A comparison of different surround sound recording and reproduction techniques based on the use of a 32 capsules microphone array, including the influence of panoramic video

FABIO MANOLA, ANDREA GENOVESE, ADRIANO FARINA

Department of Electronics, University of York, Heslington, York, UK

This paper provides a comparison between the operational results obtained reproducing a three-dimensional sound field by means of traditional 1st order Ambisonics, and employing for the first time the virtual microphone technique 3DVMS. Audio and video were recorded at the same time, employing 32-capsules spherical microphone arrays and a panoramic video capture system of our design. In both cases, a matrix of FIR filters was employed, for deriving the standard 4 B-format components (Ambisonics), or 32 highly-directive virtual microphones pointing at the same directions of the 32 loudspeakers (3DVMS).
A pool of test subjects was employed for comparative listening tests, evaluating some standard psycho-acoustical parameters. Furthermore, the same tests were repeated with and without the accompanying panoramic video.
The tests were performed inside the 3Sixty room in the University of York, an immersive space with all-around video projection and a 32 speakers array. A complete set of IRs has been measured placing a microphone in the center of the room and sending a sine sweep test signal to each of the 32 loudspeakers. These IRs have been employed for equalizing individually each loudspeaker for Ambisonics and 3DVMS playback. In addition to this, we experimented with the capture of a live performance inside the room and with the virtual reconstruction of the complete audio-visual experience.

INTRODUCTION

The goal of this work is very practical: to assess comparatively the 3D sound reproduction capability of the 3Sixty room, employing the traditional 1st-order Ambisonics method (considered the reference), and a modern, high resolution method which does not rely anymore on the complex math related to spherical harmonics and the like, but instead employs the simple, yet powerful, approach known as “Virtual Microphone”.
In particular, in this case the 3DVMS method was employed [1,2,3]: this is a “theory-less” approach, based on the numerical inversion of a massive matrix of impulse responses, measured when the microphone array receives a test signal coming from hundredths of different directions.
The 3DVMS method, indeed, has not been employed yet for attempting to reproduce a complete 3D soundscape inside a properly-equipped listening room: the virtual microphones obtained by this technique have only been employed, till now, as if they were the signals coming from a number of real microphones, for example for capturing the various sections of an orchestra. These signals were sent to a digital mixer, and subsequently processed in a quite traditional approach (exactly as if the were real microphones capturing different sound sources).
In this work, instead, we attempted to derive a “complete set” of virtual microphones, covering the spherical horizon as uniformly as possible, in such a way that the sound being radiated by the corresponding loudspeakers merges in a realistic reproduction of the original Soundfield.

Again, this processing is NOT based on any complex mathematical theory: the initial plan was to test two competing pragmatic methods, but the short time available forced us, for now, to employ only the first of the following two approaches:

"empirical" approach, creating virtual microphones pointing in the same directions as the corresponding loudspeakers, as "seen" from a notional point located at the center of the listening room, and all having the same directivity (4th-order pure cardioid)

"inverse matrix" approach, which is substantially the same approach employed for inverting the matrix of the impulse response measured over the spherical microphone array, for deriving the "encoding" filters. In this case, we want to create a set of "decoding" filters, and this can be done inverting the matrix of impulse responses measured inside the listening room

It is well known, for example by the work of Bruce Wiggins [4] for optimizing Ambisonics decoders, that the first approach is suboptimal, and works perfectly only in the rare case of a loudspeaker rig shaped as a

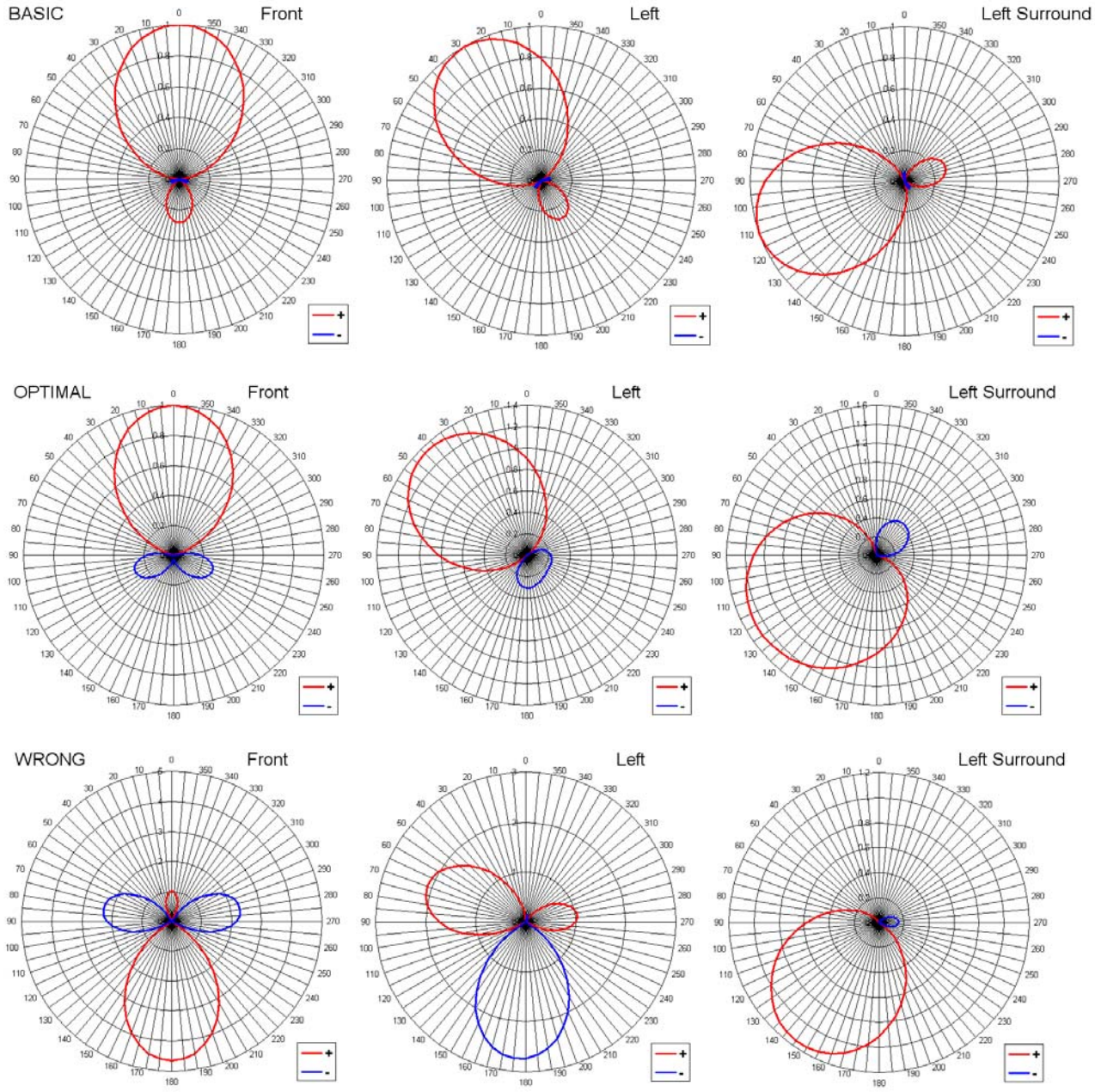


Fig. 1 - Polar patterns of the virtual microphones for 2nd-order Ambisonics decoders, from [6]

perfectly-regular polyhedron. Indeed, this approach is very simple and very robust, so we wanted to give it a try.
Of course, whenever the loudspeaker rig is geometrically irregular (as it happens to be inside the 3Sixty room), it is advisable to employ virtual microphones which have different directivity patterns (narrower where the loudspeakers are closer each other, and possibly asymmetrical when the angular distance between loudspeakers is not uniform). Furthermore, the optimal aiming of these irregular virtual microphones is not, in general, exactly in the same direction as the corresponding loudspeaker.
In a theory-less approach, the optimization of directivity pattern and aiming of the virtual microphones can be obtained employing trial-and-error, or other "heuristic" algorithms well known in the field of operational research.
It is important to understand that, in general, a direct, "brute force" approach to the matrix inversion can result in completely crazy decoding filters. A well known example of such a failure is represented by the second-order 2D Ambisonics decoding coefficients for the ITU 5.1 "surround" loudspeaker layout developed by Richard Furse [5]: it was shown in [6] that these coefficients correspond to a set of 5 virtual microphones which display a completely wrong polar pattern, albeit they exhibit the nominally-correct value (1) in the direction of the corresponding loudspeaker, and zero in the direction of the other 4.
The following figure, taken from [6], compares the three sets of virtual microphones corresponding to the **basic**, constant directivity approach (Gerzonics' Decopro plugin), to the "**optimal**" approach of Bruce Wiggins (Wigware plugin) and to the "**wrong**" matrix inversion approach of Richard Furse (as implemented in the Gerzonic's Emigrator plugin).

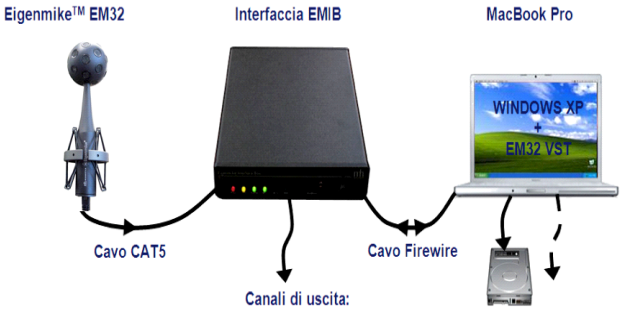


Fig. 2 – the Eigenmike™ panoramic recording system

HARDWARE SYSTEM

The system is composed of a 32-capsules spherical microphone array, and of a parabolic-mirror panoramic video camera. We tested two different units for each:

- Eigenmike™ professional microphone array from MH Acoustics [7]
- DIY spherical microphone array developed by us (with some external help) employing cheap capsules (Knowles) and ADC preamplifiers/converters (Berhinger)
- 0-360 professional glass parabolic mirror, mounted on 2Mp, Carl-Zeiss optics webcam (Logitech)
- Sony Bloggie camcorder, with its own parabolic mirror for panoramic capture

The Eigenmike

The Eigenmike is the first commercially-available, broadcast-quality spherical microphone array, developed by Gary Elko at MH Acoustics. As shown in the following figure, the system is made of a microphone probe which includes, inside the 80mm sphere, also preamplifiers, AD converters, and a audio-over-ethernet chipset. At the other end of the network cable (which also carries power and control signals to the microphone), an EMIB Firewire interface allows to connect to a PC/Mac/Linux computer by means of standard ASIO or Jack drivers.

The "virtual microphone" software coming with the Eiegnemike, indeed, is quite unsatisfactory for a number of reasons: first of all, it is controlled by means of sliders, which are definitely impractical for setting directivity and aiming of a large number of virtual microphones.
Second, it allows to process up to 16 virtual microphones maximum.
Third, it operates in realtime only if connected with the Emib interface.
Fourth, the virtual microphones have very limited control on their directivity patterns.
And fifth, it operates by means of 3rd-order Ambisonics formulation, which has severe constraints (bandwidth, S/N ratio) when a virtual microphone with very sharp polar pattern is being synthesized.
For these reasons, we preferred to employ a different approach, processing the Eigenmike recordings by means of a set of 32x32 FIR filters, and making use of the free Xvolver VST plugin [8] for performing this in realtime.

The DIY Microphone Array

The experimentation described in this paper was realized mostly using a microphone array of our production. As shown in Figure 3, this microphone is a sphere of expanded polyurethane with 32 capsules placed on its surface. These capsules were originally intended for hearing-aids (Knowles Electronics), therefore their dynamic response is more suited for

speech than music, albeit their frequency response is almost flat over a wide frequency range. This is a prototype, and therefore has certain problems. The main chassis is built in a light, cheap and easy to work material, which is also quite fragile, and requires caution when handling.
The preamplification is done externally, by means of a rack of 4 Behringer AD-8000 converters. As the unbalanced multicore connection cable is too long, it easily captures noise and electrical disturbances.
Finally, the calibration procedure is time consuming and quite delicate, but it is required to be performed very often, as the gain knobs of the converters loose their setting quite easily.

Fig. 3 – the DIY spherical microphone array (close-up)

Fig. 4 – calibration of the DIY spherical microphone by means of an earplug

The 0-360 Parabolic Mirror Camera

This unit was developed in order to be employed in realtime together with the "instant steering" software developed at the University of Parma ofr the RAI research center [1,2,3]. However, we managed to borrow this unit, together with one of their Eigenmikes, for doing a number of audio-video "location recordings" in the town of Barcelona, Spain.

The mirror and the Logitech hi-res camera are mounted inside a plexiglass tube, for protection against dust and finger oil, as shown in fig. 5. The Eigenmike is mounted just above the mirror.

Fig. 5 – the 0-360 parabolic mirror camera mounted under the Eigenmike

The image captured by the camera is a "donut" image, as shown in fig. 6. We developed a specific software tool, written in the "Processing" JAVA-based scripting language, for performing realtime "de-warping" of the "donut" image to a standard rectangular video with 2:1 aspect ratio, as shown in fig. 6.

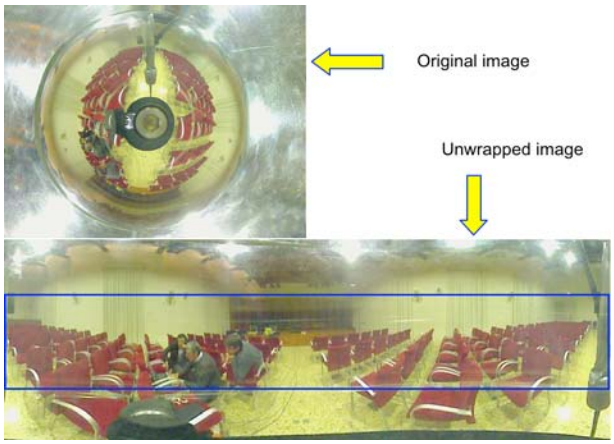


Fig. 6 – unwrapping the "donut" image to a standard panoramic video image

The angular coverage of this system is 360° horizontally by 120° (+/- 60°) vertically, resulting in a video footage which covers great part of the spherical horizon.
Unfortunately, this huge vertical range is not going to be useful inside the 3-Sixty room, which is equipped of a multiple video projection system with a total aspect ratio of 64:10 (4 beamers at 16:10 each). Of consequence, we needed to cut a thin slice in the unwrapped video, as shown in the rectangle overplotted

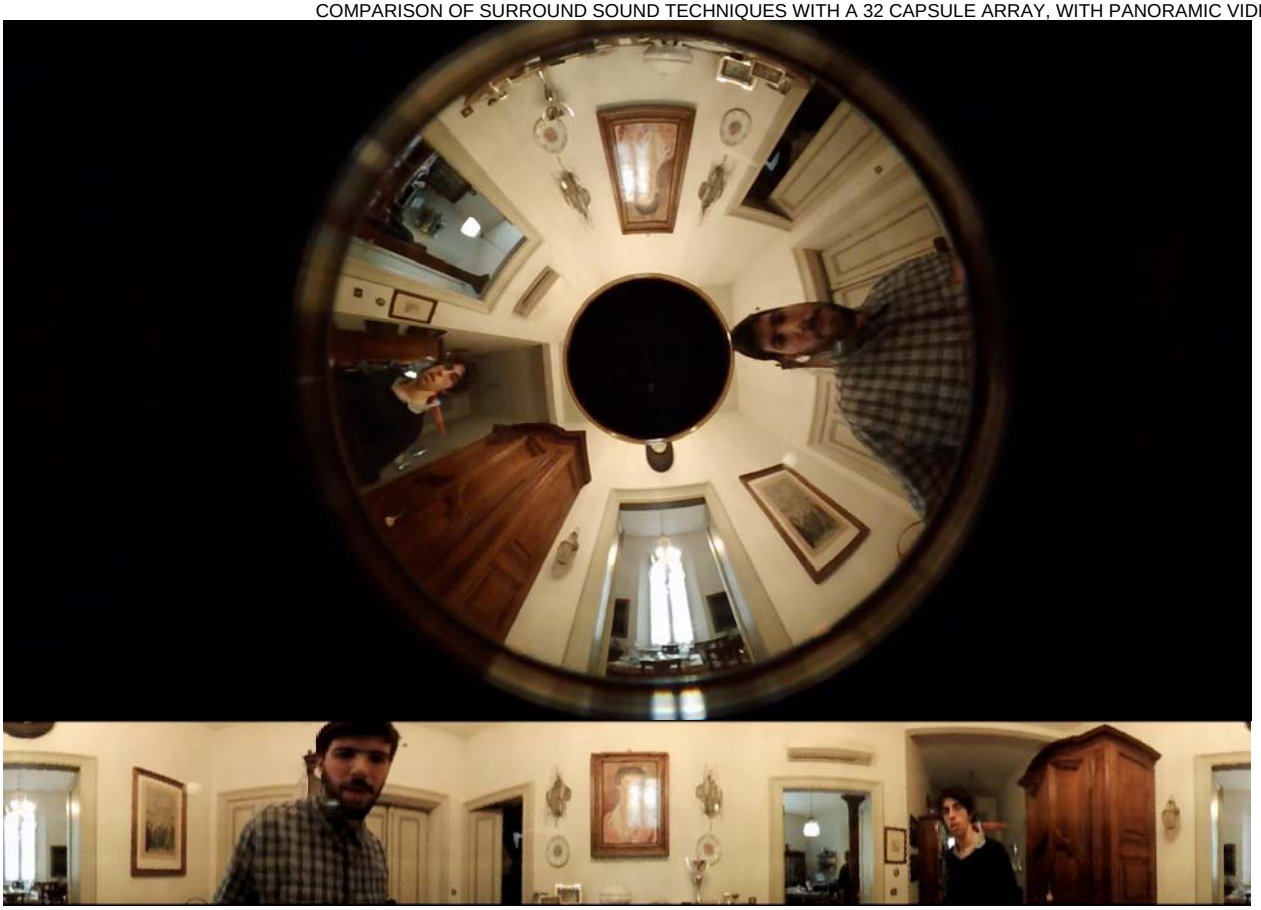


Fig. 8 – “donut” and unwrapped images – Sony Bloggie software

in fig. 6, wasting more than half of the precious pixels captured..
This fact, jointly with the low resolution employed during the location recordings (just 800x600, albeit the camera was capable of a much more impressive 1600x1200), resulted in a quite poor quality of the video rendered inside the 3-Sixty room, despite the cost of this professional optical system.

The Sony Bloggie Camera

For making easily panoramic video recordings together with our DIY microphone probe, we purchased a cheap camcorder, the Sony Bloggie MHS-TS20K, as shown in fig. 7. It is an inexpensive full-HD camcorder, which comes with a small parabolic mirror for panoramic video recordings.
The camera also includes software for unwrapping the donut video image to standard wide-ratio rectangular video at 1280x720 resolution (with letterboxing, the real video image is 1280x192, which matches almost perfectly the 64:10 aspect ratio of the 3-Sixty room).

Fig. 7 – the Sony Bloggie panoramic camcorder

For location recordings it was necessary to build an enclosure cage for the DIY spherical microphone, for

protecting its delicate sphere, for mounting a suitable windscreen, and for suspending the Bloggie camcorder just above it, as shown in fig. 9.

Fig. 9 – the Sony Bloggie mounted above the DIY spherical microphone array

SOFTWARE SYSTEM

The standard software employed both for recording and for processing the signals was Plogue Bidule, employing the free VST plugin Xvolver, capable of performing massive convolution with a matrix of FIR filters of size up to 32x32. This software is available both for Windows and for Mac, but we prefer the latter, due to better stability and performances.
The recordings were always performed in the W64 file format, which is mandatory for allowing the files size to exceed 2 Gbytes (and recording 32 channels at 48 kHz, 24 bits, in a single file, this happens quite easily...).
The matrix of FIR filters required both for Ambisonics processing and for 3DVMS processing were computed by means of a small Matlab script, which loads the matrix of measured impulse responses of the chosen microphone array, and computes the matrix of FIR filters which transform the 32 signals coming form the capsules in the single signal of the virtual microphone having the chosen directivity and aiming.

$$y_v(t) = \sum_{m=1}^{M} x_m(t) * h_{m,v}(t)$$

It must be noted that this way the FIR filters are static, whilst the most advanced software described in [1,2] allows for dynamic change of the filtering coefficients, with real-time control by means of the mouse or the joystick.
But, for the playback of the recordings over a static set of loudspeakers, a static set of filters is required...

Ambisonics decoding

In this case we need to generate just 4 virtual microphones, which have the standard directivity patterns known as “B-format” (WXYZ), as shown in fig. 10.

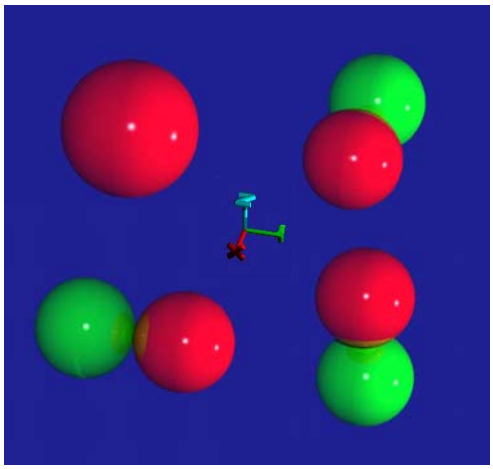


Fig. 10 – polar patterns of the 4 virtual microphones (1st order Ambisonics)

Fig. 11 shows the set of 32x4 FIR filters employed for converting the signals coming form the capsules to these 4 standard output signals WXYZ:
It can be seen how all the 32 capsules contribute almost equally to the omnidirectional (W) signal, whilst their contribution to the 3 directive virtual microphones (X,Y,Z) is quite different.
Finally, fig. 12 shows X-Volver, whilst performing the B-format conversion in realtime.

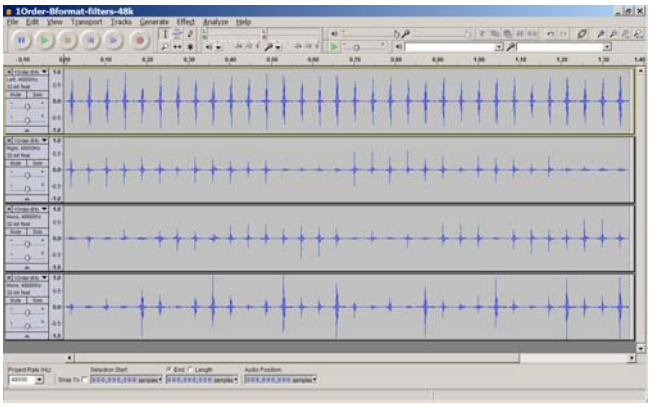


Fig. 12 – The 32x4 FIR processing filters

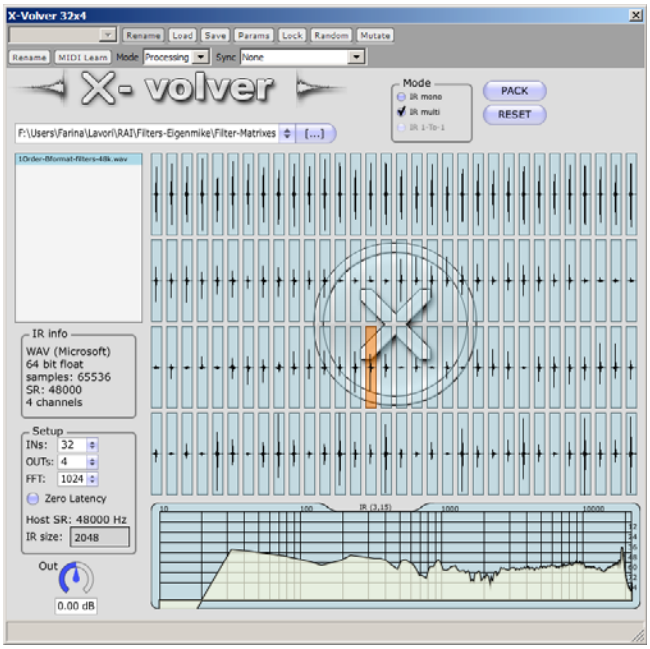


Fig. 12 – X-volver processing 32 inputs to 4 outputs

After the 4 standard signals have been obtained, they are sent to a traditional Ambisonics decoder (Gerzonic Decopro) which feeds just 16 loudspeakers, chosen as follows:

- 4 loudspeakers in the lower ring
- 8 loudspeakers in the medium ring
- 4 loudspeakers in the top ring

The selected loudspeakers are outlined in fig. 13

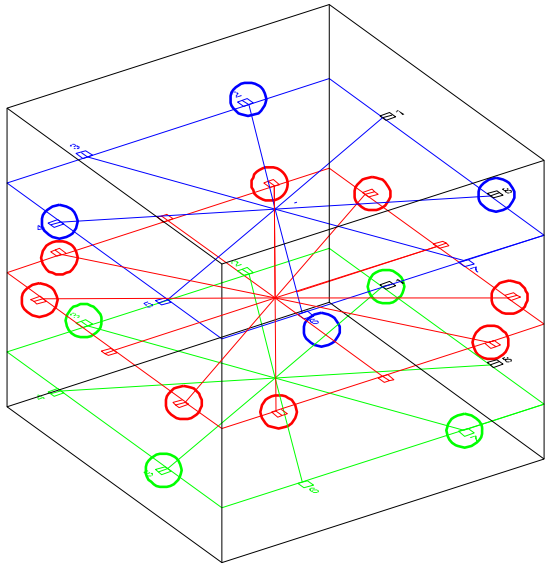


Fig. 13 – 16 loudspeakers selected for 1st order Ambisonics among 28 main speakers

Entering the 3D coordinates of the 16 selected loudspeakers, Decopro automatically configures the decoding coefficients for the selected geometry.

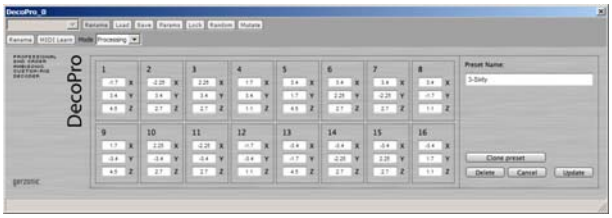


Fig. 14 – Decopro feeding 16 loudspeakers for the 3-Sixty room

The order of the loudspeakers has been selected so that they are wired at odd numbers of the standard channel numbering of the 3-sixty room, that is, at outputs number 1, 3, 5...31.
Fig. 15 shows the patch employed for Ambisonics processing and playback in the room 3-sixty.

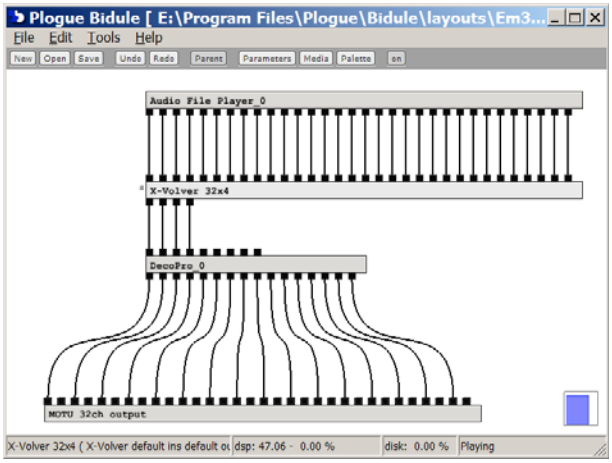


Fig. 15 – The processing patch in Plogue Bidule for Ambisonics decoding

3D – VMS decoding

Our current Matlab script allows for the creation of a number of virtual microphones (up to 32).
The directivity pattern is always of the Cardioid type (no side or rear lobes), according to the following equation:

$$Q_1(\vartheta,\varphi) = (0.5 + 0.5\cdot\cos(\vartheta)\cdot\cos(\varphi))^n$$

Where n is the order of the cardiod.
Fig. 16 shows the polar patterns of such variable-order cardioids:

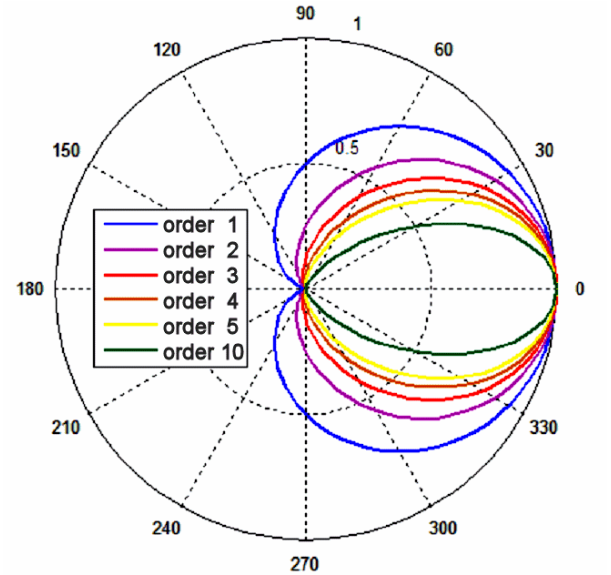


Fig. 16 – polar patterns of cardioids of order 1 to 10

After some experiments, we decided to employ 4th-order cardioids, which are a good compromise between channel separation and absence of "holes" between the directions subtended by the loudspeakers.

As our Matlab script requires to enter the directions by means of polar spherical coordinates (azimuth and elevation), it was necessary to compute these angles from the knowledge of the Cartesian coordinates of the 32 loudspeakers, as shown in the following table.

Angular and Cartesian coordinates of the 32 loudspeakers in the 3-Sixty room

N.	Azimuth (°)	Elevation (°)	Radius (m)	X (m)	Y (m)	Z (m)
1	116.5651	41.81031	5.1	-1.7	3.4	4.5
2	63.43495	41.81031	5.1	1.7	3.4	4.5
3	123.4952	21.42696	4.379783	-2.25	3.4	2.7
4	90	25.20112	3.757659	0	3.4	2.7
5	56.50482	21.42696	4.379783	2.25	3.4	2.7
6	116.5651	0	3.801316	-1.7	3.4	1.1
7	63.43495	0	3.801316	1.7	3.4	1.1
8	135	-12.8858	4.932545	-3.4	3.4	0
9	26.56505	41.81031	5.1	3.4	1.7	4.5
10	-26.5651	41.81031	5.1	3.4	-1.7	4.5
11	33.49518	21.42696	4.379783	3.4	2.25	2.7
12	0	25.20112	3.757659	3.4	0	2.7
13	-33.4952	21.42696	4.379783	3.4	-2.25	2.7
14	26.56505	0	3.801316	3.4	1.7	1.1
15	-26.5651	0	3.801316	3.4	-1.7	1.1
16	45	-12.8858	4.932545	3.4	3.4	0
17	-63.4349	41.81031	5.1	1.7	-3.4	4.5
18	-116.565	41.81031	5.1	-1.7	-3.4	4.5
19	-56.5048	21.42696	4.379783	2.25	-3.4	2.7
20	-90	25.20112	3.757659	0	-3.4	2.7
21	-123.495	21.42696	4.379783	-2.25	-3.4	2.7
22	-63.4349	0	3.801316	1.7	-3.4	1.1
23	-116.565	0	3.801316	-1.7	-3.4	1.1
24	-45	-12.8858	4.932545	3.4	-3.4	0
25	-153.435	41.81031	5.1	-3.4	-1.7	4.5
26	153.4349	41.81031	5.1	-3.4	1.7	4.5
27	-146.505	21.42696	4.379783	-3.4	-2.25	2.7
28	180	25.20112	3.757659	-3.4	0	2.7
29	146.5048	21.42696	4.379783	-3.4	2.25	2.7
30	-153.435	0	3.801316	-3.4	-1.7	1.1
31	153.4349	0	3.801316	-3.4	1.7	1.1
32	-135	-12.8858	4.932545	-3.4	-3.4	0

Fig. 17 shows the 32x32 FIR filters computed by the Matlab script, for generating 32 4th-order virtual cardioids aimed at the directions shown in the table above.

Fig. 17 – the matrix of 32x32 FIR filters, creating 32 4-th order cardioids

The nice thing of the 3DVMS approach is that no further decoding is required. The outputs of the processing are already the required speaker feeds, provided that the virtual microphones being synthesized are aimed at the correct directions.
So, as shown in fig. 18, the Plogue Bidule patch is very simple in this case.

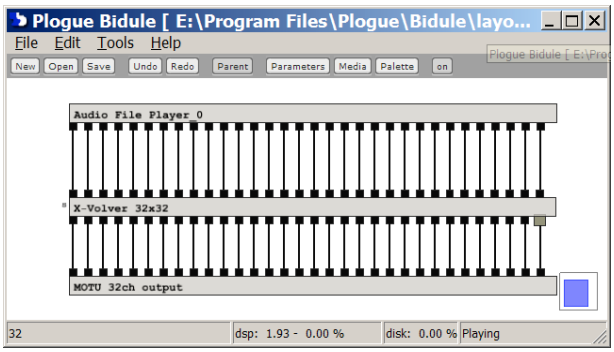


Fig. 18 - The processing patch in Plogue Bidule for 3DVMS decoding

EXPERIMENTS

Due to the difficulty of obtaining a large number of test subject, we limited ourselves to 4 cases. Starting from 3 different recordings, we generated an Ambisonics and a 3DVMS rendering. These were played back with and without the accompanying panoramic video track. The main focus of the subjective tests was on the perception of spatiality. The questions focused on where each sound source appeared to be. The recordings had the sound sources around the room, with the recording apparatus in the center. When playing it back we had the test subjects as close as possible to the original microphone location. Normal blind testing procedures were followed.

We recorded three different test examples, one purely instrumental, one completely vocal and a mix of the two (consisting in a pop band). We expected the pop band to be the most easy to be localized, because of the different timbre characteristics. In general we expected the video to be of strong help in the localization. This gave a total of 12 testing condition, resulting from 2 audio renderings of each signal and the playback with and without the support of the video.
Each test subject was made to listen to each of the 12 signals, always leaving the video version last, in order not to create a bias. Furthermore, the order was randomized. The spatialization question enquired about the location of a particular sound source in our recording. A short snippet of the source in question was played at the beginning, in order to help to identify it. We thought that people might be unable to recognize the sound of, for example, an oboe, if they were asked to localize it without hearing it first. The localization was then represented by the subjects as a cross on a map of the room. Their position was then measured and averaged. Furthermore, questions of sound quality were asked appropriately for each signal (for example, about intelligibility for speech).
We further expect the panoramic video to influence the perception of the surround sound in a different way than simply giving and removing the sense of sight to a subject in normal conditions. In fact, in the real world we always perceive sounds coming from all around us, but only see what is in our field of view. This is compounded in our expectations of an audio/visual performance by our experience of everything from theatre to cinema. In this case, however, the video is blatantly put forward as panoramic. Therefore the subjects will be much more likely to look around than they would be in normal circumstances, altering both their visual and acoustic perception of the event.

The statistical elaboration of these tests is still in progress and it will be completed in time for the presentation.

CONCLUSIONS

The 3Sixty room features an interesting approach to multimedia immersion, providing a seldom seen panoramic video playback paired with surround sound.

However, just by learning how to use it we found a couple of considerations worth pointing out.
First of all, the room seems to be designed to serve both a multimedia (video playback) and an exhibition purpose. When facing the presenter and the main screen for a presentation, for example, the vision of the side screens is deeply inhibited, and that of the back one is completely precluded. At the same time the sound system provides quite a good localization for sound and video, but it cannot easily create separate soundscapes for accompanying different parts of a walk-around exhibitions. It is therefore best suited for creating immersive experiences, in which the spectator looks around freely. An interesting example of such an experience is that created for Coca Cola in Istanbul by Boogy [9].
Such experiences are usually custom-created for the room in which they will be played back. However, if one was attempts to simply playback a panoramic video (created with a consumer system, for example) inside the 3-Sixty room, he would find the center of the image to be hidden by the edge between two screens. It could be very interesting to create a draft specification for the compatibility of rooms of this kind, allowing to transport easily audio-video presentations among them.
This, indeed, would require reconfiguring the 3-Sixty room, for making it more consistent with standard audio-video practice (video centered in the middle of the "main" screen, more standard locations for loudspeakers and for audio channel ordering).

REFERENCES

1. A. Capra, L. Chiesi, A. Farina, L. Scopece, 2010. A spherical microphone array for synthesizing virtual directive microphones in live broadcasting and in postproduction. Proceedings of 40th AES International Conference, Spatial audio: sense of the sound of space, Tokyo, Japan, October 8-10 2010
2. L. Scopece, A. Farina, A. Capra. 360 Degrees Video And Audio Recording And Broadcasting Employing A Parabolic Mirror Camera And A Spherical 32-Capsules Microphone Array - IBC 2011, Amsterdam, 8-11 September 2011
3. A. Farina, M. Binelli, A. Capra, E. Armelloni, S. Campanini, A. Amendola. Recording, Simulation and Reproduction of Spatial Soundfields by Spatial PCM Sampling (SPS) - International Seminar on Virtual Acoustics, Valencia (Spain). 24-25 November 2011
4. Wiggins, B. The Generation of Panning Laws for Irregular Speaker Arrays Using Heuristic Methods. Proceedings of the 31st International AES conference, London, UK (2007).
5. http://www.muse.demon.co.uk/ref/speakers.html
6. http://pcfarina.eng.unipr.it/Public/B-format/5_1_conversion/5_1_decoders.htm
7. http://www.mhacoustics.com
8. http://pcfarina.eng.unipr.it/Public/Xvolver/
9. https://vimeo.com/36321631

EXAMPLE QUESTIONNAIRE

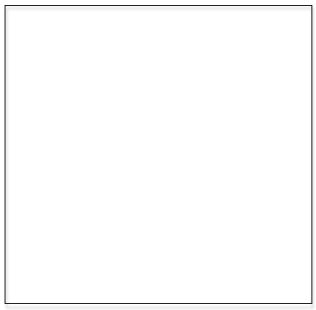
Appendix 1

Test signal #1

Subject name: _____

You will now hear a brief example of a sound source, which will be part of the complete test signal.

In the following square, mark with a cross the approximate position where you think the sound from the example originated.



Rate according to the following attributes

Like						Dislike
Intelligibile						Confused
Crisp						Messed up
Natural						Forced

Don't hesitate do ask for clarification of any parameter, preferably before the test begins.