AUDIENCE – Audio Immersion Experiences in the CAVERNA Digital

Regis Rossi A. Faria[1], Leandro F. Thomaz[1], Luciano Soares[1], Breno T. Santos[1], Marcelo K. Zuffo[1], João Antônio Zuffo[1]

[1]LSI – Laboratório de Sistemas Integráveis – Universidade de São Paulo
Av. Luciano Gualberto, 158 tv.3 – 05508-900 – São Paulo – SP – Brasil

{regis,lfthomaz,lsoares,brsantos,mkzuffo,jazuffo}@lsi.usp.br

Abstract. *In this paper we introduce the AUDIENCE project undergoing in the CAVERNA Digital of the University of São Paulo, whose main purpose is to implement flexible and scalable multichannel spatial audio solutions for this CAVE environment, to permit navigation through a 2D/3D audiovisual scene with both visual and auditory immersion. An architecture for spatial audio production has been proposed to build auralizators, and a whole infrastructure has been designed and installed in the CAVE, so to support several speaker array setups. We present our activities towards the construction of an Ambisonics auralizator, outline some details and challenges of the implementation. We also cover recent achievements of the project and future directions of investigations.*

1. Introduction

Most previous audio systems in immersive virtual reality environments (such as CAVE systems) were more concerned about having some sonification or accompanying sound than with realistic and accurate spatial sound production and auralization. Very few works have addressed spatial audio for CAVE's [Ogi, 2003], [Eckel, 1998]. Frequently these relied upon amplitude panning techniques, and did not go for real sound field rendering, such as with Ambisonics [Gerzon, 1973] or Wave Field Synthesis (WFS) techniques [Berkhout, 1993]. Very often sound was approached as a secondary or complimentary task, addressing stereo audio support only.

As of a consequence of the popularization of the multichannel systems due to 5.1 standard issued by ITU-T, there was a reborn of interest in sound field rendering in the last decade, and many groups are working on multichannel approaches addressing large and stable sweet spots for larger audiences and environments, such as cinemas theaters and auditoriums. Take for recent examples the IOSONO [IOSONO, 2005] and CARROUSO [Carrouso, 2005] project approaches, both relying on WFS techniques.

In this paper we introduce the project AUDIENCE – Audio Immersion Experience by Computer Emulation [Faria, 2004], which is undergoing in the CAVERNA Digital of the University of São Paulo, Brazil [Zuffo, 2001]. The project aims the implementation of a flexible and scalable system for 2D/3D audio production and displaying through multichannel techniques.

Objectives include since the implementation of decoding systems for the traditional formats and surround configurations (such as 5.1, 7.1, DTS® and Dolby®) up to the development and deployment of more sophisticated formats for 2D/3D audio generation and reproduction, such as Ambisonics and WFS. In the AUDIENCE project we are interested in giving, for more than one user inside the CAVE, the auditory immersion experience similar to the one achieved in the visual domain with stereoscopy techniques. The possible universe of applications is infinite, but we can stress some interesting examples. Imagine for instance the experience of navigating through a symphony orchestra playing on one of your favorite theaters. Or a more modern approach to enjoy alternative bands, primarily setting up a stage, focusing timbres and gestures and locating them in space (associating them to certain positions) and assigning different acoustics properties to instruments or regions in space.

The idea surpasses the situation where users can set up a desired stage and a position in the audience, but goes much further, making it possible to be within the orchestra, e.g. closer to instruments, to experience the conductor position, to create acoustic dimensions and locate timbres in space. One ultimate degree of freedom in this way is to make available such an edition power for the multichannel sound track engineer and also for the final consumer. To investigate such advanced scenarios we are addressing mainly sound field auralization techniques in immersive audiovisual environments.

The following sections of this paper will address activities of the 1st and 2nd phases of the AUDIENCE project, mainly about the sonorization infrastructure and software-based auralizator proposals for sound field generation in CAVE's. These reproduce novel results in the area, since construction of audio infrastructures and multichannel auralization in CAVE's are not well covered in literature. Our approach in audio infrastructure relies upon high-quality commodity equipment found in the marketplace. For the auralization engine we bet on software builds, which are remarkably more flexible and do not show computational disadvantages when cluster parallel computers are available to host processes, as it is the case at the CAVERNA Digital.

2. Sonorization infrastructure for AUDIENCE project

Sound field projection techniques are more difficult to set up in auditory restrictive environments such as CAVE's, and this is one of the main reasons this has not been addressed before in immersive cubes. However, there are several ways to systematic attack some traditional impeachments which render almost impracticable stable sound field projection in these environments. One important thing is to have a proper audio hardware and flexible patch bays.

In the AUDIENCE project we have designed and constructed a complete audio setup, from the soundcard selection up to the speaker array mounting. There was no previous system in marketplace adequate for the purposes of the AUDIENCE project, although Lake Huron (www.lake.com.au) and AuSIM (www.ausim3d.com) have some sonorization systems for CAVE's.

The patch hardware consists of 3 levels: (1) the *rack bay*, where digital audio interface and amplifiers are mounted, (2) the *multicable distribution* section, where balanced electric routes are sent from rack to the CAVE backstage, and (3) the *terminal panels*, from where terminal cables are connected via TRS plugs to the speakers.

We are using in the 1st phase of the project 8 to 16 analog hi-fi LANDO speakers (www.lando.com.br), high quality and low noise Sankya multichannel amplifiers (www.sankya.com.br), and “Cabos Golden” cables and distribution panels (www.cabosgolden.com.br).

Sonorization of CAVE’s are somehow complicated due to extensive forbidden areas where loudspeakers cannot be positioned due to back visual projection. Speakers should preferably be invisible to users, lay behind screens, and irradiate properly towards the centre of the CAVE but covering a large listening volume inside, around the center.

3. Auralization architecture

Spatial audio in interactive electronic media have roughly three mainstreams, one addressing the game industry, one addressing the multichannel surround market, and finally one addressing high precision/realistic rendering of acoustical phenomena. A lack of a reference architecture linking these segments is an important cause of several individual proposals in the field, without cross-references.

Faria (2005) proposed an architecture for referencing spatial audio production, based on four functional layers, from scene composition up to rendering and sonorization processes. Figure 1 shows the layers reference model of the architecture.

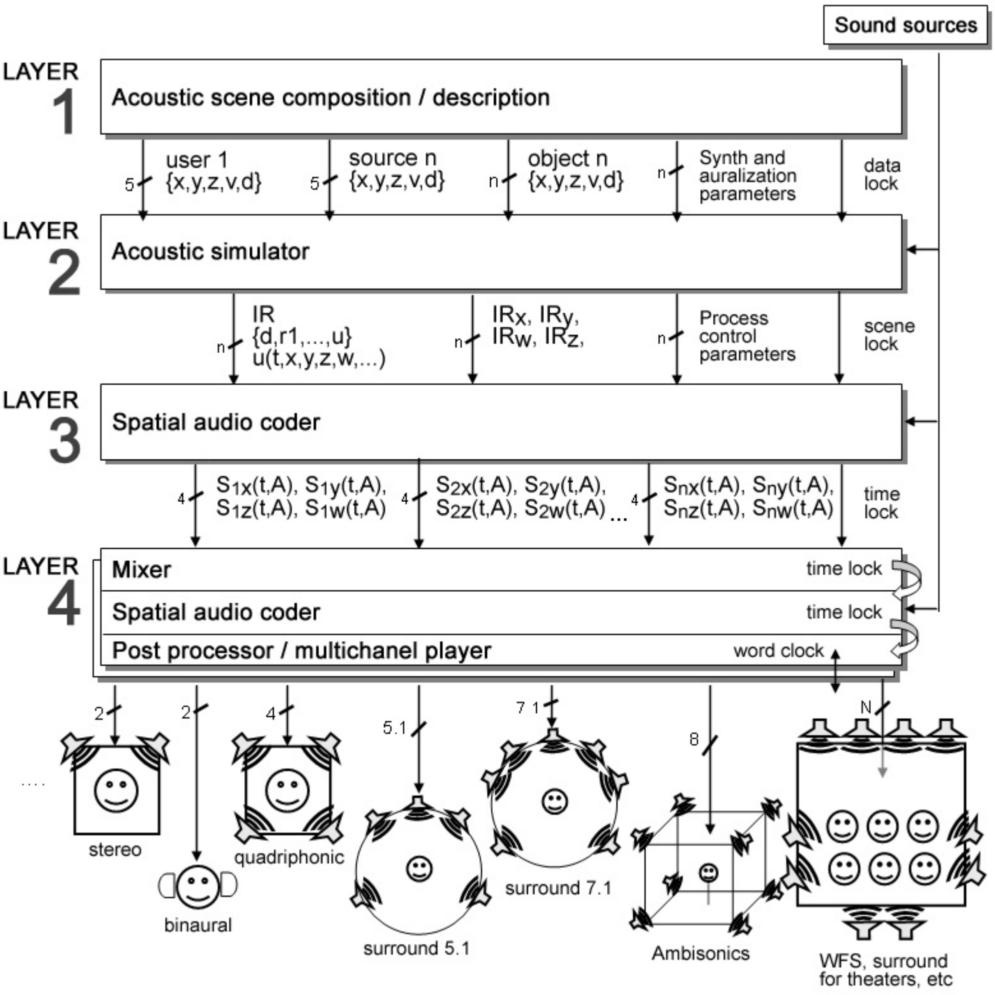
Sound sources
LAYER 1
Acoustic scene composition / description
5 user 1 {x,y,z,v,d}
5 source n {x,y,z,v,d}
n object n {x,y,z,v,d}
n Synth and auralization parameters
data lock
LAYER 2
Acoustic simulator
n IR {d,r1,...,u} u(t,x,y,z,w,...)
n IRx, IRy, IRw, IRz,
n Process control parameters
scene lock
LAYER 3
Spatial audio coder
4 S1x(t,A), S1y(t,A), S1z(t,A), S1w(t,A)
4 S2x(t,A), S2y(t,A), S2z(t,A), S2w(t,A) ...
4 Snx(t,A), Sny(t,A), Snz(t,A), Snw(t,A)
time lock
LAYER 4
Mixer
time lock
Spatial audio coder
time lock
Post processor / multichanel player
word clock
2
2
4
5.1
7.1
8
N
....
stereo
binaural
quadriphonic
surround 5.1
surround 7.1
Ambisonics
WFS, surround for theaters, etc

Figure 1 – Layers reference model

Usually tools for programming spatial audio in electronic platforms rely on proprietary or non-complete API's which do not offer flexibility for developers to adopt one or another component from different vendors. For example, they offer tools for selecting positions and attributes of sources, environment and receptors, but not tools for selecting acoustic propagation/simulation techniques and spatial audio coding formats.

This architecture aims a clear identification of signals in between functional layers, so that tools from different developers can still interoperate. As we pursue the highest possible interconnection and interchange of tools avoiding the need for a whole system re-write due to a change in one function, we are building auralization engines following this strategy. Several possible output configurations and speakers arrays are possible for final sonorization, as seen in the figure 1.

4. Building an auralizator

Based on the architecture above, in the first phase of the project we are building an auralizator using image-source techniques for modeling acoustic simulation and the Ambisonics technique for spatial coding the sounds.

Figure 3 shows a functional block diagram of the auralization engine. Three major blocks are detached: (1) the (audiovisual) VR (virtual reality) application, (2) the auralizator (which comprises the acoustic simulator and spatial format coder) and (3) the sonorizator node, where decoding, post-processing and multichannel reproduction take place.

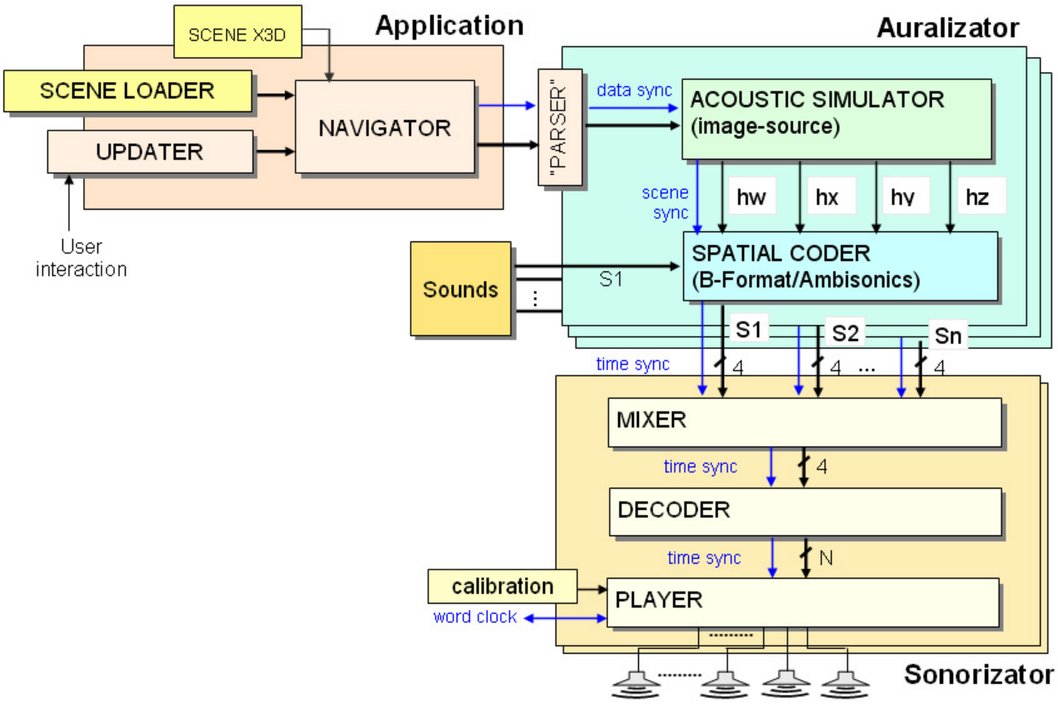

Figure 3 – Auralization engine block diagram

We are using the Pure Data (PD) software as framework to build auralizator blocks and make the necessary interconnections [Puckette, 2005]. PD is a real time graphical programming environment for audio and music applications, widely used by related communities [Noisternig, 2003], [Fraunberger, 2003].

Having a flexible and real-time capable tool for the audio subsystem and having proper software components for the *glue-logic* with visual and CAVE management subsystems are two major issues. These are being addressed as two concurrent task forces. A remarkable tool contributing to this interconnection is a shared format for audiovisual scene description, and a synchronization hierarchy similar to that used in the visual domain.

Blocks that perform specific tasks are connected to implement a function: patches are created producing chains of several processing algorithms through a chain of connections. When PD is told to compute audio, it starts to pass audio signal chunks through the blocks.

PD is used in the AUDIENCE project as the tool that binds together all the different modules and renders the audio. It also permits these modules to communicate with the VR navigator application and the operating system by built-in network blocks: *netsend* and *netreceive*. These blocks permit also exchange audio information between the navigator and the PD path through TCP sockets.

4.1. An auralizator patch in PD

Basically, four processing layers are shown in the patch presented in Figure 4 below.

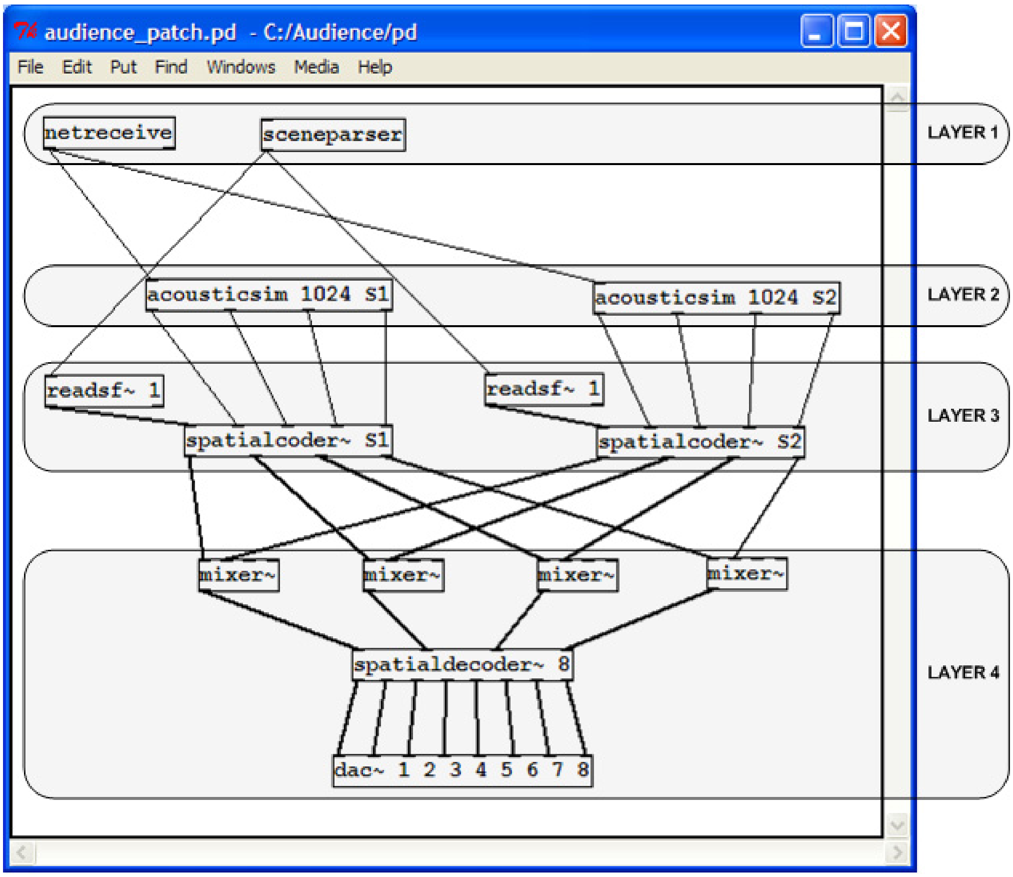

Figure 4 – An example of auralization patch for Ambisonics

The first one, *sceneparser*, receives and parses acoustic attributes from audio nodes within the scene graph, generated by the 3D VR navigator application. Next section will introduce this application, and cover the scene description layer with more detail.

In a second layer, *acousticsim* receives the acoustical attributes, environment dimensions, and listener and source's positions, and render the acoustic simulation using an image-source algorithm modified from Allen's (1979), producing as outputs multidimensional impulse responses (IR_W, IR_X, IR_Y and IR_Z) actually coded in B-format. In other words, we have developed an image-source to B-format acoustic renderer.

Next layer is formed by the *spatialcoder~* block, which basically convolves the sound sources (S1 and S2 in Figure 4) with a B-format set of impulse responses, producing an Ambisonic-coded sound source.

If more than one sound source is being rendered at a time, their outputs in B-format may be combined to generate a single 4-channel set. This is done in a mixing layer. Final B-format signals are then routed to the *spatialdecoder~* block, which then produces the speakers' outputs. Item 4.5 ahead describes this layer with more details.

4.2. Acoustic scene parsing

In the CAVERNA Digital we have developed a 3D virtual reality browser called Jinx, which permits to navigate in a virtual scene described in the X3D format [X3D, 2005], using commodity clusters [Soares, 2004]. For the acoustic scene, the Jinx parses acoustic data from the scene graph, and sends it to the PD auralization patch.

Jinx runs in each node of the cluster. Some nodes can take care of graphics, I/O devices etc. At least one node must take care of sound, and host the auralization tree. The digital sound interface must also be installed in this server. The node running the sound server can share resources with other processes (like video) or it can run in a dedicate fashion. Auralization processes share data with visual scene graph, and may also actually run distributed in more than one node. Figure 5 shows how processes communicate inside Jinx, including sound. The current AUDIENCE auralization patch is designed to be called by Jinx.

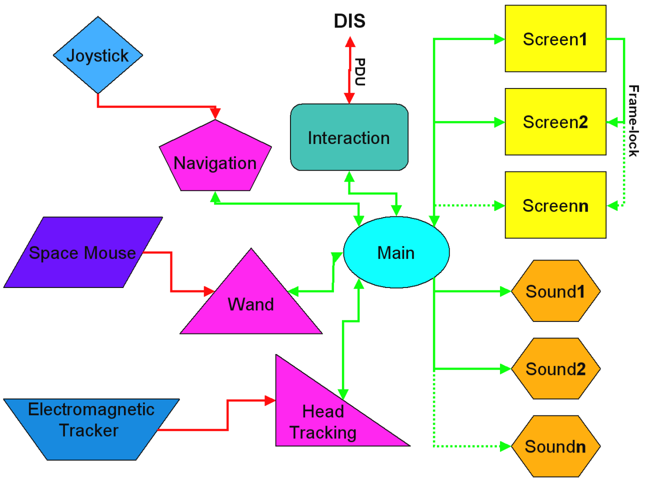

Figure 5 – Jinx Hierarchy

If Jinx is not able to find AUDIENCE auralizator, it is going to use Fmod, an open source sound library [Fmod, 2005]. Fmod supports some features for 3D sound, but with hidden implementations for layer 2 and 3. It may also be used as multichannel player in layer 4. Jinx loads Fmod with the sound files (like wave) with objects' position and orientation.

If AUDIENCE auralizator is available, Jinx parses the file and sends the sound configuration to the *sceneparser* block in PD thought network. *Sceneparser* is an external dynamic library block built in PD, performing layer-1 functions. During normal operation Jinx updates the user position and orientation for each frame in the *sceneparser*. This is locked with graphical output by means of data-lock/scene-lock sync signals, using *netreceive* block (see Figure 4). At the end of navigation, Jinx sends and end tag to finish sound processes.

For the acoustic scene we found the current version of X3D standard not capable of conveying all the information we needed to transmit to subjacent levels of processing, such as material acoustic attributes. To hold sound information for the environment it was then proposed an extension to X3D scene graph. Figure 6 presents an idea of a scene graph tree with a special sound node called `AcousticScene`, with has information as environment dimensions, objects' acoustic attributes, etc.

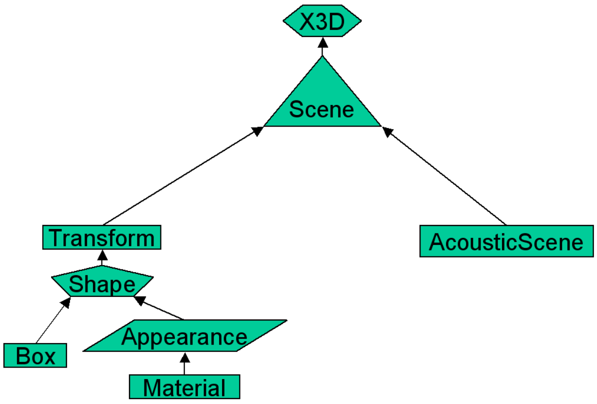

Figure 6 – Scene Graph with Acoustic Information

We also found necessary to add additional children nodes, e.g. the `AcousticMaterial` node. For example take the following code passage:

```
<Transform DEF="floor" center="7.5 0.05 10">
  <Shape>
    <Box size="15 0.1 20"/>
    <Appearance>
      <Material diffuseColor="1 1 1"/>
    </Appearance>
  </Shape>
  <Sound>
    <AcousticMaterial coefref=".8" freqref="1000"/>
  </Sound>
</Transform>
```

In the above passage we define for the floor not only its dimensions but also its acoustic reflection coefficient, which are both needed in the next processing layer. The `AcousticMaterial` node was added to permit our parser block to extract these parameters for the acoustic simulator (the *acousticsim* block).

The MPEG-4 Advanced AudioBIFS (Binary formats for Scenes) is a more comprehensive tool for describing acoustic and sound relevant parameters [Väänänen, 2002], and this is one focus for future works. Spatial coding for several applications do require metadata to be transmitted aside the media itself, which can describe scene, set up decoding and mastering/editing parameters for final channel production in the terminal gear. This is, for example, being explored in the DAB (Digital Audio Broadcast) and SAC (Spatial Audio Coding) initiatives for standardizing multichannel and spatial audio coding/transmission schemes.

4.3. Acoustic simulator

The acoustical simulation is maybe the most important task in the auralization process. Spatial perception and quality are directly associated with this. There are several methodologies for modeling acoustical propagation, often relying on two different approaches for modeling sound: ray-based or wave-based.

For validating the strategy for building and integrating acoustic simulators within our architecture, our first approach was to consider a simple geometry and a precise technique to calculate reflections and obtain artificial impulse responses, such as the image-source method, which is a ray-based technique.

We have developed an acoustic simulator based on Allen's image-source technique for rectangular spaces [Allen, 1979]. A reflection in this technique is supposed to come from a virtual image source, which is located behind the reflective wall, tracked as in optical geometry laws. Figure 7 shows this concept.

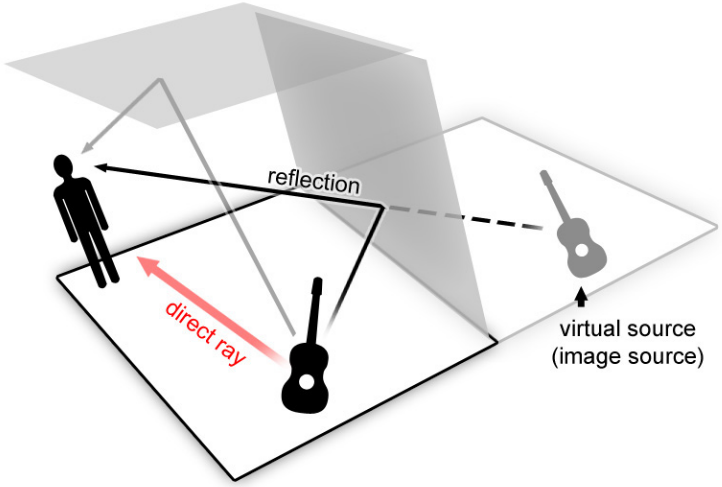

Figure 7 – Image Source ray tracing concept

4.4. Spatial audio format coder

Several formats can be exploited to produce final audio channels comprising both temporal and spatial sound attributes. One of the most elegant, however, is the Ambisonics B-Format. Originally developed in the 1970's by Gerzon and others, its functional mechanism to register a 2D/3D sound field may be explained either via mathematical approach or via an extension of Blumlein's stereo techniques for a 2D and 3D microphone setup. Initially aimed at the record industry, Ambisonics emerged in a time when quadraphonic systems were in decline. This fact, and also a lack of a proper multichannel media to record and transmit all necessary channels at that time, contributed to its low popularity.

We have designed an Ambisonics 1st order coder, which takes the outputs of the acoustic simulator coded in B-Format (4 channels) and convolve them with the anechoic sound source, so to produce a final B-format spatial encoded audio signal. A B-Format set of signals for 3D audio coding is comprised of four channels, W, X, Y and Z, which register the temporal and spatial properties of sound.

4.5. Spatial audio decoder and multichannel player

The spatial audio decoder developed in the current auralization engine is a basic 1st order Ambisonics decoder. Figure 8 shows its block diagram. In the first stage, four shelf filters, one for each B-format channel (W, X, Y, Z) are applied, with the purpose of adjusting the psychoacoustic quality of the sound for enhancing auditory localization cues [Gerzon, 1974]. Next, there is an amplitude matrix stage. Depending on the number of speakers and their positions, it applies different gains to each one of the channels and produces as output a number of channels corresponding to the number of speakers.

A very important component of the decoder is an inverse filter to minimize the effects of the CAVE's projection screens. They are reflective, and have a diffuser effect on transmitting sound from the speakers, located behind them, to the center of the CAVE. This filter is usually designed using the impulse response of the speaker-screen system, measured with a flat microphone inside the CAVE. This convolution is also made within the implemented decoder block.

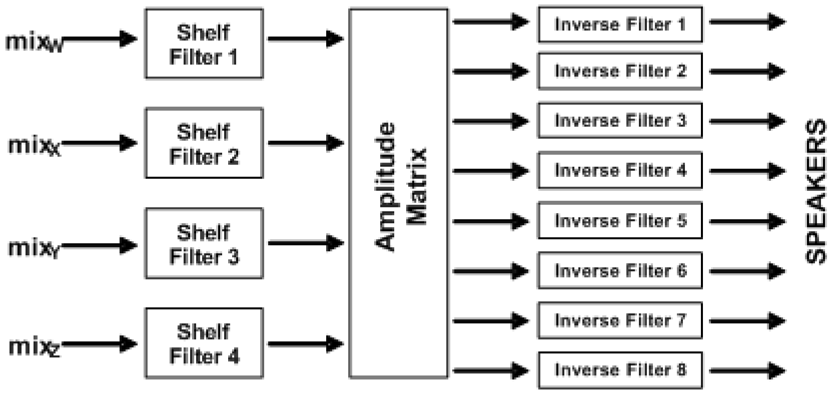

Figure 8 – Block diagram of *spatialdecoder~*

The Ambisonics decoder was implemented as a PD block, named *spatialdecoder~*. The gains were obtained from previous calculations made by Richard Furse [Furse, 2005]. It takes as input a 4-channel mixdown (from a previous mixing stage, where all B-format encoded sound sources available were mixed). It also receives parameters, as the intended type of speaker configuration (cube, octagon etc).

5. Conclusions

The AUDIENCE project 1st phase objectives were achieved with a very flexible and high-quality audio infrastructure setup. In the 2nd phase, very good initial results in perceived spatial quality showed that our current approach to build auralization engines shall lead to progressive refinements. The open/layered architecture permits us to invest independently in any phase of the spatial sound production.

In the scene layer, for instance, the standardization of more powerful nodes for X3D seems a prospective and important task, for a better sound description in virtual environments. The integration of multiple acoustic simulation methods, such as geometrical-based and wave-based techniques, is expected to resolve acoustical phenomena in a large bandwidth and in a complementary way: one technique covering the weakness of the other.

Several spatial audio formats may also be addressed, such as WFS/MPEG-4 AAC (Advanced Audio Coding) and commercial formats, as 5.1/6.1/7.1/10.2 etc. Finally, many decoders, including commercial ones, may be employed in different situations, depending on operating system, application requirements, etc.

We have been testing two speaker configurations: a horizontal (2D) octagonal array, and a cubic (3D) array, both with eight speakers surrounding the CAVE. For the octagonal rig, we have tested it outside the CAVE and surrounding the CAVE, in an experiment that showed the challenges in treating the acoustic paths and interferences due to screen and backstage superimposed acoustics.

A simple experiment was performed where a single source and the listener were positioned in a virtual room with 20x15x8m dimensions. Two timbers were tested, a real flute and a synthetic battery set, both playing a small passage with several important musical gestures for spatial quality evaluation. Figure 9 shows a picture of a testing octagon rig mounted outside the CAVERNA and later mounted surrounding it.

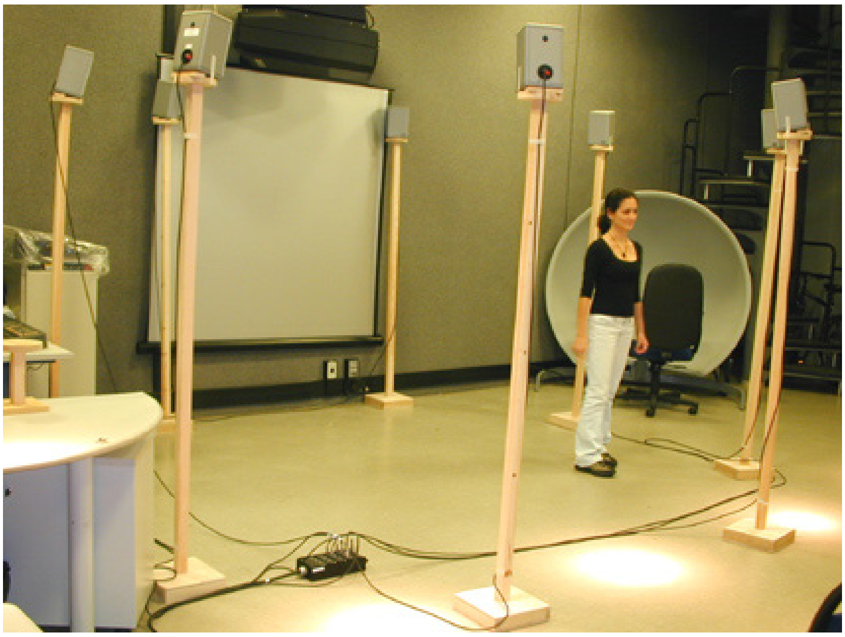

Figure 9 – Octagonal (2D) horizontal speaker array for Ambisonics auralization

Excellent localization was achieved with the auralizator rig outside the CAVE. Inside the CAVE, without any special active or passive treatment, the perception of position was less precise, although the direction cue was always preserved. The cubic array was only tested in the CAVE. Direction was also preserved in this rig, although the elevation difference between source and listener was not easy to perceive. Better speaker positioning however will be soon possible, due to a new elevated support to hold speakers in the upper part of the cube.

Faria discusses some results of these experiments in [Faria, 2005] and also lists future directions and works. Several other experiments are being prepared, for example an audiovisual scene with 4 instruments playing on stage (a flute, a violin, a trumpet and a cello) where we intend to explore complete real-time audiovisual virtual navigation in the environment.

A future phase of the PD Ambisonics decoder is to implement an auto-configuring decoding matrix in a way that, given an arbitrary array of speakers and their positions, the software configures automatically the gains. To accomplish this feature, a numerical solution of the cylindrical Bessel wave equations/functions will be required, which is the basic mathematical concept behind the Ambisonics technique in reconstructing plane waves.

For the WFS phase of the project, the higher number of speakers and other requirements of the technology pose a lot of additional challenges beyond the Ambisonics system. Different speaker arrays and audio distribution will be proposed for this phase, to be addressed in future papers.

References

Ogi, T. et al. “Immersive sound field simulation in multi-screen projection displays”. In: International Immersive Projection Technologies Workshop, 7. Eurographics Workshop On Virtual Environments, 9. Zurich, 2003. Proceedings. Zurich: Eurographics, 2003. p.135-142.

Eckel, G. A spatial auditory display for the CyberStage. In: ICAD’98, 5th International Conference on Auditory Display, Glasgow, 1998. Proceedings. Glasgow: ICAD, 1998.

Gerzon, M. “Periphony: With-height sound reproduction”, JAES Jan./Feb. 1973, Vol.21, No.1

Berkhout, A. J. et al. A wave field extrapolation approach to acoustical modeling in enclosed spaces. Journal of the Acoustical Society of America, v.93, n.5, p.2764-2778, May 1993.

IOSONO Wave Field Synthesis System. In: http://www.iosono-sound.com/ Access in: 20 May 2005.

CARROUSO Project: Creating, assessing and rendering in real time of high quality audio-visual environments in MPEG-4 context: system specification and functional architecture. In: http://www.idmt.de/projects/carrouso/index.html. Access in: 20 May 2005.

Faria, R. R. A. “AUDIENCE – *Audio Immersion Experience by Computer Emulation*”. In: http://www.lsi.usp.br/interativos/nem/audience/. 2004. Accessed in: 20 May 2005.

Zuffo, J. A et al. “CAVERNA Digital – Sistema de Multiprojeção Estereoscópico Baseado em Aglomerados de PCs para Aplicações Imersivas em Realidade Virtual. In: 4th Symposium of Virtual Reality, Florianópolis, 2001. Proceedings.

Faria, R. R. A. “Auralização em ambientes audiovisuais imersivos”. Thesis (Ph.D.). Electronic Engineering. Polytechnic School, University of Sao Paulo. 2005.

Puckette, M. “Pd Documentation”. In: http://crca.ucsd.edu/~msp/Pd_documentation/. Accessed in: 20 May 2005.

Noisternig, M., Sontacchi, A., Musil, T. and Höldrich, R. “A 3D Ambisonic Based Binaural Sound Reproduction System”. In: Audio AES 24th Conference, Banff, 2003.

Frauenberger C., Ritsch, W., Höldrich, R., “Internet Archive For Electronic Music IAEM-IARS (Internet Audio Rendering System)” In: Audio AES 24th Conference, Banff, 2003.

Allen, J. B.; Berkley, D. A. “Image method for efficiently simulating small-room acoustics”. Journal of the Acoustical Society of America, v.65, n.4, Apr. 1979, p.943-950.

“X3D”. In: http://www.web3d.org/. Accessed in: 20 May 2005.

Soares, L. P. and Zuffo, M. K. “JINX: an X3D browser for VR immersive simulation based on clusters of commodity computers’. In Proceedings of the ninth international conference on 3D Web technology, Monterey, California, USA, p.79-86. 2004.

“Fmod music & sound effects system”. In: http://www.fmod.org. Accessed in: 28 Aug. 2005.

Väänänen, R., Huopaniemi, J. “SNHC audio and audio composition”. In: Pereira, F.C.N, Ebrahimi, T. “The MPEG-4 Book”. New Jersey, IMSC Press/Prentice Hall, 2002.

Gerzon, M. “Surround-sound psychoacoustics”. In: Wireless World, Dec. 1974.

Furse, R. “First and Second Order Ambisonic Decoding Equations”. In: http://www.muse.demon.co.uk/ref/speakers.html. Accessed in: 20 May 2005.