normal hearing in a familiar studio environment. Because several rooms can be stored in the BRS-processor, a multichannel control room or a surround-sound movie theatre as well as an "average" listening room can be chosen.

(Felderhoff et al., 1999: 3)



Binaural Room Scanning scheme. Derived from Felderhoff et al., 1999.

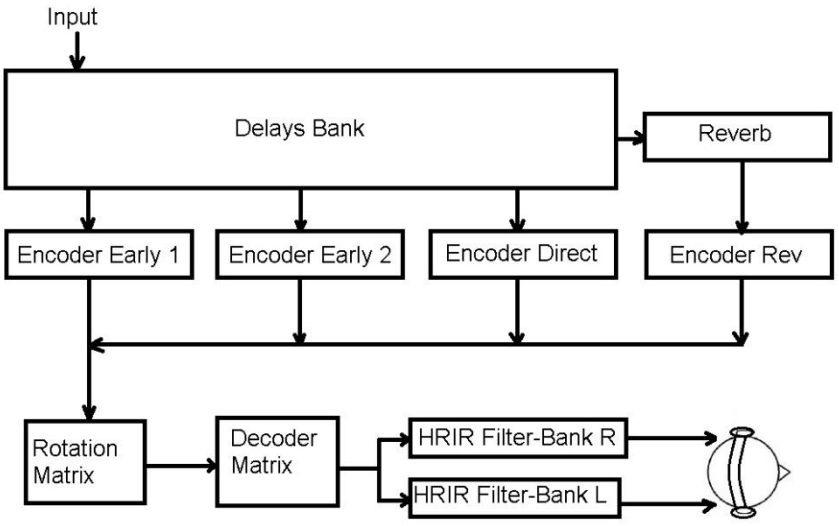
In Höldrich et al. works, from 2003, they developed a system for virtual Ambisonic array reproduction as well as a head tracking implementation based on a gyroscope and using the PureData platform. According to them,

To overcome the problem of high-quality, time-varying interpolation between different HRTFs in time-variant binaural sound reproduction systems, a virtual Ambisonic approach is used. This approach is based on the idea to decode Ambisonic to virtual loudspeakers. Then the binaural signals are created by convolving the virtual loudspeaker signal with HRTFs appropriate to their position in space. Now, the filtered signals are superimposed to create left and right ear headphone signals.

(Höldrich et al. 2003a: 2-3)

In their implementation of the head tracking system, the Ambisonic domain approach is taken into account where simple rotation matrices are performed around the z axis, e. g. changing the horizontal plane orientation. The head tracking device is mounted on the headphone and for implementing the room simulation, simple delays are calculated for the first reflections of a rectangular room. They also argue that an advantage of using Ambisonic is that the number of sources to be spatialized does not increase processing of HRTFs.

They point out that due to the fact that high quality time varying interpolation of HRTFs to reproduce sound source movements or implement head tracking is very cpu expensive for real time applications, the Ambisonic approach is proposed, resulting in the use of time invariant HRTFs.



Block diagram of the Ambisonic based binaural system developed by Höldrich et al. Derived from Höldrich et al. 2003a.

In an application of their system, Höldrich et al. (2005a), describe a recreation of the Varèse Philips pavilion in a virtual environment added to a real time virtual visualization of the building. In this description, the listener or composer is able to walk around the

pavilion as well as compose new pieces to be played in such environment. In another article (Höldrich et al., 2005b) the addition of an implementation of a controller based on video tracking of objects, representing the listener and the sound sources placed on a table, is described.

As be seen from above, discussions about whether to perform the head tracking in the Ambisonic domain or in the binaural domain have been extensively pursued by many authors. Head-tracking systems acting on signals in the binaural domain can be implemented by changing the loudspeaker position correspondent HRTFs to be convolved with the outputs. In the Ambisonic domain, it is achieved by using fixed HRTFs corresponding to the loudspeaker positions and manipulating the matrices that control rotating, tumbling and tilting movements of the B-format signal. It seems that the B-format approach to simulate listeners' head movement has been widely accepted, mainly due to the fact that it significantly reduces the cpu processing power required.

The first approach is described, for instance, by Wenzel (1996) while referring to the Convolvotron system that has head tracking implemented in the binaural domain. The second approach is described in many publications such as Malham (1993, 1999b) where the author argues that when reproducing a complete 3D sound field through HRTFs and headphones, a lot of computational power is needed and, by using B-format Ambisonic signals, this computation can be reduced and operations like rotation, tilt, tumble and mirroring can be performed.

> By placing all the sound sources in a B format soundfield including, if required, full complexity natural soundfields recorded with a Soundfield microphone, the processing involved in rotating, tilting, etc. the full soundfield as required in head tracking configuration is significantly simplified compared to that involved when directly processing HRTF's. The B-format signals can then be decoded to virtual loudspeakers feed signals and only these need to be passed through HRTF filters. Since these are limited to a single fixed set of HRTFs, it is possible to do all necessary operations on standard hardware, even when full head tracking is in use.
>
> (Malham, 1999b: 26)

Jot et al. (1995), when observing the need for efficient dsp for real time applications in binaural reproduction, also argue that it can be best achieved by performing head tracking in the Ambisonic domain since it is cheaper in terms of cpu usage. Travis also discusses this Ambisonic approach for head tracking and points at it as a natural choice:

> The Ambisonic representation has a bonus feature, which is that it allows relatively easy rotation of the sound field. This open up the possibility of making the HRTFs filters static and accommodating the user's head movements by dynamic rotation of the ensemble. The filter interpolation and commutation issues would then go away, making for a simpler and hopefully more efficient implementation.
>
> (Travis, C. 1996a: 115)

Jot et al. (2006) suggests that for the various applications of three-dimensional sound, the adoption of a form of rendering audio that is completely independent of the playback setup is required. This idea fits with the implementation of a virtual studio using Ambisonic B-format signals. They also quote Malham 1993 and Travis 1996b, when the authors present that the solution for using less cpu for rendering binaural signals would be the use of pre-processing in the Ambisonic domain and convolving signals from the virtual loudspeakers from the decoding process, then one would have static binaural processing and could perform the dynamic process in Ambisonic domain that is much less cpu expensive.

Daniel (2003) also mention the possibility of rendering Ambisonic B-format signals to headphones through 'virtual loudspeaker process, which consists in the binaural simulation of each loudspeaker for a centred listener position, i.e. the filtering of its signal by the corresponding HRTF (Head related Tranfer Function)'. Goodwin and Jot (2007) also describe the virtualization of multichannel audio based on HRTFs processing. According to them the 'virtual 3-D audio reproduction of a two-channel or multichannel recording traditionally aims at reproducing over headphones the auditory sensation of listening to the recording over loudspeakers' (Goodwin and Jot, 2007: 2-3) and consists of applying HRTFs filters to the loudspeakers signals and summing them afterwards. George et al. (2009) argue that

> Currently, most existing audio content is produced for loudspeaker reproduction. In order to render such material for headphone reproduction, each loudspeaker can be represented by a virtual source placed at a defined location with respect to the listener. This is typically achieved by binaural processing.
>
> (George et al., 2009: 2-3)

Another author that discusses this approach is Nicol (2010), who also gives us interesting considerations on virtual loudspeakers reproduction over headphones. She points out the difference between the 2 channel and the multichannel implementation of binaural synthesis. In the 2 channel implementation there are as many HRTFs convolutions as sound sources, multichannel implementation, on the other hand, is related to the concept that of virtual loudspeakers which

> consists in simulating a multi-channel loudspeaker setup by headphones. Each virtual loudspeaker is synthesized by binaural synthesis with the appropriate pair of filters. In other words, this is a solution that enables us to adapt multichannel content (5.1, 6.1, 7.1, 10.2, 22.2...) to headphone reproduction, which allows one to listen to spatial audio without the need of any loudspeaker equipment, be it simple or complex. That's why it is referred to as *binaural downmix*. The spatialization of the sound source is controlled by the primary multi-channel format (for instance intensity panning), which defines the spatial functions, whereas the filters are determined by the HRTF corresponding to the specific location of the virtual loudspeakers.
>
> (Nicol, 2010: 41)

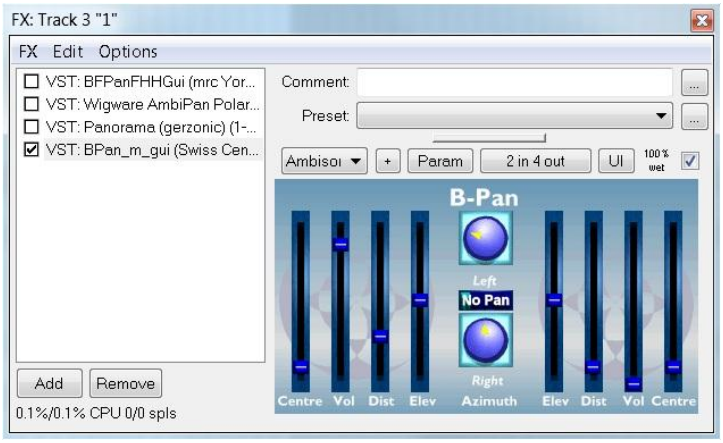
According to the author, 'for consumer equipment, such as home cinema or music listening, binaural offers spatial enhancement through the technology of virtual loudspeakers (or binaural downmix) that allows one to listen to multichannel (5.1, 6.1, 7.1, 10.2, ... or even Ambisonics) content over headphones' (Nicol, 2010: 63). She also quotes some developments that have similar approach directed to 5.1 material reproduced over headphones such as the Dolby®Headphone or Dolby®Virtual Speaker and the Beyerdynamics®Headzone. She concludes that 'one great advantage of binaural technology is its compactness, which makes it appropriate for any handheld device' (Nicol, 2010: 63).

Boland et al. (2010) quote other works that have been demonstrating auditory artefacts caused by wave discontinuity in the convolved binaural signal obtained after the process, when the HRTFs are interpolated according to head movement, a reason for the adoption of the flexible approach of virtual loudspeakers processing where the listener can be placed at the centre of the imaginary array.

The present development of the 'virtual studio' where, through headphones, one can monitor multichannel mixes is based on all the studies previously described and involves the use of five main tools chosen due to their easy access by composers and sound engineers:

Ambisonic encoder (VST or MAX object)

Intended to process a mono sound source into a B-format Ambisonic audio signal, the encoder (VST) can be inserted in a track in the DAW and used to send to a MAX patch the four signals correspondent to the 1st order B-format. Optionally the user can have a player in MAX that send a mono signal to the encoder object (also in MAX) before the processing stage. As described in the previous chapter there are a lot of free downloadable VST encoders and MAX objects that perform this task. In the present implementation, VST plugins were chosen due to the familiarity most of the sound engineers have with DAW manipulation of panners and the simplicity of the automation process of their parameters performed as any other VST plugins.



FX section of a mono track in Reaper, where the Ambisonic encoder can be chosen.

Special attention must be taken when routing signals in Reaper, mono track with mono sound files have Ambisonic B-format panners inserted, and then have four outputs. These four outputs must be routed through the four first channels in ReaRoute to be received in MAX by an adc~ 1 2 3 4 object, which are monitored and sent to the processor, as must be sent any four channel track with recordings made with a Soundfield microphone. MAX has to be set to use ReaRoute driver and its outputs are send back to Reaper through the two first ReaRoute outputs. A two channel track in reaper must receive these and set for monitoring through the computer physical interface driver

Ambisonic processor

The role of the processor is to perform changes in the orientation of the three-dimensional virtual environment, in the three axes – X, Y and Z – according to the head tracking system outputs. The processor used in this implementation was the ambiproc~ object ported to MAX by Matthew Paradis from the VST plugin developed by Dave Malham. This choice was made due to the need of integration with the headtracker object, developed for MAX platform.

The ambiproc~ object is receiving the four audio outputs from the adc~ object and processing them according to the ‘p control’ patch output signal. The ‘p control’ is receiving signal from the three scaled signals of the headtracker object and performing an average of the stream of numbers that are coming, using a bucket object and a mean object in each one of the three inputs – X, Y and Z. The average is needed since the flow of numbers coming from the headtracker is very instable and sensible to fast movements.

Ambisonic decoder

The decoder generates the loudspeakers audio signals for the chosen array which can be directed either to the real loudspeakers or to the binaural processing stage. For this implementation the decoder chosen was the one from the ICST set of tools and the main reason of this choice is that it allows easy changing between different loudspeaker arrays by sending simple messages that can be set by a pop-up menu.

In the pop-up menu, one can select between six available loudspeaker arrays and this happens through bang messages sent by the menu to previously established messages and by turning the decoder outputs on and off (also by messages). The pop-up menu also triggers predefined set of messages to set SIR vst plugin presets for binaural convolution.

Binaural processor

The binaural processor will convolve each signal from each loudspeaker with a correspondent HRTFs and sum them together to obtain a two channel signal routed to the headphone. This process can be achieved with multiple VST convolvers, such as the SIR, that is freeware, or with the spat~ object in MAX, which also gives the possibility of performing the head movement processing in the binaural domain by HRTFs interpolations instead of in the Ambisonic domain. Some differences were noted between the two processes but there were not deeply analysed as the option of using the spat~ object would limit the access to the patch since it is not a freely downloadable object. On the other hand, the choice of using the SIR VST plugin limited the patch to eight loudspeakers arrays setups, since the processing gets very cpu expensive with more instances of this VST plugin.

Several presets were previously saved (.fxp files) to be loaded when the SIR vst plugin receives proper messages. These presets load a HRTF filter (wave impulse response files) corresponding to the loudspeaker position in each one of the eight SIR vst~ plugin objects and adjusts wet (-6dB) and dry signal (to 0). An observation need to be made at this point: due to the fact that these presets were created in my personal computer, it will look for the HRTF database at the following driver path C:\Users\fabiojanhan\Documents\York\Research\Virtual studio\HRTF Database\, where the folder 'HRTF Database' must be placed by the user. The messages received by the SIR vst plugin are structure as follows:

Ex.: S1CT045P315, where

S1 is the number of the subject the HRTF is from (subject 1)

C means if the HRTF is from the compensated set or raw (compensated)

T045 is the azimuth angle according to the LISTEN nomenclature (45 degrees)

P315 is the elevation angle according to the LISTEN nomenclature (315 degrees)

The messages are created by 'sprintf' objects which receive signal from the HRTF pop-up menu, the compensated-raw pop-up menu and already established positions of the loudspeakers, triggered by the loudspeaker array pop-up menu. A message 'wet 0' are triggered according to the selected loudspeaker array to assure that those SIR vst~ that are not in use are off. All the 51 subjects HRTF found in the LISTEN database are available for the user to fell free and try different ones, looking for the one that fits best to him.

Head tracking

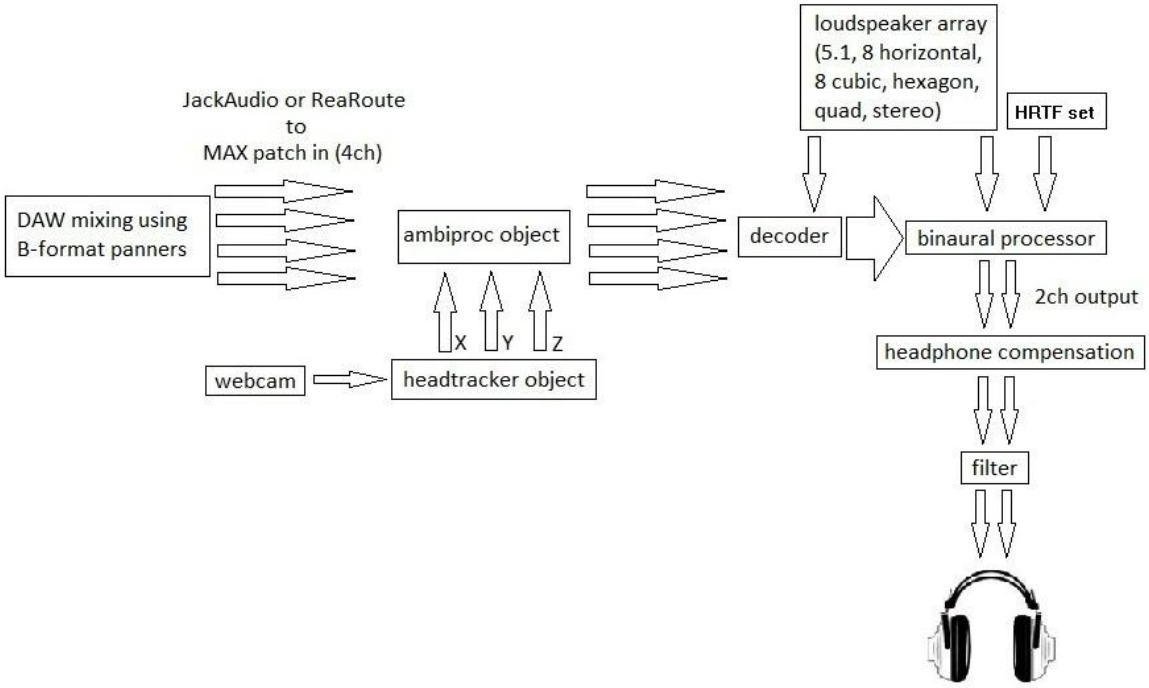
The first attempt at head tracking was based on fiducial marks, but due to their slow response to movements other possible options were examined. The headtracker object in MAX (unfortunately available only for windows users) was chosen due to its ease of usage and no need of hardware device. It allows the loudspeaker array to respond to the mixing engineer or composer's head movements by processing a webcam signal. The outputs used in the present patch implementation are the three signals corresponding to rotation on the three axes (X, Y and Z). These are scaled empirically for better sensibility and considering that the region the headtracker works is different from the region ambiproc~ does. Since the webcam has a sensibility of around 60 degrees and the ambiproc~ receive signals to vary between 0 and 360 degrees, the scale process is mandatory. The signal is then sent to the three inputs of the ambiproc~ object for its manipulation, in this case performed in the Ambisonic domain. A control for switching the head tracking processing on and off was implemented to allow comparisons to be made by the users.

Another implementation made in the patch was the headphone compensation for the use of an AKG 271 mkII, attempting to cancel its tonal coloration, performed by convolving the system output with an inverse impulse response obtained from the same database as the HRTFs set (<https://dev.qu.tu-berlin.de/projects/measurements/files>), also loaded into a

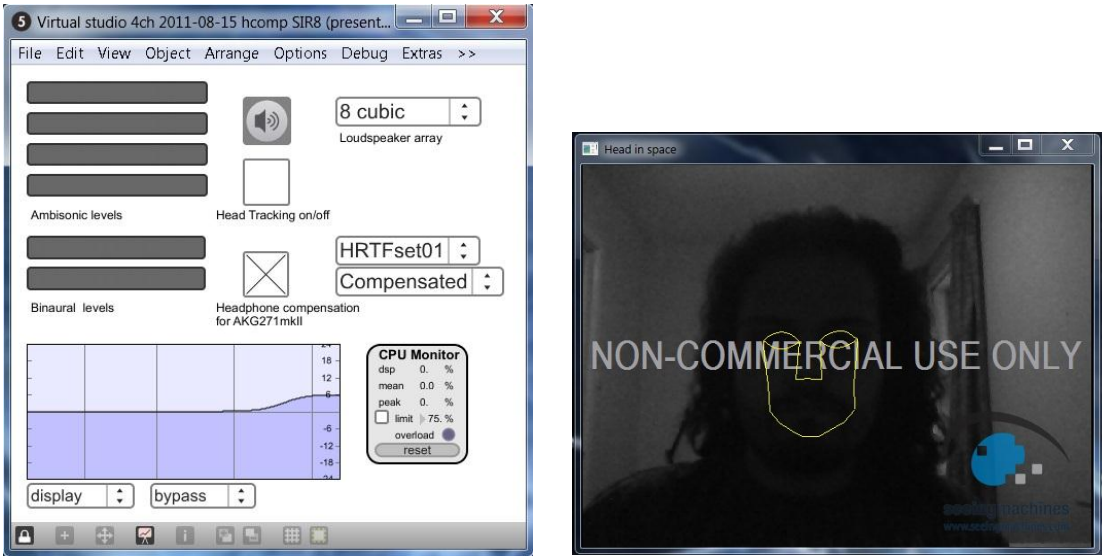
SIR vst~ plugin. A switch to turn the headphone compensation on and off was included in the user interface as well as level meters for monitoring both Ambisonic inputs and binaural outputs. A cpu monitor was included in the patch as well as a filter that can be set for high shelf boost, an attempt to compensate for the lack of spatial resolution of Ambisonic systems and improve sound sources localization by boosting the high frequency region.

Although some works describe the need for simulating room reverberation for better reproduction of virtual environments, this was not included in this implementation. This choice was made due to the fact that one of the main objectives was to reproduce recordings where the original reverberation of the recorded place was already present or was obtained by convolution with pre-recorded B-format impulse responses files. Adding room simulation would include undesired coloration to the reproduction through headphones. In other words, this implementation is trying to simulate an ideal mixing environment, absent of any acoustic interference.

In the user interface, one have access to six different loudspeaker arrays (stereo, quad, 5.1, hexagon, 8 cubic and 8 horizontal) and the set of HRTFs (raw and compensated) to be chosen according to personal choice or by just picking one of them and getting used to it, procedure already reported by some authors to be efficient. The B-format signal can be sent from a DAW to MAX by using JackAudio, a virtual driver, or by ReaRoute, virtual driver inside the DAW Reaper.



Block Diagram of the developed system.



Virtual Studio MAX patch user interface and Head in Space interface for head tracking system monitoring.

During a year lots of recordings were made at the music department of the University of York to evaluate the system. These recordings as well as two compositions are going to be described in the next sections. Also aiming to evaluate the patch developed, several students were invited to a session where they could try their own musical material (in development) through the binaural system, and compare to the original reproduction

through a loudspeaker array in the Trevor Jones Studio (Music Research Centre of the University of York). A questionnaire was delivered, so they could express their considerations on the advantages and fidelity of the system.

Due to the fact that insufficient students were present at the listening session, it was not possible to perform a quantitative analysis and few considerations could be extracted from the answered questionnaires. Considerations were verbally expressed and opened to the present group. The fact that the 'virtual studio' MAX patch was useful and easy to use was agreed, as well as the possibility of using it as a supporting tool for future developments. The naturalness with which it can reproduce loudspeakers arrays in virtual (binaural) environments still need some data to be analysed properly, but it was agreed that mixing or producing in such a system is not definitive and all the material would need to be heard through loudspeakers to be finalized, probably due to the fundamental concept and nature of headphone reproductions.

Head tracking definitely improves the virtual reproduction of sound sources, even if individualized HRTFs are not being used, but the headphone compensation and the high shelf boost still need some detailed analysis to be considered an important improvement. One of the listeners, for instance, definitely preferred the binaural reproduction without the headphone compensation, which lead us to thoughts about the real advantages it brings to the system. Overall the listeners were satisfied with the results and the patch proved to be a useful tool for future multichannel audio productions.

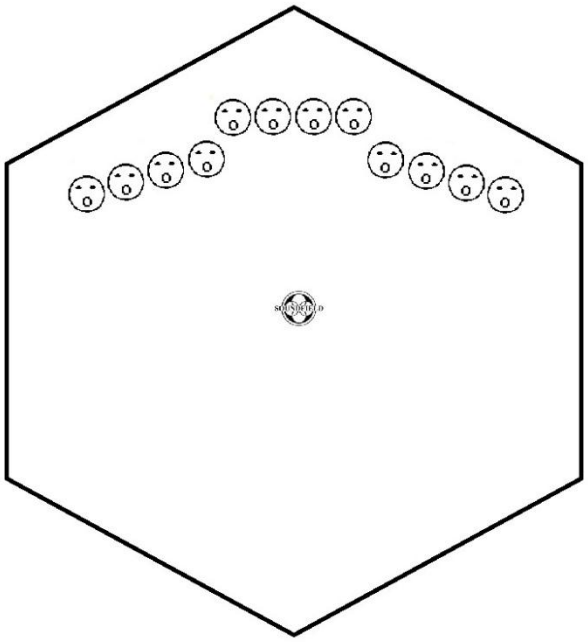
3.2 Producing and composing for Ambisonics and monitoring through headphones

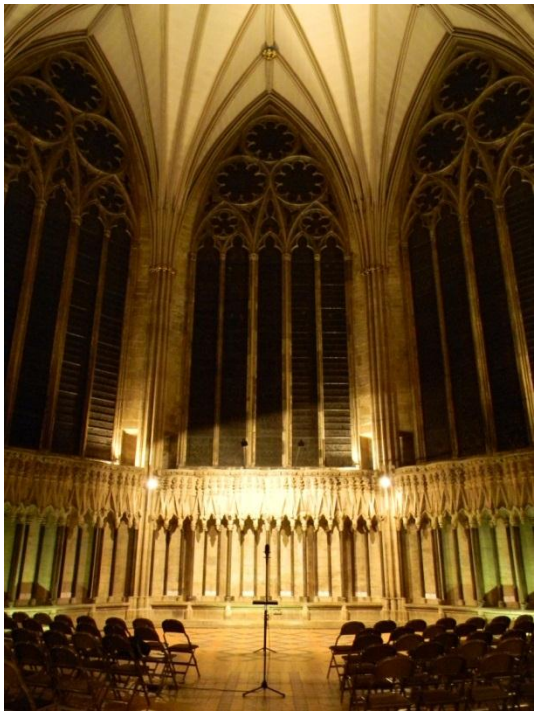
> Ambisonics is a unification and extension of prior art in sound field recreation.
>
> (Gerzon, 1985: 859)

3.2.1 Musical productions

Once the 'virtual studio' patch is implemented there was a clear need to try working with this tool. Recordings were made in order to try out different musical materials and different environments, and mixed using the virtual environment reproduced through the headphones. Some preliminary recordings were done for the experiment described in section 3.1.1. These involved field recordings and concert recordings using the B-format signal from the Soundfield microphone to obtain a binaural reproduction for headphones. The reproduction of such content proved not to be straight forward and some high pass filtering was needed to make it sounds more natural, particularly in the case of the field recordings, as well as a choice for reproducing a smaller loudspeaker array configuration to avoid the 'phasing' effect. Recordings of concerts made indoors proved being very dependent on the reverberation of the environment where it was recorded and the positioning of the microphone in relation to the sound sources to improve externalization.

Another attempt still using only the Soundfield microphone was done in the recording of a choir concert, on the 29th of January of 2011, in the Chapter House of the York Minster, an environment more reverberant than the Sir Jack Lyons Concert Hall where the previous recordings were made, positioning the microphone in a place where direct sound and reverberant sound would be more balanced. In this recording the Soundfield microphone ST250 B-format signal was recorded on an Edirol R-4 four track portable recorder, in the same way as the preliminary field and concert recordings were done. The reproduction of this new material through the patch and with the head tracking system switched on proved to be more realistic than the previous recordings and also provided a reasonable level of externalization of the sound sources. The head tracking allowed good frontal localization cues. The better sensation of envelopment is attributed to the environment itself, the musical material (pure vocal music) and the microphone positioning.

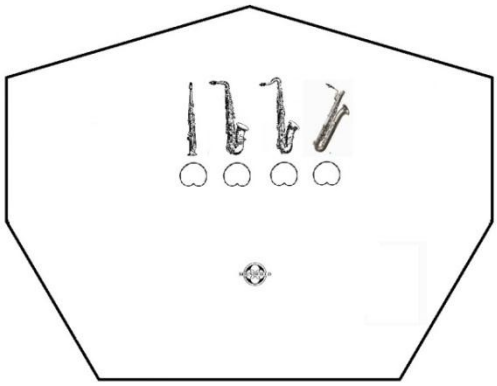




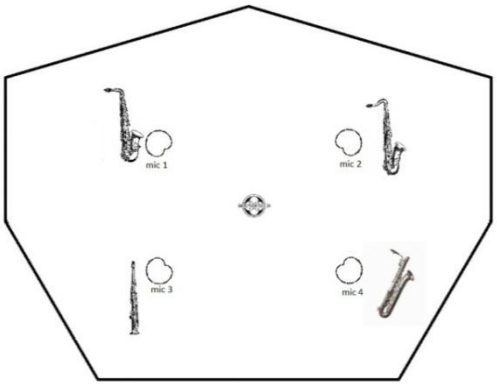
Positioning of the choir and the Soundfield microphone and picture of the Chapter House in the York Minster.

A series of multitrack recording sessions were performed in the Sir Jack Lyons Concert Hall, intending to test the capabilities of the 'virtual studio' developed during the mixing process, while being monitored through headphones. On the 14th of March of 2011, a saxophone quartet was recorded and two approaches to spatialization were adopted: the first one with the musicians playing as they would do in a stage performance and a second one as if they were surrounding the listener. The Soundfield microphone was positioned in the centre of each soundscape and spot microphones (Rode NT-1) placed close to the instruments. Four of the eight signals were recorded with an Edirol R-4 portable recorder and the other four with a M-box Pro 3 in ProTools software. After the recording, the signals were put together in Reaper software and mixed to a B-format 1st order file. A problem was faced while recording with two different systems: syncing the 8 recorded tracks proved not to be as simple as expected since no word clock or timecode sync signal was recorded simultaneously, and this problem added some unreliability on the final mix results.

Positioning the spot microphones signals into the three-dimensional soundscape established by the Soundfield microphone proved to be quite successful even through headphone monitoring. The second approach allowed it to be done more easily, probably due to the bigger separation of the sound sources (unfortunately a practice not welcomed by the musicians as it presented them with some difficulties as a result of playing far away from each other). The distance parameter for the spot microphones panners were the hardest to set, but this phenomenon is similar to mixing in stereo or in surround 5.1 through loudspeakers, so this proved that monitoring through headphones was close to the natural experience of other mixing processes. Some practice was needed to achieve good results and comparing mixing through headphones and loudspeakers was found to be essential to achieve a good final product.



Positioning scheme and picture of the first recording approach of the saxophone quartet, with the instruments in a frontal stage.



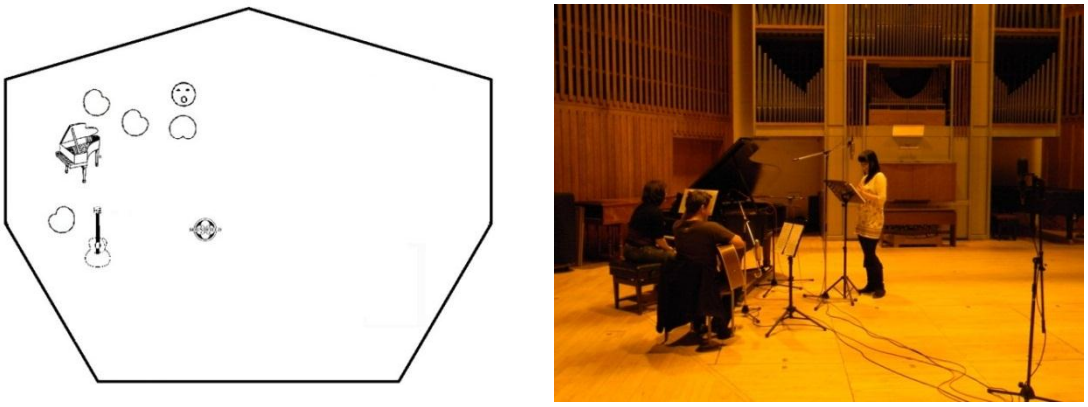
Positioning scheme and picture of the second recording approach of the saxophone quartet, with the instruments surrounding the listener.

The recording of a trio, composed by a singer, a piano and an acoustic guitar, was performed on the 19th of March of 2011 in the Sir Jack Lyons Concert Hall. The same two approaches used for the saxophone quartet recording session were also adopted in this recording, but the equipment used was changed to avoid the sync problem that occurred in the previous session. The Soundfield microphone was also positioned in the centre of both soundscapes, a pair of Rode NT-5 positioned inside the piano (pointing to the centre of the low and high strings), a Rode NT-1 in front of the singer and another Rode NT-1 near the guitar. The eight signals were routed through a small Behringer mixer, recorded using a MOTU 828 mkII audio interface and mixed in a Reaper digital audio workstation.

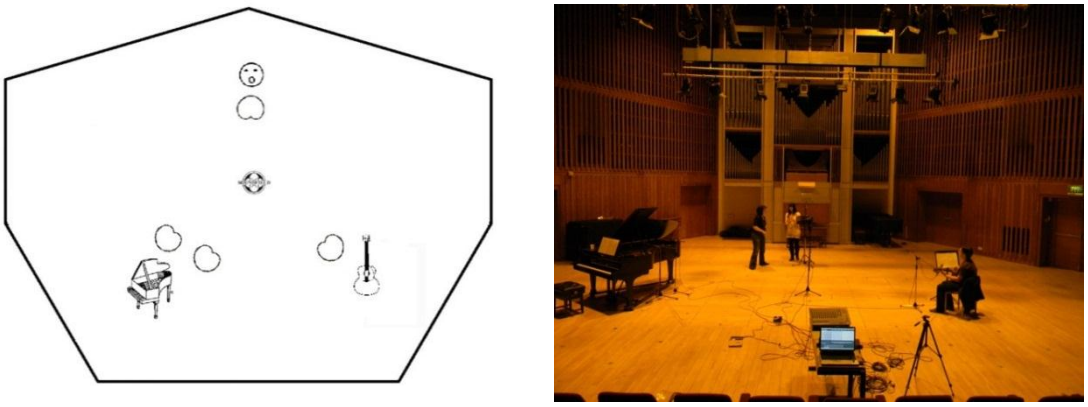
Some of the previous observations can also be applied to this experience but this recording session proved to be easier to mix than the previous one, probably due to the different instruments involved. The Soundfield microphone positioning was noticed to be very important in these processes where the sound engineer approaches the recording using the reverberant signal of such a microphone together with the direct sound from the spot microphones. If the Soundfield microphone is very close to the sound sources, capturing more direct sound than reverberated sound, the mixing process becomes more difficult and the results are not as satisfactory as in the case where the Ambisonic signal contains more reverberated sound than direct sound.

An interesting observation to make about all these recording sessions is that the parameters defined as the final mix, such as volume and panning parameters of the spot microphones signals, do not change by changing the loudspeaker array in the decoding stage, but the final result heard through the headphones for each setup is completely different, mainly in terms of tonal coloration that seems to be particular for each chosen array, as reported by other authors, for instance Fredriksson and Zacharov, 2002. One may ask which loudspeaker array is the best to be used and this question could not be answered in the present research due to its scope, which does not intend to analyse loudspeakers arrays properties and capabilities, but to reproduce all of them with their particularities in

such a way that the composer or engineer can choose between them before defining which one to use.



Positioning scheme and picture of the first recording approach of the trio, with the instruments positioned as in a frontal stage even in left side of the concert hall. It can be noticed that the front of the Soundfield microphone points to them with the piano in the centre.



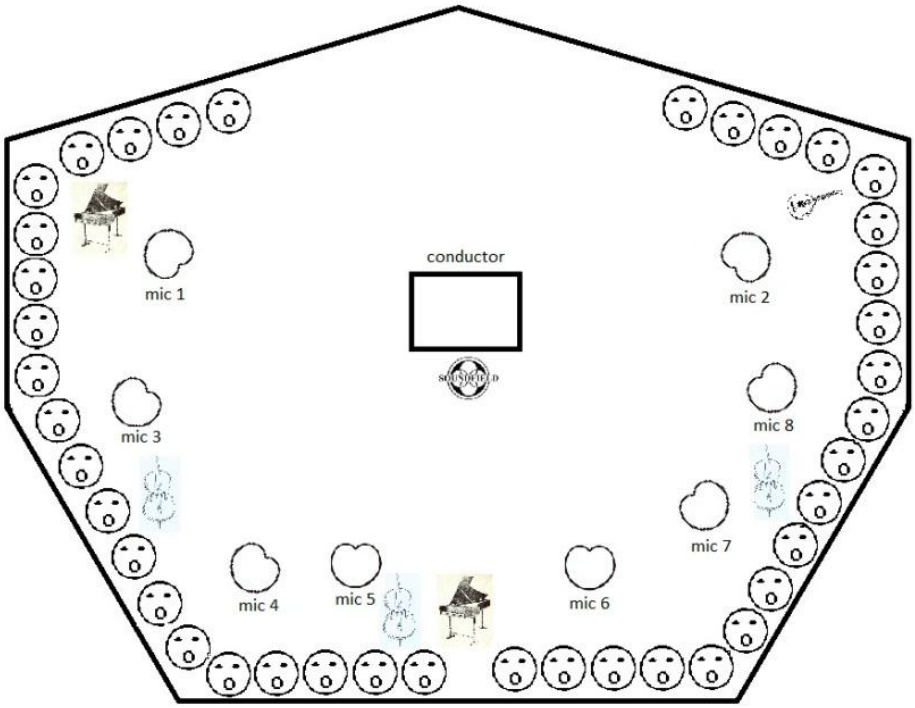
Positioning scheme and picture of the second recording approach of the trio, with the instruments surrounding the listener.

A recording of a larger group was made taking the opportunity of a workshop of the 40 parts mass from the 16th century, composed by Alessandro Striggio, performed by students of the music department of the University of York, on the 21st of June of 2011. Conducted by Robert Hollingworth, the piece portrays the listener centred approach as it was originally conceived at that period, an approach best known with the Thomas Tallis 'Spem

in Alium' 40 voices motet, which made this recording session an interesting opportunity to try the potential of the Ambisonic system out.

The Soundfield microphone was positioned in the centre of the soundscape and eight spot microphones around the choir and instruments. As illustrated by the following scheme, microphones 1 and 2 were Neumann U87, microphones 3, 7 and 8 were Calrec CM652, microphone 4 was a Calrec CM1050, and microphones 5 and 6 AKG C414. The four channels of the Soundfield microphone as well as the microphones 3, 4, 5 and 6 were routed through a Focusrite Octopre and then to the MOTU 828 mkII audio interface line inputs. The microphones 7 and 8 were connected straight to the MOTU microphone channels and the microphones 1 and 2 were routed through a MICO Audient pre amplifier and converter and then to the MOTU SPDIF inputs. The session was recorded and mixed in Reaper software.

In this session, positioning the spot microphone signals into the Ambisonic soundscape proved to be quite hard, probably due to the limitations of the equipment used, small number of spot microphones and the medium hall being not very coherent with such a big group. The musical material, consisted mainly of voices and just a few other instruments (two harpsichords, a lute and two cellos), would be better captured in a bigger hall in which the performers would be more away from the centre where the soundfield microphone was positioned, presenting more reverberant signal than direct signal. The final result, even when reproduced through headphones, was a massive envelopment reproduction of the group but with few details, which could be better achieved in a larger production with more microphones positioned closer to the instruments and groups of voices in an attempt to capture a more isolated direct sound (mainly of the instruments).

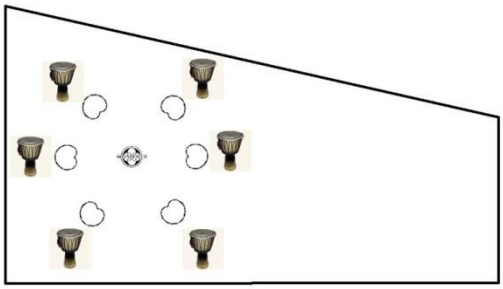


Positioning scheme of the Striggio 40 parts mass workshop, with the choir and instruments surrounding the listener.

Due to previous experiences with drums providing better localization cues, a drum session was planned with the West African Percussion Group of the University of York in the Trevor Jones studio, on the 28th of June of 2011. In this session, as in the previous, a soundscape approach with the instruments surrounding the listener was adopted and the Soundfield microphone positioned in the centre of the group. Six dynamic microphones (five Shure SM57 and an AKG D112) were positioned as spot microphones about 20 cm near the drums' skin. For some songs, in which one of the drums was substituted by a set of three bigger and lower drums, the SM57 microphone was replaced by a Rode NT-1 positioned a little further away from the drum set. The Soundfield microphone four signals and four of the spot microphones were routed through a Focusrite ISA 828 pre-amplifier and then to the MOTU HD192 audio interface, the other two spot microphones were routed through a Grace 201 pre amplifier and then to the MOTU interface. The session was recorded and mixed in Reaper software. The particular room where the group was recorded

allows the sound engineer to change the amount of reverberation in the recording by changing movable absorbers panels on the walls. In this case, all the panels of the room were completely closed intending to get the more reverberant response of the room.

As expected, the localization of this kind of instrument is better reproduced through headphones than the previous sessions' musical material, which made positioning the spot microphones signals into the soundscape very easy. This can be observed mainly when the players switch from playing the Djembe to the cowbells and shakers, due to their high frequency content. The soundscape obtained from the Soundfield microphone was, however, not as good as the previous experiences in the sense that the reverberation of the room where they were recorded is not as interesting as those of the previous sessions. The final result was a very localizable three dimensional environment but with a room coloration that will probably not be very appreciated by many listeners, as is not by the present author.



Positioning scheme and picture of the West African Percussion Group recording session, with the instruments surrounding the listener.

A sample of all the final mixes described above is presented in a DVD attached to this work as a 1st order B-format wav file, with the channels ordered as normal (W, X, Y and Z). The MAX patch developed is also on the DVD allowing readers to try out the binaural reproduction and compare it to the multichannel reproduction obtained from a straight reproduction of the 4 channels file, over a loudspeaker array. Orientation for installing the

necessary tools to use the patch is presented in a .txt file, also on the DVD.

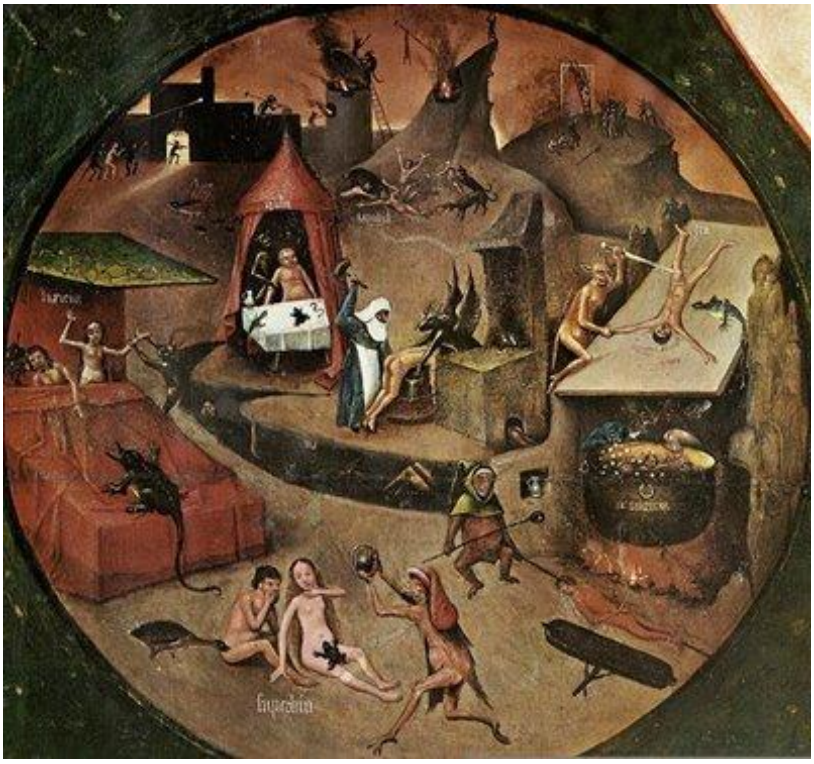
Although one of the objectives of this development is to make mixes that can work in any loudspeaker array, no deep investigation was performed aiming the modifications the choice of different loudspeaker array can bring. It was perceived that there are clear differences between them but those are not in the scope of this work, as have been of many other works, and leave opened another possibility for the engineer or composer to choose the array best fits to his ideal of multichannel reproduction.

Following are the description of the two first author's compositions focused on exploring the space in a musical and meaningful way. 'Meetings' – for three electric guitars, was intended to be performed live and explores the development of an extension of the well-known stereophonic guitar effect 'rotary' adapted for Ambisonic loudspeaker array with height. 'The Seven Sins' is a MAX patch designed to playback Gamelan recorded samples in a very well structured way full of rhetoric meanings. The submitted compositions exemplify thoughts on using the points in space from which particular sounds are played back in a musical way, 'pushing the boundaries' of contemporary multichannel electroacoustic music (Otondo, 2008), instead of just trying to enhance the listening experience or creating varied perspectives (Austin and Smaley, 2000) as have been observed in recent concerts and personal research. Choices made upon the use of Ambisonic encoding are due to the portability it offer the composer while performing pieces through different loudspeaker arrays (Austin and Field, 2001).

3.2.2 'The Seven Sins'

The performance of 'The Seven Sins' is intended to be played differently each time it is performed. This indeterminacy is ruled by a very well defined structure based on geometrical shapes the sound can draw in the space evoking rhetoric meaning not only due to the shapes themselves but also due to the sounds chosen and the numbers permeating the

whole piece. Inspired by other works with the same theme and quotations from ancient scriptures this algorithm composition tries to translate these ideas into sounds reproduced in space in a meaningful way.



'The Seven Deadly Sins' (Hieronymous Bosch) – oil on panel 1480-90 Prado, Madri. Take a look also at 'The Seven Deadly Sins of Modern Times' (Susan Dorothea White) – acrylic on wooden table 1993, see reference link.

Snow White and The Seven Sins (Judy Fox) – sculptures and exhibition. Extracted from Fox, 2007.

Seven things are disgusting to Him.

(The Holy Bible, Proverbs 6:16)

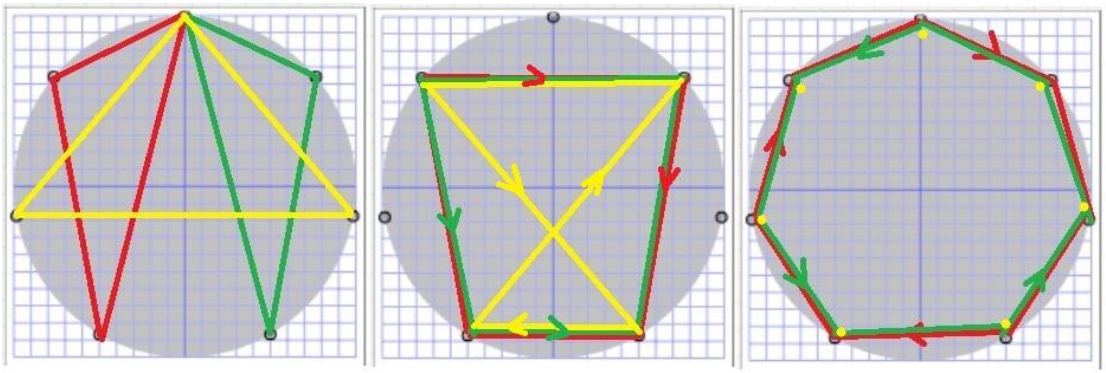
Divided into seven sections and played by what the composer called three 'virtual players', the piece evolved with the same 14 Gamelan samples divided into two groups with similar characteristics. Each time a 'player' starts playing it takes one of the two groups of samples. A random ordered set of samples is played each time in such a way that their organization can be considered serial in the sense that a sample never plays again before all the seven samples are played once by the same 'virtual player'. The sets of intervals between the samples are treated in a similar way. The seven points in space were chosen in such a way that one can draw three different triangles, a square, a circle and a seven tips star, and this set of positions are defined by the shapes each section is related to.

> (…) and the earth was waste and without form.
>
> (The Holy Bible, Genesis 1:2)

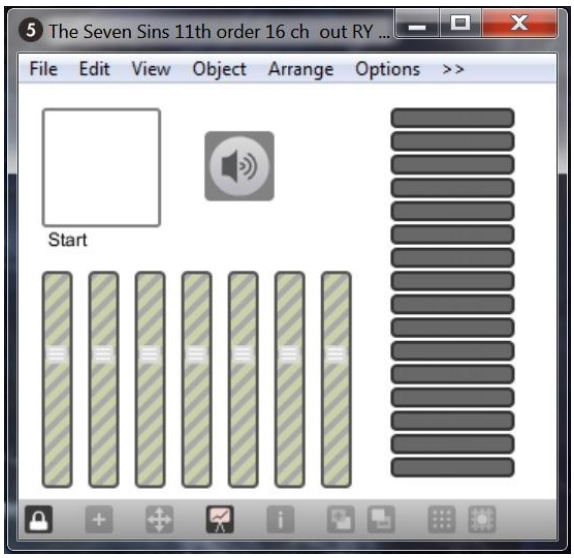
The 'players' start playing a section one at a time, and in the next section adding one to the other, the only exception is the first section when a certain chaos is intentionally reproduced. Each player has their own points in the space to play their notes and those are chosen according to a particular shape related to that particular section. The first section, however, has no particular shape and that fits with the chaos that everything began with. The second and third sections are based on three different sets of positions that create triangle shapes in the auditorium space, the fourth and fifth are based on three sets of position that creates square shapes and the sixth and seventh creates circle shapes.

> This square figure has not only an allusion to the progress of a craftsman in the science and arts; but more especially to the advancement of the good man in the paths of virtue and religion. His progress is said, by the wisest of men, to be as the shining light which shineth brighter and brighter unto the perfect day.
>
> (Akerman, 1875: 19)

This piece was conceived to be played through any horizontal loudspeaker array and the possibility of playing it back through a 5.1 system or even stereo proved a very interesting option, as well as the attempt at having a recorded version in binaural format. For spatialization, the ICST MAX externals were used.



Locations ‘player 1’ (red), ‘player 2’ (green) and ‘player 3’ (yellow) play in the space, according to the music sections – upper view.



'The Seven Sins' MAX patch screenshot.

While developing this piece, a variation of the ‘virtual studio’ patch, described in the previous section, was extremely useful. Due to the musical material used in the piece and the concept of developing it as an algorithmic piece, where the positions where the samples were played was the most important factor to be monitored, the whole process could be monitored through headphones even without the implementation of any reverb algorithm and the performance matches almost perfectly what was expected.

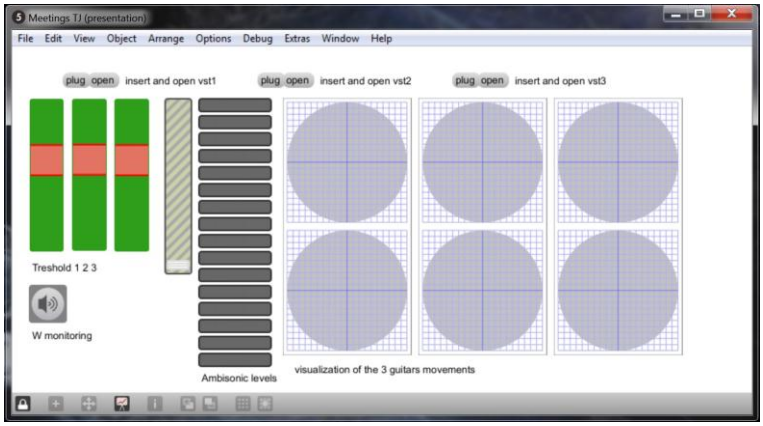
3.2.3 ‘Meetings’

To perform the composition ‘Meetings’, a MAX patch was developed using pitch and envelope recognition to define the signals trajectory through the sphere boundaries of the space reproduced by an Ambisonic system.

A ‘new development of harmony of space’ (Stockhausen, 1997) is approached where notes of specific pitches are played in specific heights in the auditorium space and the concept of envelopment is expanded (Rumsey, 2002; Austin and Smalley, 2000) in the sense that the movement of sounds through the boundaries of the sphere start from the frontal stage point and go circling the audience according to the duration between two attack detections.

The instruments, electric guitars, were chosen due to the ease its sound can be captured with no leakage or interference between the signals for processing purposes. The choice of having three players was established aiming at performing a counterpoint that covers the whole height of the auditorium from the bottom to the ceiling.

The piece itself is structured into three sections and favours the use of intervals of a second, most of times augmented or diminished. The first one present three main themes and starts in a fugue style followed by dissonant chords that compose the second theme, the third theme has thirds and forth intervals as main characteristic. A development can be observed when notes start to be played away from each other in time domain and try to speed up their start points to reach the other two. It is also divided into three very similar sections that follow the same principles. The third section is a retrograde movement of the first one (not very strict) that culminates on unison, in space, time and pitch domains, on the very long end note which is reached through an octave step.



'Meetings' MAX patch screenshot for live performance.

The ICST MAX externals were chosen to perform the spatialization due to their ease of use and flexibility while setting up loudspeakers and sound sources positions. Pitch detection is obtained from the yin~ MAX object and the envelope follower is based on peak amplitude detection. In the live performance patch, three threshold controllers can be adjusted according to the incoming signals. A master volume control for Ambisonic signals is available as well as 16 VU meters for the 16 loudspeaker 3rd order decoded signals. A visualization of the three guitar signal movements is also available to check if the correct movements are being performed and consequently adjust the effectiveness through the threshold controls. An optional VST plugin can also be inserted into the three guitar channels in such a way that run parallel to the signals directed to pitch and amplitude detection without disturbing them.

While monitoring the reproduction of this piece through headphones some interesting observations were made. The first one is that the elevation cannot be so easily recognized as the azimuth when monitoring through headphones and this observation matches with lots of other research that argue our capabilities of distinguishing elevation is inferior to distinguishing azimuth changes. In the binaural reproduction this seems more evident. Another observation is that, since two signals are performing the same trajectory in opposite sides of the loudspeaker array, it is really hard to perceive such movement in the binaural reproduction. This can be improved with head tracking especially if the listener

positions their head such that it is oriented around 30 degrees to one side, thus making the sound trajectories less similar. This piece needed to be reviewed by reproducing it through a loudspeaker array.

Conclusions and future work

A useful tool for working in an Ambisonic virtual environment was developed and put into practice through many recording sessions, two compositions developments and hearing tests with the participation of other music and technology students. The influence of many factors such as the music material, the individual experience and physiology, the loudspeaker array to be reproduced, among others, were understood and need to be taken into account by any music production or composition, as well as in the process of monitoring such production through headphones. The initial hypothesis that any loudspeaker array can be reproduced through headphones could not be confirmed since binaural systems proved to be very sensitive to the many elements that constitute the system and the process itself. However, monitoring larger loudspeakers systems through headphones proved to be very useful, even if not definitive.

Despite the lack of spatial resolution, first order four channel Ambisonic B-format recording and processing proved to be a very powerful tool for music productions so long as the composer or sound engineers are aware of its limitations and usability in the context of the final product objectives. Using Ambisonic B-format signal as an intermediate format to other formats, such as binaural, proved to be an interesting choice but limited by the system known limitations, to be adapted or considered in each production.

New tools for working with Ambisonic systems, such as delays and impulse response based reverberators, amongst others, are already being developed and presenting very interesting and promising results. The next step is to encourage the creation of more art with the already free available software, by making composers and sound engineers aware of them.

Further research needs to focus into reproducing the directivity pattern of loudspeakers and sound sources to simulate both of them through loudspeaker arrays or virtual environments reproduced through headphones and, consequently, improving them. High

order Ambisonics still needs to make available a practical microphone solution as well as a solution that can increase the spatial resolution but keep a low number of channels to be transmitted or stored. Hybrid techniques that can keep the use of only four Ambisonic channels as output and improve the spatial resolution would be very welcome. Some research also need to be done in respect to other psychoacoustic phenomena that influences the perception of the sound in space, as observed in section 1.2.3.

Reference List

3D60. Accessed at <http://www.3d60.co.uk/index.php> (on 06 Feb. 2011).

Aarts, R. M.; Schobben, D. W. E. 'Personalized multi-channel headphone sound reproduction based on active noise cancellation'. *Acta Acustica United with Acustica,* 91 (2005), 440-450.

Akerman, W. *The Bible and the square.* Montreal: Notre Dame Street, 1875. Facsimile accessed at <http://www.archive.org/stream/cihm_03243#page/n25/mode/2up> (on 27 April 2011).

Algazi, V. R.; Avendano, C.; Duda, R. O. 'Modeling the contralateral HRTF'. *Convention paper presented at the 16th AES convention*, Rovaniemi (Finland), 10-12 Apr 1999.

Algazi, V. R.; Angel, E. J.; Duda, R. O. 'On the design of canonical sound localization environments'. *Convention paper presented at the 113th AES convention*, Los Angeles (USA), 5-8 Oct 2002.

Algazi, V. R.; Duda, R. O.; Thompson, D. M. 'Motion Tracked Binaural Sound'. *J.Audio Eng. Soc.* 52 / 11 (2004), 1142-1156.

Ambisonic external for MAX/MSP. Accessed at <http://www.grahamwakefield.net/soft/ambi~/index.htm> (on 20 Jun 2011).

Ambisonic objects. Accessed at <http://www.york.ac.uk/music/mrc/software/objects/> (on 20 Jun 2011)

Ambisoncs - BBC soundfield experience. Accessed at <http://www.ambisonic.net/sfexp.html> (on 25 Jun 2011).

Amerson, T. L.; Desaulniers, D. H.; Kiefer, S. A.; Mershon, D. H. 'Visual capture in auditory distance perception: proximity image effect reconsidered'. *The journal of auditory research*, 20 (1980), 129-136.

Anderson, M. R.; Begault, D. R.; Wenzel, E. M. 'Direct comparison of the impact of head-tracking, reverberation, and individualized head related transfer functions on the spatial perception of a virtual speech source'. *J. Audio Eng. Soc.* 49 / 10 (2001), 904-916.

Angus, J. A. S.; Evans, M. J.; Tew, A. I. 'Analyzing head-related transfer function measurements using surface spherical harmonics'. *J. Acoust. Soc. Am.* 104 / 4 (1998), 2400-2411.

Ascension Flock of Birds. Accessed at <http://www.ascension-tech.com/realtime/RTflockofBIRDS.php> (on 25 Jun 2011).

Asselot, M.; Dellepiane, M.; Pietroni, N.; Scopigno, R.; Tsingos, N. 'Reconstructing head models from photographs for individualized 3D-audio processing', *Pacific Graphics*, 27 / 7 (2008).

Austin, L. and Field, A. 'Sound diffusion in composition and performance II: an interview with Ambrose Field' in *Organised Sound*, 25 / 4 (2001), 21-30.

Austin, L. and Smalley, D. 'Sound diffusion in composition and performance: an interview

with Denis Smalley', *Organised Sound*, 24 / 2 (2000), 10-21.

Avdelidis, K. A.; Dimoulas, C. A.; Kalliris, G. M.; Papanikolaou, G. V. 'Improved localization of sound sources using multi-band processing of Ambisonic components'. *Convention paper presented at the 126th AES convention*, Munich (Germany), 7-10 May 2009.

Azzali, A.; Cabrera, D.; Capra, A.; Farina, A.; Martignon, P. 'Reproduction of auditorium spatial impression with binaural and stereophonic sound systems'. *Convention paper presented at the 118th AES convention*, Barcelona (Spain), 28-31 May 2005.

B-mic VST plugin. Accessed at <http://www.york.ac.uk/inst/mustech/3d_audio/vst/bfmic_help.htm> (on 22 Jan. 2011).

Bamford, J. S.; Vanderkooy, J. 'Ambisonic sound for us'. *Convention paper presented at the 99th AES convention*, New York (USA), 6-9 Oct 1995.

Barrett, N.; Berge, S. 'A new method for B-format to binaural transcoding'. *Convention paper presented at the 40th AES International conference*, Tokyo (Japan), 8-10 Oct 2010.

Barton, G. J.; Gerzon, M. A. 'Ambisonic surround sound mixing for multitrack studios'. *Convention paper presented at the 2nd AES international conference*, California (USA), 11-14 May 1984.

Barton, G. J.; Gerzon, M. A. 'Ambisonic decoders for HDTV'. *Convention paper presented at the 92nd AES convention*, Vienna (Austria), 24-27 Mar 1992.

Batke, J.-M.; Keiler, F. 'Using VBAP derived panning functions for 3D ambisonics decoding'. *Procedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris (France), 6-7 May 2010.

Bauck, J.; Cooper, D. H. 'Generalized Transaural Stereo and Applications'. *J. Audio Eng. Soc.*, 44 / 9 (1996), 683-705.

Bauck, J.; Cooper, D. H. 'Prospects for Transaural Recording'. *J. Audio Eng. Soc.*, 37 / 1/2 (1989), 3-19.

Bauer, B. B. 'Stereophonic earphones and binaural loudspeakers'. *J. Audio Eng. Soc.*, 9 / 2 (1961), 148-151.

Baume, C.; Churnside, A. 'Upping the auntie: a broadcaster's take on Ambisonics'. *Convention paper presented at the 128th AES convention*, London (UK), 22-25 May 2010.

Begault, D. R. 'Challenges to the successful implementation of 3-D sound'. *J. Audio Eng. Soc.,* 39 / 11(1991), 864-870.

Begault, D. R. 'Auditory and non-auditory factors that potentially influence virtual acoustic imagery'. *Convention paper presented at the 16th AES international conference*, Rovaniemi (Finland), 10-12 Apr 1999.

Belar, H.; Olson, H. F. 'Acoustics of sound reproduction in the home'. *J. Audio Eng. Soc.,* 8 / 1 (1960), 7-11.

Bern, M.; Eppstein, D. 'Mesh generation and optimal triangulation', in *Computing in*

Euclidean geometry. Singapore: World Scientific, 1992.

Berry, S.; Lowndes, V.; Paterson-Stephens, I.; Wiggins, B. 'The design and optimisation of surround sound decoders using heuristic methods'. Accessed at <http://www.brucewiggins.co.uk/?page_id=3> (on 11th Jun 2011).

Bformat2Binaural. Accessed at <http://www.kvraudio.com/get/1685.html> (on 23 Jun 2011).

Binaural Panner. Accessed at <http://www.ece.ucdavis.edu/binaural/binaural_tools.html> (on 31 May 2011).

Binaural Room Simulator. Accessed at <http://iem.at/Members/noisternig/bin_ambi> (on 23 Jun 2011).

Blake, Randolph; Sekuler, Robert. *Perception*. New York: McGraw Hill, 2006.

Blauert, J. *Räumlisches Hören.* Stuttgart: Hirzel Verlag, 1974. Repr. Trans. *Spatial Hearing: the Psychophysics of Human Sound Localization.* London: MIT Press, 1983, 1997.

Blauert, J. 'Sound localization in the median plane'. *Acoustica*, 22 (1969/1970), 205-213.

Blue ripple sound. Accessed at <http://www.blueripplesound.com/> (on 31 May 2011).

Blumlein, A. D. 'Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems'. UK Patent GB 394325, 1933.

Boland, F.; Gorzel, M.; Kearney, G.; Rice, H. 'Depth perception in interactive virtual acoustic environments using higher order ambisonic soundfields'. *Procedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Pairs (France), 6-7 May 2010.

Bosch, Hieronymous. *The Seven Deadly Sins*. Oil on panel. Prado, Madri 1480-90. Accessed at <http://awhitecarousel.com/2010/the-art-of-hieronymus-bosch/> (on 28 April 2011).

Braasch, J.; Leung, K. L.; Leung, N. M.; So, R. H. Y. 'A low cost, non-individualized surround sound system based upon head related transfer functions: an ergonomics study and prototype development'. *Applied Ergonomics,* 37 (2006), 695-707.

Braasch, J.; Matthews, T.; McAdams, S.; Peters, N. 'VIMIC – a novel toolbox for spatial sound processing in MAX/ MSP'. *International computer music conference proceedings*, 2008.

Bresson, J.; Schumacher, M. 'Compositional control of periphonic sound spatialization'. *Procedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris (France), 6-7 May 2010.

Bridge, H.; Jagger, D. S. 'Recent developments and improvements in soundfield microphone technology'. *Convention paper presented at the 75th AES convention*, Paris (France), 27-30 Mar, 1984.

Bronkhorst, A. W. 'Localization of real and virtual sound sources'. *J. Acoust. Soc. Am.*, 98

/ 5 (1995), 2542-2553.

Brookes, T.; Kim, C.; Mason, R. 'An investigation into head movements made when evaluating various attributes of sound'. *Convention paper presented at the 122nd AES convention*, Vienna (Austria), 5-8 May 2007.

Bruno, R.; Laborie, A.; Montoya, S. 'A new comprehensive approach of surround sound recording'. *Convention paper presented at the 114th AES convention*, Amsterdam (The Netherlands), 22-25 Mar 2003.

Cába, C. 'Surround audio that lasts: future proof Ambisonic recording and processing technique for the real world'. *Convention paper presented at the 112th AES convention*, Munich (Germany), 10-13 May 2002.

Caroll, Lewis. *Symbolic Logic*. 1897. Accessed at <http://www.gutenberg.org/ebooks/28696> (on 31 May 2011).

Carpentier, T.; Nguyen, K.-V.; Noisternig, M.; Warusfel, O. 'Calculation of head related transfer functions in the proximity region using spherical harmonics decomposition: comparison with measurements and evaluation'. *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris (France), 6-7 May. 2010.

CATT Acoustics. Accessed at <http://www.catt.se/> (on 04 Jul 2011).

Clark, D. F.; Horsburgh, A. J. 'Ambisonic decoders: is historical hardware the future?' *Convention paper presented at the 128th AES convention*, London (UK), 22-25 May 2010.

Chen, F. ‘The reaction time for subjects to localize 3D sound via headphones’. *Convention paper presented at the 22nd AES international conference*, Espoo (Finland), 15-17 Jun 2002.

Cheng, C. I; Wakefield, G. H. ‘Moving sound sources synthesis for binaural electroacoustic music using interpolated head-related transfer functions (HRTFs)’. *Computer Music Journal* 25 / 4 (2001), 57-80.

Cooper, D. H.; Shiga, T. ‘Discrete-matrix multichannel stereo’ *J. Audio Eng. Soc.* 20 / 6 (1972), 346-360.

Craven, P. G.; Gerzon, M. ‘Coincident microphone simulation covering three dimensional space and yielding various directional outputs’. US patent 4042779, 1977.

CREATE Signal Library. Accessed at <http://fastlabinc.com/CSL/index.html> (on 20 Jun 2011).

Dalenbäck, B.-I.; Kleiner, M.; Svensson, P. ‘Auralization – an overview’. *J. Audio Eng. Soc.* 41 / 11 (1993), 861-875.

Dalenbäck, B.-I.; McGranth, D. S.; Reilly, A. ‘Using auralisation for creating animated 3-D sound fields across multiple speakers’. *Convention paper presented at the 99th AES convention*, New York (USA), 6-9 Oct 1995.

Damasle, P.; Mellert, V. ‘Ein Verfahren zur riáhtungstreuen Schallabbildung des oberen halbraumes über zwei Lautsprechern’. *Acustica,* 22 (1969), 154-162.

Daniel Courville Ambisonic Studio. Accessed at <http://www.radio.uqam.ca/ambisonic/> (on 21 Jun 2011).

Daniel, J.; Rault, J.-B.; Polack, J.-D. 'Ambisonics encoding of other audio formats for multiple listening conditions'. *Convention paper presented at the 105th AES convention*, San Francisco (USA), 26-29 Sep 1998.

Daniel, J. 'Représentation de Champs Acoustiques, Application à la Transmissionet à la Reproduction de Scènes Sonores Complexes dans un Contexte Multimédia'. Ph.D. diss., Université de Paris, 2000.

Daniel, J. 'Spatial sound encoding including near field effect: introducing distance coding filters and a viable, new Ambisonic format'. *Convention paper presented at the 23rd AES international conference*, Copenhagen (Denmark), 23-25 May 2003.

Eickmeier, G. C. 'An image model theory for stereophonic sound'. *Convention paper presented at the 87th AES convention,* New York (USA), Oct 1989.

Eickmeier, G. C. 'Comments on the distinction between stereophonic and binaural sound'. Letters to the editor. *J. Audio Eng. Soc.*, 39 / 4 (1991), 261-266.

ep.binSpat~ 0.16. Accessed at <http://www.eude.nl/maxmsp/> (on 22 Jun 2011).

FaceAPI. Accessed at <http://www.seeingmachines.com/product/faceapi/downloads/> (on 24 Jun 2011).

Faure, J. 'Les systèmes de head tracking'. Technical report, France Telecom, 2004.

Felderhoff, U.; Horbach, U.; Mackensen, P.; Pellegrini, R.; Theile, G. 'Binaural room scanning – a new tool for acoustic and psychoacoustic research'. 1999. <http://www.hauptmikrofon.de/theile/BRS_DAGA_1999_Paper.PDF> (accessed 09 Jan. 2011).

Fels, J.; Masiero, B. 'Perceptually robust headphone equalization for binaural reproduction'. *Convention paper presented at the 130th AES convention*, London (UK), 13-16 May 2011.

Fiducial symbols. Accessed at <http://reactivision.sourceforge.net/data/fiducials.pdf> (on 26 Jun 2011).

Fox, Judy. *Snow White and The Seven Sins*. Sculptures and exhibition. New York, USA. 2007. Accessed at <http://www.artnet.com/magazineus/features/kuspit/kuspit11-21-07.asp> (on 28 April 2011).

Frank, M.; Sontacchi, A.; Zotter, F. 'Localization experiments using different 2D Ambisonics decoders'. *25th Tonmeistertagung – VDT international convention*, Leipzig (Germany), 13-18 Nov 2008.

Fredriksson, M.; Zacharov, N. 'Natural reproduction of music and environmental sounds'. *Convention paper presented at the 112th AES convention*, Munich (Germany), 10-13 May 2002.

Gan, W.-S.; Tan, C.-J. 'Direct concha excitation for introduction of individualized hearing cues'. *J. Audio Eng. Soc.* 48 / 7/8 (2000), 642-653.

Geier, M.; Raake, A.; Spors, S.; Wierstorf, H. 'A free database of head-related impulse response measurements in the horizontal plane with multiple distances'. *Convention paper presented at the 130th AES convention,* London (UK), 13-16 May 2011.

George, S.; Neugebauer, B.; Plogsties, J.; Silzle, A. 'Binaural processing algorithms: importance of clustering analysis for preference tests'. *Convention paper presented at the 126th AES convention,* Munich (Germany), 7-10 May 2009.

Gerzon, M. A. 'Periphony (with height sound reproduction)'. *J. Audio Eng. Soc.* 21 /1 (1973), 2-10.

Gerzon, M. A. 'The design of precisely coincident microphone array for stereo and surround sound'. *Convention paper presented at the 50th AES convention*, London (UK), Mar 1975a.

Gerzon, M. A. 'N. R. D. C. Ambisonic Technology Report n°3 - Pan pot and field controls'. Unpublished technical report, 1975b.

Gerzon, M. A. 'N. R. D. C. Ambisonic Technology Report n°4 - Artificial reverberation and spreader devices'. Unpublished technical report, 1975c.

Gerzon, M. A. 'Practical periphony: the reproduction of full sphere sound'. *Convention paper presented at the 65th AES convention*, London (UK), 25-28 Feb 1980.

Gerzon, M. A. 'Ambisonics in Multichannel Broadcasting and Video'. *J. Audio Eng. Soc.* 33 / 11 (1985), 859-871.

Gerzon, M. A. ‘General metatheory of auditory localisation’. *Convention paper presented at the 92nd AES convention*, Vienna (Austria), 24-27 Mar 1992.

Gerzonic DecoPro. Accessed at <http://www.gerzonic.net/> (on 21 Jun 2011).

Gierlich, H. W. ‘The application of binaural technology’. *Applied Acoustics*, 36 (1992), 219-243.

Goodwin, M. M.; Jot, J.-M. ‘Binaural 3-D audio rendering based on spatial audio scene coding’. *Convention paper presented at the 123rd AES convention*, New York (USA), 5-8 Oct 2007.

Griesinger, D. ‘Binaural technique for music reproduction’. *Convention paper presented at the 8th AES convention*, Washington D.C. (USA), May 1990.

Hammershoi, D. ‘Fundamental aspects of the binaural recording and synthesis techniques’. *Convention paper presented at the 100th AES convention*, Copenhagen (Denmark), 11-14 May 1996.

Hammershoi, D.; Jensen, C. B.; Moller, H.; Sorensen, M. F. ‘Transfer characteristics of headphones’. *Convention paper presented at the 92nd AES convention*, Vienna (Austria), 24-27 Mar 1992.

Hammershoi, D.; Jensen, C. B.; Moller, H.; Sorensen, M. F. ‘Transfer characteristics of headphones measured on human ears’. *J. Audio Eng. Soc.*, 43 / 4 (1995), 203-217.

Hammershoi, D.; Jensen, C. B.; Moller, H.; Sorensen, M. F. 'Binaural technique: do we need individual recordings?' *J. Audio Eng. Soc.*, 44 / 6 (1996a), 451-469.

Hammershoi, D.; Jensen, C. B.; Moller, H.; Sorensen, M. F. 'Using a typical human subject for binaural recording'. *Convention paper presented at the 100th AES convention*, Copenhagen (Denmark), 11-14 May 1996b.

Han, H. L. 'On the relation between directional bands and head movements'. *Convention paper presented at the 92nd AES convention*, Vienna (Austria), 24-27 Mar 1992.

Härmä, A.; Karjalainen, M.; Tikander, M. 'Acoustic positioning and head tracking based on binaural signals'. *Convention paper presented at the 116th AES convention*, Berlin (Germany), 8-11 May 2004.

Harpex VST plugin. Accessed at <http://www.harpex.net/> (on 21 Jun 2011).

Hartmann, W. M.; Wittenberg, A. 'On the externalization of sound images'. *J. Acoust. Soc. Am.*, 99 (1996), 3678-3688.

Head in space. Accessed at <http://www.lim.dico.unimi.it/HiS> (on 24 Jun 2011).

Head tracking for desktop virtual reality displays using the wii remote. Accessed at <http://johnnylee.net/projects/wii/> (on 25 Jun 2011).

Hebrank, J.; Wright, D. 'Spectral cues used in localization of sound sources on the median plane'. *J. Acoust. Soc. Am.,* 56 (1974), 1829-1834.

Held, R. ‘Shifts in binaural localization after prolonged exposures to atypical combinations of stimuli’. *The American Journal of Psychology,* 68 / 4 (1955), 526-548.

Hess, W. ‘Influence of head tracking on spatial perception’. *Convention paper presented at the 117th AES convention*, San Francisco (USA), 28-31 Oct 2004.

Höldrich, R.; Lorenz, R.; Musil, T.; Noisternig, M.; Sontacchi, A. ‘Hearing Varèse's Poème Électronique inside a virtual Philips pavilion’. *Proceedings of ICAD 05-Eleventh meeting of the international conference in auditory display*, Limerick (Ireland), 6-9 Jul 2005a.

Höldrich, R.; Musil, T.; Noisternig, M.; Sontacchi, A. ‘A 3D ambisonic based binaural sound reproduction system’. *Convention paper presented at the 24th AES international conference*, Banff (Canada), 26-28 Jun 2003a.

Höldrich, R.; Musil, T.; Noisternig, M.; Sontacchi, A. ‘A 3D real time rendering engine for binaural sound reproduction’. *Proceedings of the 2003 International Conference on Auditory Display*, Boston (USA), 6-9 Jul 2003b.

Höldrich, R.; Musil, T.; Noisternig, M.; Sontacchi, A. ‘3D binaural sound reproduction using a virtual ambisonic approach’. *Convention paper presented at the International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, Lugano (Switzerland), 27-29 Jul 2003c.

Höldrich, R.; Musil, T.; Zmölnig, J. M.; Zouhar, V. ‘Virtual audio reproduction engine for spatial environments’. *International computer music conference proceedings*, 2005b.

Holman, T.; Kyriakakis, C. ‘Video based head tracking for improvements in multichannel loudspeaker audio’. *Convention paper presented at the 105th AES convention*, San Francisco (USA), 26-29 Sep 1998.

Hope, Adrian. ‘From rubber disc to magnetic tape’. *New Scientist,* 77 / 1085 (1978), 96-97. Accessed at <http://books.google.com/books> (on 31 May 2011).

Horbach, U.; Karamustafaoglu, A.; Mackensen, P.; Pellegrini, R.; Theile, G. ‘Design and applications of a data-based auralization system for surround sound’. *Convention paper presented at the 106th AES convention*, Munich (Germany), 8-11 May 1999.

Huron Technical Manual. Accessed at <http://www.sheffield.ac.uk/content/1/c6/01/79/48/HuronMan32.pdf> (on 21 Jun 2011).

ICST Ambisonic external for MAX website. Accessed at <http://www.icst.net/research/downloads/ambisonics-externals-for-maxmsp/> (on 27 Apr 2011).

IEEEghn website. ‘Alan Dower Blumlein’. *IEEE Global History Network,* 2009. Accessed at <http://www.ieeeghn.org/wiki/index.php/Alan_Dower_Blumlein> (on 31 May 2011).

InertiaCube – InterSense. Accessed at <http://www.intersense.com/categories/18/> (on 25 Jun 2011).

Impulse response measurements. Accessed at <https://dev.qu.tu-berlin.de/projects/measurements/files> (on 23 Jun 2011).

Jack audio router. Accessed at <http://jackaudio.org> (on 27 Feb. 2011).

Jin, C.; Martin, A.; Schaik, A. van. 'Psychoacoustic evaluation of systems for delivering spatialized augmented reality audio'. *J. Audio Eng. Soc.*, 57 / 12 (2009), 1016-1027.

Jot, J.-M.; Larcher, V.; Pernaux, J.-M. 'A comparative study of 3-D audio encoding and rendering techniques'. *Convention paper presented at the 16th AES international conference,* Rovaniemi (Finland), 10-12 Apr 1999.

Jot, J.-M.; Larcher, V.; Vandernoot, G. 'Equalisation methods in binaural technology'. *Convention paper presented at the 105th AES convention*, San Francisco (USA), 26-29 Sep 1998a.

Jot, J.-M.; Larcher, V.; Wardle, S. 'Approaches to binaural synthesis'. *Convention paper presented at the 105th AES convention*, San Francisco (USA), 26-29 Sep 1998b.

Jot, J.-M.; Larcher, V.; Warusfel, O. 'Digital signal processing issues in the context of binaural and transaural stereophony'. *Convention paper presented at the 98th AES convention*, Paris (France), 25-28 Mar 1995.

Jot, J.-M.; Philip, A.; Walsh, M. 'Binaural simulation of complex acoustic scenes for interactive audio'. *Convention paper presented at the 121st AES convention*, San Francisco (USA), 5-8 Oct 2006.

Jot, J.-M.; Warusfel, O. 'Spat~: a spatial processor for musicians and sound engineers'. CIARM: International Conference on Acoustics and Music Research, Ferrara (Italy), 1995. Accessed at <http://articles.ircam.fr/textes/Jot95a/> (on 03 May 2011).

Karamustafaoglu, A.; Spikofski, G. ‘Binaural room scanning and binaural room modelling’. *Listening demonstration at the 19th AES convention*, Bavaria (Germany), 21-24 Jun 2001.

Kendall, G. S. ‘A 3-D Sound Primer: Directional Hearing and Stereo Reproduction’. *Computer Music Journal*, 19 / 4 (1995), 23-46.

Kistler, D. J.; Wightman, F. L. ‘Headphone simulation of free-field listening II: Psychophysical validation’. *J. Acoust. Soc. Am.*, 85 /5 (1989), 868-878.

KLT3D. Accessed at <http://gpa.lps.ufrj.br/index.php/pt_BR/KLT3D-VST> (on 23 Jun 2011).

Kocher, P.; Schacher, J. C. ‘Ambisonics spatializations tools for MAX/ MSP’. *International computer music conference proceedings*, 2006.

Lindau, A.; Hohn, T.; Weinzierl, S. ‘Binaural resynthesis for comparative studies of acoustical environments’. *Convention paper presented at the 122nd AES convention*, Vienna (Austria), 5-8 May 2007.

LISTEN HRTF Database. Accessed at <http://recherche.ircam.fr/equipes/salles/listen/> (on 23 Jun 2011).

Ludovico, L. A.; Mauro, D. A.; Pizzamiglio, D. ‘Head in space: a head tracking based binaural spatialization system’. 2010. <http://smcnetwork.org/files/proceedings/2010/54.pdf> (accessed 09 Jan. 2011).

Lyon, E. ‘Conditions for development of an interchange format for spatial audio’. *International computer music conference proceedings*, 2008.

Malham, D. G. ‘Computer control of Ambisonic soundfield’. *Convention paper presented at the 82nd AES convention*, London (UK), 10-13 Mar 1987.

Malham, D. G. ‘Ambisonics – a technique for low cost, high precision, three dimensional sound diffusion’. *International computer music conference proceedings*, Glasgow (UK), 1990, 118-120

Malham, D. G. ‘Experience with large area 3D Ambisonics sound system’. *Proceedings of the Institute of Acoustics* 14 / 5 (1992), 209-216.

Malham, D. G. ‘3D sound for virtual reality systems using Ambisonic techniques’. Presented at the 4th Virtual Reality conference (VR93), London (UK), Apr 1993. Accessed at <http://www.york.ac.uk/inst/mustech/3d_audio/vr93papr.htm> (on 9 Mar 2011).

Malham, D. G. ‘Ambisonics in the new media – the technology comes of age’. *Convention paper presented at the 11th AES UK conference*, London (UK), 25-26 Mar 1996.

Malham, D. G. ‘Approaches to spatialisation’. *Organised Sound* 3 / 2 (1998a), 167-177.

Malham, D. G. ‘The role of the single point soundfield microphone in surround sound systems’. *Convention paper presented at the 13th UK AES regional conference*, London (UK), Mar 1998b.

Malham, D. G. ‘Higher order Ambisonic systems for the spatialisation of sound’. *International computer music conference proceedings*, 1999a.

Malham, D. G. ‘Homogeneous and non-homogeneous surround sound systems’. *Proceeding of the AES 14th UK conference: Audio – the second century*, London (UK), 7-8 Jun 1999b.

Malham, D. G. ‘Spherical harmonic coding of sound objects – the ambisonic 'O' format’. *Convention paper presented at the 19th AES international conference*, Bavaria (Germany), 21-24 Jun 2001a.

Malham, D. G. ‘Toward reality equivalence in spatial sound diffusion’. *Computer Music Journal* 25 / 4 (2001b), 31-38.

Malham, D. G.; Myatt, T. ‘3-D sound spatialization using Ambisonic techniques’. *Computer Music Journal* 19 / 4 (1995): 58-70.

Malham, D. G.; Orton, R. ‘Progress in the application of 3-dimensional Ambisonic sound systems to computer music’. *International computer music conference proceedings*, 1991.

Manning, P. *Electronic and computer music*. Oxford University Press, 2004.

McGrath, D.; McKeag, A. 'Sound field format to binaural decoder with head tracking'. *Convention paper presented at the 6th Australian regional AES convention*, Melbourne (Australia), 10-12 Sep 1996.

McGrath, D.; Reilly, A. ‘Real time auralization with head tracking’. *Convention paper*

presented at the 5th AES Australian regional convention, Sydney (Australia), 26-28 Apr 1995.

McGrath, D.; Reilly, A. 'A suite of tools for creation, manipulation and playback of soundfields in the huron digital audio convolution workstation'. *Convention paper presented at the 100th AES convention*, Copenhagen (Denmark), 11-14 May 1996.

Minnaar, P.; Pedersen, J. A. 'Evaluation of a 3D audio system with head tracking'. *Convention paper presented at the 120th AES convention*, Paris (France), 20-23 May 2006.

Moller, H. 'Fundamentals of binaural technology'. *Applied Acoustics,* 36 (1992), 171-218.

Moller, H. 'Reproduction of artificial head recordings through loudspeakers'. *J. Acoust. Soc. Am.*, 37 / 1/2 (1989), 30-33.

Monro, G. 'In-phase corrections for Ambisonics'. *International computer music conference proceedings*, 2000.

Moore, D.; Wakefield, J. 'Optimisation of the localisation performance of irregular Ambisonic decoders for multiple off centre listeners'. *Convention paper presented at the 128th AES convention*, London (UK), 22-25 May 2010.

Movement sensor device for head tracking. Accessed at <http://vimeo.com/22727528> (on 14 Jun 2011).

Musil, T.; Ritsch, W.; Sontacchi, A.; Zmölnig, J.; Zotter, F. 'Remote 3D audio performance with spatialized distribution'. *International computer music conference proceedings*, 2008.

Murphy-Chutorian, E.; Trivedi, M. M. 'Head pose estimation in computer vision: a survey'. *IEEE Transactions on pattern analysis and machine intelligence.* 31 / 4 (2009), 607-626.

Nettingsmeier, J. 'General purpose ambisonic playback systems for electroacoustic concerts – a practical approach'. *Procedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris (France), 6-7 May 2010.

Neukom, M. 'Ambisonic Panning'. *Convention paper presented at the 123rd AES convention*, New York (USA), 5-8 Oct 2007.

Neukom, M.; Schacher, J. C. 'Ambisonic equivalent panning'. *International computer music conference proceedings*, 2008.

Nicol, Rozenn. 'Binaural Technology'. AES Monograph, 2010.

Nicol, R.; Moore, A. H.; Tew, A. 'Headphone transparification: a novel method to investigating the externalisation of binaural sounds'. *Convention paper presented at the 123rd AES convention,* New York (USA), 5-8 Oct 2007.

Offenhauser, W. H. Jr. 'Binaural and stereophonic sound in retrospect'. *J. Audio Eng. Soc.,* 6 / 2 (1958), 67-69.

Otondo, F. 'Contemporary trends in the use of space in electroacoustic music'. *Organised Sound,* 13 / 1 (2008), 77-81.

Phoenix technologies. Accessed at <http://www.ptiphoenix.com/Products.php> (on 25 Jun 2011).

Plack, Christopher J. 'Spatial hearing', in: *The Sense of Hearing.* England: Lawrence Erlbaum Associates, 2005.

Poletti, M. A. 'Three-Dimensional surround sound system based on spherical harmonics'. *J. Audio Eng. Soc.* 53 / 11 (2005), 1004-1025.

Poletti, M. A. 'Robust two-dimensional surround sound reproduction for nonuniform loudspeaker layouts'. *J. Audio Eng. Soc.* 55 / 7/8 (2007), 598-610.

Polhemus Innovation in motion. Accessed at <http://www.polhemus.com/> (on 24 Jun 2011).

Pollow, M. 'Applying extrapolation and interpolation methods to measured and simulated HRTF data using spherical harmonic decomposition'. *Procedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris (France), 6-7 May 2010.

Postal, J. 'Audio aspects of the SMPTE 74th convention – a review'. *J. Audio Eng. Soc.,* 2 / 2 (1954), 119-142.

Pulkki, V. 'Virtual sound source positioning using vector base amplitude panning'. *J. Audio Eng. Soc.* 45 / 6 (1997), 456-466.

Pulkki, V. 'Generic panning tools for MAX/ MSP'. *International computer music conference proceedings*, 2000.

Pulkki, V. ‘Evaluating spatial sound with binaural auditory model’. *International computer music conference proceedings*, 2001.

Ramakrishnan, C. ‘Zirkonium: non-invasive software for sound spatialisation’. *Organised Sound.* 14 / 3 (2009), 268-276.

ReacTable. Accessed at <http://www.reactable.com/> (on26 Jun 2011).

ReacTIVision. Accessed at <http://reactivision.sourceforge.net/> (on 26 Jun 2011).

Reaper DAW software. Accessed at <http://www.reaper.fm/download.php> (on 12 march 2012).

Rumsey, Francis. *Spatial Audio.* Oxford: Focal Press, 2001.

Rumsey, Francis. ‘Spatial quality evaluation for reproduced sound: terminology, meaning and scene-based paradigm’. *J. Audio Eng. Soc.* 50 / 2 (2002), 651-666.

Seeber, B. ‘A new method for localization studies’. *Acta Acoustica*, 88 (2003), 446-450.

Shewchuk, J. R. ‘Triangle: engineering a 2D quality mesh generator and Delaunay triangulator’, in *Applied computational geometry: towards geometric engineering*. Berlin: Springer-Verlag, 1996.

SIR VST plugin. Accessed at <http://www.knufinke.de/sir/sir1.html> (on 05 Jul 2011).

Sound externalization headphone. Sursound email list posts, 26 May 2011.

Soundfield Technology. Accessed at
<http://www.soundfield.com/downloads/b_format.pdf> (on 03 Apr 2012).

Soundflower. Accessed at <http://code.google.com/p/soundflower/> (on 27 Feb. 2011).

Sousa, F. W. J. 'Mixagem em surround 5.1 e transcrição para audição binaural de músicas eletroacústicas espacializadas: o caso de "4 sketches em movimento"'. *Convention paper presented at the 8th AES brazilian convention*, São Paulo (Brazil), 4-6 May 2010.

Sousa, F. W. J. 'Subjective comparison between stereo and binaural processing from B-format Ambisonic raw audio material'. *Convention paper presented at the 130th AES convention*, London (UK), 13-16 May 2011.

Snow, W. B. 'Basic principles of stereophonic sound'. *IRE transactions on Audio*, 3 / 2 (1955), 42-53.

SPS200 Surround Zone software. Accessed at
< http://www.soundfield.com/products/sps200_s_zone.php > (on 22 Jun 2011)

Stern, Richard M., Brown, Guy J. and Wang, Deliang. 'Binaural Sound Localization', in *Computational Auditory Scene Analysis – Principles, Algorithms and Applications*. New Jersey: IEEE Press Wiley Interscience, 2006.

Stockhausen, K. interview broadcast in 'Access All Areas' Paul Bronowsky, presenter. ABC television, 4 May 1997. Australia.

Sunier, J. *The story of stereo: 1881-*. Gernsback Library, Inc. New York (USA), 1960.

Swiss center for computer music VST plugins. Accessed at <http://www.dmalham.freeserve.co.uk/vst_ambisonics.html> (on 30 Jan. 2011).

The Eigenmike microphone array. Accessed at <http://www.mhacoustics.com/mh_acoustics/Eigenmike_microphone_array.html> (on 19 May 2011).

The Holy Bible. 'Genesis' and 'Proverbs'.

The Magical Tunnel. Accessed at <http://www.futurestories.ca/osc/index.html> (on 26 Jun 2011).

Toole, F. E. 'Binaural record / reproduction systems and their use in psychoacoustic investigations'. *Convention paper presented at the 91st AES convention*, New York (USA), 4-8 Oct 1991.

Toronto Museum Project. Accessed at <http://www.futurestories.ca/toronto/> (on 26 Jun 2011).

Travis, C. 'A virtual reality perspective on headphone audio'. *Convention paper presented at the 11th AES UK conference*, London (UK), 25-26 Mar 1996a.

Travis, C. 'Virtual reality perspective on headphone audio'. *Convention paper presented at the 101st AES convention*, Los Angeles (USA), 8-11 Nov 1996b.

Universal internal audio routing for Allocator / Alloclite. Accessed at <http://www.thuneau.com/forum/viewtopic.php?t=168> (on 03 May 2011).

VBAP MAX/MSP tools. Accessed at <http://www.acoustics.hut.fi/software/vbap/MAX_MSP/> (on 20 Jun 2011).

Vicon. Accessed at <http://www.vicon.com/products/> (on 25 Jun 2011).

Wakefield, G. 'Third order Ambisonic extensions for MAX/ MSP with musical applications'. *International computer music conference proceedings*, 2006, 123-126.

Wallach, H. 'On sound localization'. *J. Acoust. Soc. Am.*, 10 (1939), 270-274.

Wallach, H. 'The role of head movements and vestibular and visual cues in sound localization'. *Journal of Experimental Psychology,* 27 (1940), 339-368.

Wenzel, E. M. 'What perception implies about implementation of interactive virtual acoustic environments'. *Convention paper presented at the 101st AES convention*, Los Angeles (USA), 8-11 Nov 1996.

White, Susan Dorothea. *The Seven Deadly Sins of Modern Times.* Acrylic on wooden table. 1993. Accessed at <http://www.susandwhite.com.au/paintings/sins.html> (on 28 April 2011).

Wiggins, B. 'The generation of panning laws for irregular speaker arrays using heuristic methods' *Convention paper presented at the 31st AES international conference*, London

(UK), 25-27 Jun 2007.

Wigware VST plugins. Accessed at <http://www.brucewiggins.co.uk/?page_id=78> (on 30 Jan 2011).

Wii head tracking in VR Juggler through VRPN. Accessed at <http://wognum.home.xs4all.nl/wii/> (on 14 Jun 2011).

Wiimote Head-Tracking. Accessed at <http://rpavlik.github.com/wiimote-head-tracker-gui/> (on 14 Jun 2011).

Worrall, D. 'Space in sound: sound of space'. *Organised Sound*, 3 / 2 (1998), 93-99.

Zahorik, P. 'Estimating sound source distance with and without vision'. *Optometry and vision science*, 78 (2001), 270-275.

Zirkonium. Accessed at <http://www.zkm.de/zirkonium> (on 19 Jun 2011).

ZLB webpage. Accessed at <http://www.friendlyvirus.org/artists/zlb/@/code/> (on 23 Jun 2011).