3D audio technologies for music production

Author: Gerard Erruz
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

External Advisor: Adan Garriga, UB Advisor: Esther Pascual

Abstract: This report aims to give an overall idea of 3D audio technologies in a physics point of view. Some of the most representative methods are revised and discussed. Ambisonics technique is specially analised due to its powerful and complete approach. In order to "see" for ourselves its strengths and drawbacks we implemented and tested a basic Ambisonics encoding/decoding tool.

I. INTRODUCTION

3D audio technologies include all audio recording/reproduction methods and technologies designed for the recreation of a complete sound field around the listener. As these technologies have been developed, the notion of space has been gradually established as a new musical parameter.

There are some examples of the use of space in musical compositions. Antiphonal music, consisting in two separate choirs singing alternate verses during liturgies was a common practice by Middle Eastern Jews in biblical times and Roman Catholic Church in the fourth century. As described by Schmele [1], classical music composers such as Beethoven, Mahler or Ives had used some space notions in their pieces. These include the use of separate orchestras in the same stage, the distribution of different instruments behind the audience, the evolution of melodic phrases through different instruments and so on.

The development of loudspeaker, capable of emulating any timbre and also of duplicating it using as many loudspeakers as desired, changed the musical paradigm in the early 20th century. The reproduction of the same sound using two or more loudspeakers is the basic condition to create virtual sound sources in the space between them as explained below. Music composers have progressively embraced space as a new musical parameter such as pitch, rhythm or dynamics. Nevertheless, there is still a long way to go before it can be understood not solely as a complementary effect but as an independent and meaningful musical variable.

Section II reviews some of the recently developed methods concerning 3D audio technologies.

A. Binaural

B. Amplitude Panning

C. Ambisonics

There are two basic approaches to three-dimensional audio reproduction. The first one considers the physiological mechanisms that allow us to hear sounds in space (e.g. Binaural) while the second focuses in the synthesis of the sound field around the listener (e.g. VBAP, Wave Field Synthesis).

An Ambisonics implementation using Eurecat's audio system is explained and discussed in Section III.

II. STATE OF THE ART

A. Binaural

As stated by Rayleigh in [2], our brain compares the signals reaching both ears in order to know the direction of arrival of sound. The human head and torso act as obstacles so the intensity difference between ears becomes a good indicator of directionality. This is called Interaural Level Difference (ILD) and is only noticeable for high frequencies (>1.5 kHz) unless the source is located very close to one ear [3]. For frequencies below 700 Hz (when wavelength doubles the distance between ears), the phase difference has the leading role in the localisation of sound sources. This phase difference is called Interaural Time Delay (ITD).

Further research showed how other mechanisms influenced this localisation. The mere use of ILD and ITD cues would lead us to the so-called cones of confusion (Fig. 1). Two sound sources placed in the surface of the cone will have the same cues. The first mechanism to avoid this situation are the small rotations of the head which break the symmetries and make the cone of confusion vanish. The other one involves our own body characteristics, especially those from pinna and ear canal, but also from our nose, head and torso. Interaction of sound with these obstacles acts as a filter that sends localization information to the brain [4].

The basic idea of binaural techniques is to reproduce the same pressure signals recorded in the ear positions directly to the ears of a listener. These sound signals can be measured using a mannequin head with two pressure microphones each placed into one ear canal to record the desired sound field. This mannequin provides the required head-shadow and the outer and inner ear filtering mentioned before.

The most common playback technique for binaural recordings is the headphone reproduction. It delivers each signal directly to the corresponding ear and acts

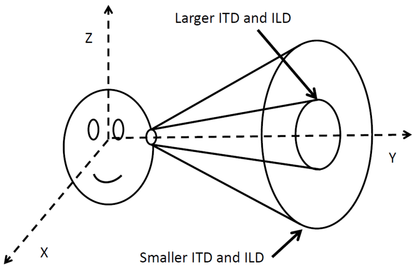

FIG. 1: Schematic illustration of the cone of confusion. Adapted from 3D audio technologies: applications to sound capture, post-production and listener perception by G. Cengarle, 2012, PhD Thesis, Universitat Pompeu Fabra, p. 11. Copyright 2012 by Giulio Cengarle.

as an acoustic barrier between the listener and unwanted external sounds. There are other approaches to binaural playback using loudspeakers and the required cross-talk cancellation techniques (to cancel the signals going to the wrong ear) such as Ambiphonic reproduction [5].

Mathematically, the signal arriving to the ear drums from a sound source is the convolution product between the original signal and the Head Related Transfer Function (HRTF). This function holds the information of the source and ear positions, the shape of the pinna and the ear canals and other contributions due to the torso, nose, etc. (i.e. the personal filtering).

B. Amplitude Panning

When we play the same audio signal using two loudspeakers, our perception places this object in a certain point between them depending on the respective gain values.

The sine law describes this situation in the case where the listener is not following the sound and stays with the head pointing directly forward. Eq. 1 shows this approach in the case where two loudspeakers are involved.

$$\frac{\sin\theta_S}{\sin\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{1}$$

where θ_s is the virtual source panning angle (in the arc between the two loudspeakers), θ_0 is the loudspeakers angular position (supposing a symmetric distribution in respect to the head pointing direction) and g_i are the gain factors.

Sine law is valid only for frequencies below 500 Hz, and also with the assumption that the loudspeakers are in the same vertical plane as the listener. Improvements of sine law are explained by Pulkki in [6].

An extra expression for the gain factors is needed in order to solve Eq. 1. The most common one is the constant loudness condition

$$\sqrt[p]{\sum_{n=1}^{N} g_n^p} = 1, \tag{2}$$

where N is the total number of loudspeakers and p is an adjustment factor for different listening room acoustics. For example, $p = 1$ maintains constant amplitude for the virtual source and fits the anechoic chamber conditions. And $p = 2$ is the constant energy case, which works fine for chambers with some reverberation.

Amplitude Panning can be extended easily to 3D loudspeaker setups, where not all the loudspeakers are in the same plane as the listener. Here, a maximum of three loudspeakers are used to locate the object inside the triangle they delimit. In 1997, Pulkki [7] developed the most employed method, named Vector Based Amplitude Panning (VBAP). In this approach, the sound source position is described by a vector p which is the sum of the three vectors pointing at the loudspeakers in the chosen triangle with magnitudes proportional to their gains (Fig. 2).

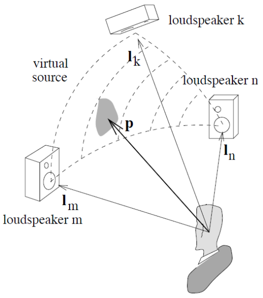

FIG. 2: VBAP virtual sound source location schematic. Loudspeakers in a triangle can change the virtual sound source location by adapting their gains. Adapted from Spatial sound generation and perception by amplitude panning techniques by V. Pulkki, 2001, PhD Thesis, Helsinki University of Technology, p. 17. Copyright 2001 by Helsinki University of Technology.

Mathematically, taking the condition of equidistance between each loudspeaker and the listener, we can define the position of the n_{th} loudspeaker as the unit Cartesian vector $\vec{l_n}$ and the sound source position $\vec{p_n}$

$$\vec{l_n} = \left(\, l_{n1}, l_{n2}, l_{n3} \, \right), \tag{3}$$

$$\vec{p_n} = \left(\, p_n, p_m, p_k \, \right). \tag{4}$$

So we can write $\vec{p}$ as a combination of the $\vec{l_i}$ in the triangle

$$\vec{p_n} = g_n\vec{l_n} + g_m\vec{l_m} + g_k\vec{l_k}, \quad \vec{p_n} = \vec{g}L_{nmk}, \tag{5}$$

from which we can get the gain factors

$$\vec{g} = \vec{p}L^{-1}_{nmk} = \left(\, p_n, p_m, p_k \, \right) \begin{pmatrix} l_{n1} & l_{n2} & l_{n3} \\ l_{m1} & l_{m2} & l_{m3} \\ l_{k1} & l_{k2} & l_{k3} \end{pmatrix}, \qquad (6)$$

It works in the case where L^{-1}_{nmk} does exist. That is, when the vector basis describes a 3D space. We are simply changing bases to represent p in the sources triangle vector basis.

Well need an extra condition for the gain factors so Eq. 2 is taken as in the 2D case.

C. Ambisonics

Ambisonics appeared in the 70s with the work by Michael Gerzon. The basic idea of his proposal was the recreation of the pressure field in the listening point by reproducing the pressure values in a certain sphere surrounding the listener. In order to reproduce all directions perfectly, one would need an infinite number of minute speakers. An alternative realistic approach is introduced by Gerzon: the directional information can be stored as a series of spherical harmonics. If we take a monochromatic sound field and write down the wave equation in spherical coordinates, the pressure at point $\vec{r}$ can be expressed as:

$$p(\vec{r}, \omega) = \sum_{m=0}^{\infty} i^m j_m(kr) \sum_{0 \le |n| \le m}^{N} A^{\sigma}_{mn}(\omega) Y^{\sigma}_{mn}(\theta, \varphi) \qquad (7)$$

where $Y^{\sigma}_{mn}(\theta, \varphi)$ are the spherical harmonics, $j_m(kr)$ are the Bessel functions of the first kind, $A^{\sigma}_{mn}(\omega)$ are the coefficients of the expansion, which describe the spatial properties of the field, and $k = \omega/c$ is the wave number.

Ambisonics is introduced as a complete model, taking into account both the encoding and the decoding processes. Its main contribution is the possibility of coding the recorded information in a certain number of channels, only depending on the order of the expansion needed. So if we truncate the series in order L, the number of Ambisonics channels stored is equal to the number of harmonics, that is, $(L+1)^2$. And the most important aspect: the number of encoded channels and their information are independent from the final loudspeaker layout. In the decoding step, the model takes this information and the layout configuration to calculate the gain of each loudspeaker in order to reproduce the recorded sound field. Notice that the exact reconstruction would need the infinite series solution.

The original article [8] introduces B-format, consisting of the four components of the first order Fourier-Bessel expansion. One can notice that these first order components have the same shape as the directional characteristics of omnidirectional and bidirectional microphones (top four diagrams in Fig. 3).

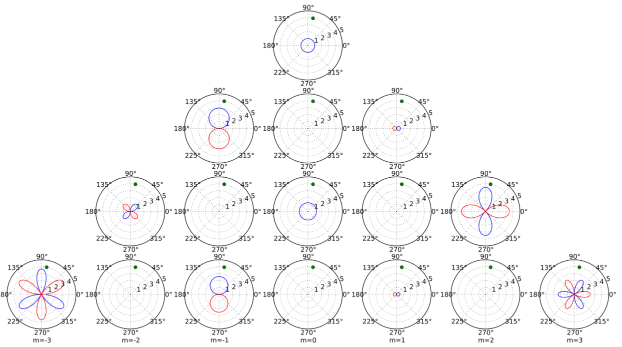

FIG. 3: 3rd order Ambisonics encoding of a punctual sound source (green) located around 85 azimuth. Positive values are in blue, and negative in red. Adapted from Real-Time 3D Audio Spatialization Tools for Interactive Performance by Andrés Pérez-López, 2014, Master Thesis, Universitat Pompeu Fabra, p. 11.

Omnidirectional component W takes the pressure at a certain point (zeroth order) while bidirectional ones (X, Y and Z) store the three orthogonal directions for the pressure gradient or velocity (first order). We can note that the B-format embraces the four physical quantities that define an acoustic field, made explicit in Euler's fluid dynamics equations.

Increasing the order of the Ambisonics expansion provides more information about the sound field near the listening point. Fig. 3 shows each harmonic contribution up to 3rd order for a punctual sound source located around 85 azimuth. Higher orders extend the listening area where the sound field remains fully reconstructed so a better listening experience is achieved.

III. AMBISONICS IMPLEMENTATION

Ambisonics techniques turned to be the more general and complete for the reproduction of any recorded or virtual sound field. In order to immerse oneself in this approach, the author has implemented a Higher-Order Ambisonics encoding/decoding system (to 3^{rd} order).

The implemented process is described next for a first order expansion:

A. Encoding

If we take a sound source S from the direction $\vec{u}$, the information is encoded in $\vec{A} = \left(\, W, X, Y, Z \, \right)$ as:

$$\begin{aligned} W &= S \\ X &= \vec{u} \cdot \vec{x}\, S = \cos\theta \cos\delta\, S \\ Y &= \vec{u} \cdot \vec{y}\, S = \sin\theta \cos\delta\, S \\ Z &= \vec{u} \cdot \vec{z}\, S = \sin\delta\, S \end{aligned} \qquad (8)$$

where $\vec{u}$ is described by the angles (θ, δ) in spherical coordinates. The spherical harmonic expressions used in this report are those described by Daniel [9]. For following discussion we introduce

$$\vec{B_u} = \left(1, \vec{u} \cdot \vec{x} , \vec{u} \cdot \vec{y}, \vec{u} \cdot \vec{z} \right), \tag{9}$$

as the characteristic vector for position $\vec{u}$.

B. Decoding

The decoding process assigns a linear combination of the encoded information channels to each loudspeaker, taking an encoded signal S from direction $\vec{u}(\theta, \delta)$ and N loudspeakers placed in directions $\vec{u_i}(\theta_i, \delta_i)$ equidistant to the centre of the listening area. So, ideally,

$$\vec{L} = D \cdot \vec{A} \quad \vec{L} = \left(L_1, L_2, ..., L_N \right), \tag{10}$$

where L_i is each loudspeaker signal and D is the decoding matrix.

Just as in the sound source encoding step, here each loudspeaker position has an associated characteristic vector. That is, spherical harmonics are used to assign a certain weight to each loudspeaker playback contribution. With this we are returning to the cartesian coordinates space.

$$\vec{B}_{u_i} = \left(1, \vec{u_i} \cdot \vec{x_i} , \vec{u_i} \cdot \vec{y_i}, \vec{u_i} \cdot \vec{z_i} \right), \tag{11}$$

The so called *reencoding matrix*, due to the recurrent spherical harmonic weighting, is then defined:

$$B_{\{u_i\}} = \left(\vec{B}_{u_1}, \vec{B}_{u_2}, ..., \vec{B}_{u_N} \right), \tag{12}$$

If we want the final sound field (right) to match the original encoded one (left), we find the expression

$$\vec{B_u} \cdot S = B_{\{u_i\}} \cdot \vec{L} \tag{13}$$

Defining gain vector $\vec{G} = \left(G_1, G_2, ..., G_N \right)$, which stores each loudspeaker final gain,

$$\vec{B_u} \cdot S = B_{\{u_i\}} \cdot \vec{G} \cdot S \tag{14}$$

Now we see that the decoding matrix D is actually the inverse form of the *reencoding matrix* $B_{\{u_i\}}$, for stability issues, the pseudoinverse form is normally used, so

$$\vec{G} = B^{pinv}_{\{u_i\}} \cdot \vec{B_u} = B^t_{\{u_i\}} \left(B_{\{u_i\}} B^t_{\{u_i\}} \right)^{-1} \cdot \vec{B_u} \tag{15}$$

In our implementation, we supposed a regular loudspeaker layout, which ensures that $\left(B_{\{u_i\}} B^t_{\{u_i\}} \right) = N \cdot I$ (where I is the identity matrix), so

$$\vec{D}_{pinv} = \frac{1}{N} B^t_{\{u_i\}} \tag{16}$$

Finally,

$$\vec{G} = \vec{D}_{pinv} \cdot \vec{B_u} \tag{17}$$

C. Decoding criteria

Gerzon defines two quantities related with the gain components in order to establish a criterion on the decoding matrix: velocity vector and energy vector.

$$\vec{V} = \frac{\sum_{i=1}^{N} G_i \hat{u}_i}{\sum_{i=1}^{N} G_i} = r_V \hat{u}_V \tag{18}$$

$$\vec{E} = \frac{\sum_{i=1}^{N} G_i^2 \hat{u}_i}{\sum_{i=1}^{N} G_i^2} = r_E \hat{u}_E, \tag{19}$$

where $r_{V,E}$ and $\hat{u}_{V,E}$ are the modulus and the direction of each quantity respectively.

There are two main approaches to Ambisonics decoding: the physical approach aims to recreate the exact recorded sound field in technical terms. So the criteria will be to maximize r_V in order to preserve the original velocity vector. The psychoacoustic approach is based in Rayleighs theories so $\vec{V}$ will be important at frequencies below 700 Hz and $\vec{E}$ over 700 Hz (maximizing r_V and r_E respectively). The two criteria show linear solutions in regular loudspeaker arrays.

D. Implemented code outline

The main goal of the code is to take a mono audio file and generate an Ambisonics weighted copy of it for each loudspeaker in the layout. As initial variables we have loudspeakers angular positions and the desired angular location of the virtual sound source.

The chosen programming language is Python (v.2.7). Specific functions from *SciPy* and *PyLab* open-source libraries are required, specially for the signal processing steps.

A basic information flow of the implemented code is introduced.

Encoding

1. Reads mono *WAV* file

2. User defines desired azimuth and elevation angles
3. Fills encoding vector with spherical harmonics expressions for the introduced angles

Decoding

4. Fills the pseudoinverse D matrix using the angular position of the loudspeakers and spherical harmonics expressions
5. Computes the gain vector from the encoding vector and D
6. Multiplies the gain vector and the initial audio file to get each loudspeaker signal
7. Writes as many *WAV* files as loudspeakers in the final layout

IV. CONCLUSIONS

The implemented code was tested in Eurecat's 3D audio system described in the Appendix. The generated audio files were played simultaneously in the corresponding loudspeaker position in order to recreate the global sound field in the central point of the studio.

Listening test suggested a good directionality performance. Some amplitude changes were noticed for different virtual source positions. In Eq.16 we supposed a regular layout situation. The loudspeaker distribution in the studio follows some simetries but it's not a regular layout. The later explains why different directions perform different playback amplitudes.

It's also important taking in account the chosen decoding criterion. The reconstruction of the final sound field by reproducing the exact original field (Eq.14) means that the velocity vector $\vec{V}$ is recreated. With this, our implementation will work better for low frequency signals.

In order to evaluate how directionality changes by adding successive orders we performed listening tests for a certain virtual source position at different series truncation. The results of this experience show an improvement in directionality and presence. That is, the same sound is better defined in space, showing less spreading effects.

Ambisonics is a very interesting approach, specially by experiencing it in the studio. Differently from other techniques such as Amplitude Panning, in an Ambisonics playback all loudspeakers (or nearly all of them) are contributing to the global field. If the listener moves closer to a certain loudspeaker, its solo contribution can be heard and the original sound directionality vanishes.

That explains why combined techniques are normally used in professional 3D audio productions. Amplitude Panning perform very good directionality while Ambisonics guarantees a complete soundfield recreation and is used in ambient effects.

Acknowledgments

I want to thank all the people that helped me in this journey. My two official advisors, Adan and Esther, for their unconditional help. I'm specially gratefull with my Eurecat team mates who embraced me in their projects: Adan, Umut, Julien, people from the image team. Special mentions to Kyran and Davide. Finally, I also want to thank my family and friends for supporting me and listening to my explanations on 3D audio during this semester.

[1] T. Schmele *"Exploring 3D Audio as a New Musical Language"*, *Master Thesis, Universitat Pompeu Fabra*, 2011.

[2] John W. Strutt (Lord Rayleigh), *"The theory of sound, volume II"*, *Macmillan and Co., pp. 440-443*, 1896.

[3] John W. Strutt (Lord Rayleigh), *"On our perception of sound direction"*, *Philosophical Magazine, vol 13, pp. 214-232*, 1907.

[4] Michael A. Gerzon, *"Surround-sound psychoacoustics"*, *Wireless World, 80(1468), pp. 483-486*, 1974.

[5] A. Farina, R. Glasgal, E. Armelloni, A. Torger, *"Ambiphonic principles for the recording and reproduction of surround sound for music"*, *Philosophical Magazine, vol 13, pp. 214-232*, 1907.

[6] V. Pulkki, *"Spatial sound generation and perception by amplitude panning techniques"*, *PhD Thesis, Helsinki University of Technology*, 2001.

[7] V. Pulkki, *"Journal of the Audio Engineering Society, 45(6):456466"*, *Philosophical Magazine, vol 13, pp. 214-232*, 1997.

[8] Michael A. Gerzon, *"Periphony: With-height sound reproduction"*, *Journal of the Audio Engineering Society, 21:2-10*, 1973.

[9] J. Daniel, *"Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia"*, *Thèse de doctorat, Université Paris 6*, 2001.

V. APPENDIX

Ambisonics implementation described in section III was tested in Eurecat's 3D audio system in Barcelona. This studio has a 25.3 loudspeaker distribution in a three-quarters-of-a-sphere layout. The information flow is shown in Fig. 4.

A. Software processing

The implementation code takes a *WAV* mono audio file and creates 25 output *WAV* files. Each file contains the original audio information weighted depending on the requested angular position.

A *Max* specific patch is used to distribute these files to their respective loudspeaker channel. *Max* is a programing language and a real-time processing audio software with a modular visual interface. It is able to handle multi-channel outputs.

The information flow is tailored by *JACK Audio Connection Kit* and a proper connection distribution. Before leaving the computer, signals are treated by a *jconvolver* set configuration. This configuration applies certain gain and delay values for each channel in order to virtually set an equidistant loudspeaker layout.

B. Hardware processing

A *RME MADIface* sound card collect computer's output and sends it via optical cable to two 25-channel *Ferrofish AD/DA* converters (working as Digital-Analog converters in this case). Finally, analog signals are sent to each loudspeaker.

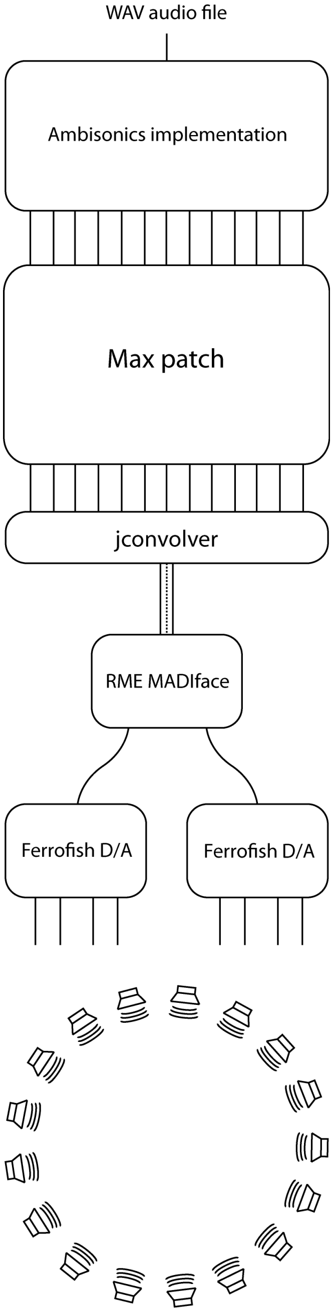

FIG. 4: Eurecat's 3D audio system schematic.