# SPATIAL PCM SAMPLING: A NEW METHOD FOR SOUND RECORDING AND PLAYBACK

**ANGELO FARINA[1], ALBERTO AMENDOLA[1], LORENZO CHIESI[1], ANDREA CAPRA[1], SIMONE CAMPANINI[1]**

[1] *Industrial Eng. Dept., University of Parma, ITALY*
angelo.farina@unipr.it

This paper presents the mathematical and physical framework of a new technology, named SPS (Spatial PCM Sampling): it is the equivalent, in a two-dimensional spherical-coordinate space, of the traditional PCM representation of a waveform (in the one-dimensional time domain). It is nowadays possible to record an SPS multichannel stream (also called P-format) by processing the signals coming from massive microphone arrays, now widely employed in the broadcasting industry and in research labs. Some types of sound processing are easy when operating on P-format signals; some, indeed, require more work. At playback, it is possible to drive loudspeaker arrays of arbitrary shape and complexity, providing in general better spatial accuracy than competing well known methods, such as Ambisonics or WFS.

## INTRODUCTION

The goal of "immersive" sound systems, employed both for recording and playback, is to capture the complete spatial information when recording and to replicate it faithfully when playing back.

Several methods were developed in the past for attempting to reach this goal, based mostly on two mathematical frameworks:

- the Ambisonics theory, which expresses the spatial information of the sound field at a single point in space, by means of a number of signals equivalent to a number of virtual microphones possessing very complex polar patterns, corresponding to spherical harmonics functions: These signals can later be recombined by means of a "decoder", which provides a number of signals for feeding the loudspeakers employed in the playback system.
- the WFS theory, in which the sound field is sampled by pressure or velocity microphones at a large number of points, covering a closed surface: later these signals are processed for feeding a corresponding array of loudspeakers, again placed on a closed surface, possibly different from the recording surface.

The method proposed here is related more strictly with the Ambisonics approach, but in this case, instead of employing spherical harmonics, a number of "spatial Dirac's Delta functions" are employed as the kernel of the system, and each signal represents consequently a virtual unidirectional microphone pointed in a direction. The whole spherical surface is sampled more or less uniformly, employing dozens of these ultradirective unidirectional microphones [1].

The resulting spatial sampling process is in some way similar to the decomposition of the original sound field in plane waves: as plane wave decomposition is the basis of WFS, it can be seen that SPS (Spatial PCM Sampling) is in between the two traditional methods.

In this paper the theory of SPS is first explained, then it is shown how to record high quality P-format signals by means of a 32-channels spherical microphone array (Eigenmike™), how to perform basic sound editing on the P-format signals, and how to create a suitable sound playback system fed with these signals.

## 1 VIRTUAL MICROPHONES

The concept of “virtual microphones” is very powerful, and it can be applied to almost any multichannel recording and playback system, either for describing the capture of the sound in the original space (or its “encoding”, when a synthetic sound field is created) and for describing the playback through a multi-loudspeaker system.

The basis of this concept is very simple: every signal, either at capture or playback, can be thought as the signal captured by a microphone placed in a specific position, with a given aiming, and a given directivity pattern (and all of this is, possibly, frequency-dependent).

In some case the signal is **really** coming from such a microphone. But in most cases, due to intermediate processing, each signal represent a **virtual** microphone, which is obtained as the mathematical combination of the signal coming from a number of physical microphones, or from a pure mathematical synthesis of an imaginary sound field.

If this is easy to understand at capture stage, the concept is very powerful also at playback stage. Whatever comes before, at the end each loudspeaker is fed with just one electrical signal: this can always be thought as the signal captured by a microphone system, which can be simple or complex.
The concept of virtual microphones is very powerful when analyzing the behaviour of a complex recording/playback system, for checking that everything works reasonably well. Let's proof this with a well-known example, that is 2nd order horizontal Ambisonics reproduction over an ITU 5.0 loudspeaker array according to the "exact" decoding formulas of Richard Furse [2].

### 1.1 The "photocopy of the photocopy" concept

This is also a good example for introducing another powerful concept, that is the "photocopy of the photocopy": as with reprographic machines, an "ideal" system replicates a copy of an image which, if further replicated, is indistinguishable form the first copy.
When this concept is applied to audio systems, and in this particular case to the playback of the 5-channels 2nd-order Ambisonics signal over an ITU 5.0 loudspeaker array, we expect that the system replicates faithfully the same 5 Ambisonics signals if we place a 2nd-order Ambisonics microphone at the centre of the loudspeaker array, as shown in fig. 1.

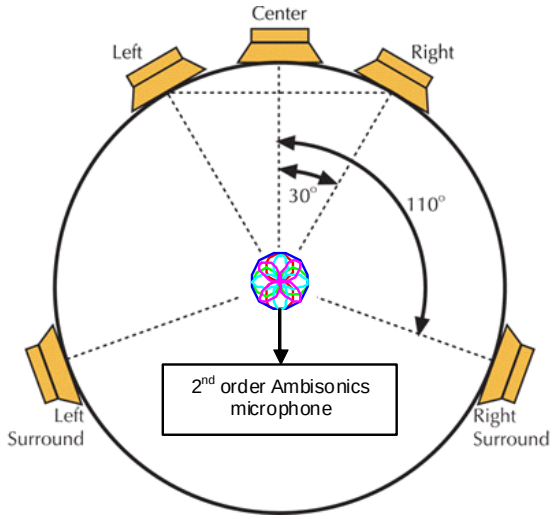


Figure 1: Re-recording 2nd-order Ambisonics signals at the centre of an ITU 5.0 loudspeaker rig.

An "exact" decoding scheme could be thought as the set of Ambisonics decoding coefficients which ensure that the 5 Ambisonics signals captured by the microphone at the centre of the reproduction rig are identical to the original Ambisonics signals.
In general, each speaker feed ***s*** of an Ambisonics system is simply the weighted sum of the 5 Ambisonics signals (*W,X,Y,U,V*):

$$s = g_w \cdot W + g_x \cdot X + g_y \cdot Y + g_u \cdot U + g_v \cdot V \quad (1)$$

So, the complete set of the decoding coefficients can be packed in a square 5x5 matrix (one row for each loudspeaker). According to Furse [2], the following set of coefficients provides the "exact" decoding solution for the standard ITU layout:

| Name | Angle | gW | gX | gY | gU | gV |
|---|---|---|---|---|---|---|
| L | 30 | 0 | 1.366 | 0.366 | -1.366 | 0.366 |
| R | -30 | 0 | 1.366 | -0.366 | -1.366 | -0.366 |
| C | 0 | 0.4714 | -1.8214 | 0 | 2.488 | 0 |
| LS | 120 | 0.4714 | -0.4553 | 0.366 | 0.122 | -0.2113 |
| RS | -120 | 0.4714 | -0.4553 | -0.366 | 0.122 | 0.2113 |

Table 1: Furse's decoding coefficients.

This set is "exact", in the sense that, if we consider 5 plane waves, each one arriving at the microphone exactly from the same direction of each loudspeaker, their signal will appear just as the feed of the corresponding speaker, whilst all the other 4 loudspeakers are muted.
However, if we look at how this "exact" decoding scheme behaves for wavefronts arriving from every other direction, we discover that the system behaves erratically. This is clearly understood if we look at the polar patterns of the virtual microphones obtained applying the decoding coefficients of table I to the standard Ambisonics signals.
For example, the resulting pattern for the signal feeding the L loudspeaker is:

$$L = 1.366 \cdot \cos(\vartheta) + 0.366 \cdot \sin(\vartheta) - 1.366 \cdot \cos(2 \cdot \vartheta) + 0.366 \cdot \sin(2 \cdot \vartheta) \quad (2)$$

The following figure 2 shows the polar patterns for C, L and LS:

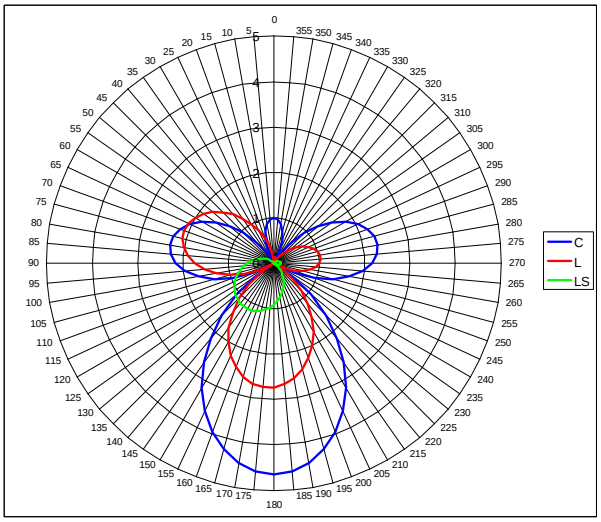


Figure 2: virtual microphones for Furse's decoder.

It is quite obvious how these virtual microphones will feed each loudspeaker with "crazy" signals. C, for example, captures very loud sound from **behind** the microphone, with a gain close to 5, whilst the sound from the frontal direction has gain 1…
This example, which appears in some way out of the scope of this paper, is indeed important for two reasons:

1. It shows how looking at the virtual microphones provides a physically-meaning picture of the behaviour of the system.
2. It demonstrates that in some cases a theoretically perfect solution is bad in practice, and this can happen easily when the computation is based on some kind of "brute force mathematical inversion".

### 1.2 "Theoryless" virtual microphones

In a previous work, we described the technology employed for deriving "virtual microphone" signals from a massive multi-capsule microphone array, without the need of solving complex equations [3].
A "virtual microphone" signal $y_v$ can be obtained as the filtered sum of $M$ "real microphone" signals $x_m$, starting from a "spatial sampling" of the sound field performed employing an array of $M$ microphones, at different locations and aiming:

$$y_v(t) = \sum_{m=1}^{M} x_m(t) * h_{m,v}(t) \tag{3}$$

Where * denotes convolution.
A "theoryless" approach is employed for obtaining the filtering coefficients $h_{m,v}$ for any virtual microphone $v$, with prescribed directivity and aiming, by imposing that its measured polar pattern deviates minimally from the ideal one.
In practice, the microphone array is first subject to a large number of anechoic impulse response measurements, from many directions, covering the whole spherical surface. Let's call $C_{m,d}$ the matrix containing the measured impulse responses, from $D$ directions and $M$ microphones.
For any of these $D$ directions, and at any frequency, the virtual microphone which we want to obtain should provide a "nominal" target gain $p_d$

$$\sum_{m=1}^{M} c_{m,d} * h_m \Rightarrow p_d \qquad d = 1..D \tag{4}$$

Of course it will be impossible to obtain **exactly** the prescribed directivity $p_d$, but a least-square system can be set up for searching the set of filtering coefficients $h_m$ which better approximate the wanted result.
As the resulting filtering coefficients are derived from measurements performed on the actual microphone array, these filters will not only provide the required directivity pattern, but they will also compensate for deviations from ideality of the magnitude and phase responses of individual transducers, and for shielding/diffraction/resonance effects caused by the mechanical structure of the array.
Please notice that in practice the target impulse responses $\boldsymbol{p_d}$ are simply obtained applying a direction-dependent gain $Q_d$ to a delayed unit-amplitude Dirac's delta function $\delta$:

$$p_d = Q_d \cdot \delta \tag{5}$$

Computation is easier in frequency domain (that is, computing the complex spectra, by applying the FFT algorithm to the $N$-points-long impulse responses $\boldsymbol{c}$, $\boldsymbol{h}$ and $\boldsymbol{p}$). Let's call $\boldsymbol{C}$, $\boldsymbol{H}$ and $\boldsymbol{P}$ the resulting complex spectra. This way, the convolution reduces to simple multiplication between the corresponding spectral lines, performed at every frequency index $k$:

$$\sum_{m=1}^{M} C_{m,d,k} \cdot H_{m,k} \Rightarrow P_{d,k} \quad \begin{cases} d = 1..D \\ k = 0..N/2 \end{cases} \tag{6}$$

Now we pack the values of $\boldsymbol{C}$, $\boldsymbol{H}$ and $\boldsymbol{P}$ in proper vectors or matrixes, taking into account all the $M$ input microphones, all the measured directions $D$ and all the $V$ outputs to create:

$$\left[H_k\right]_{MxV} = \frac{\left[P\right]_{DxV}}{\left[C_k\right]_{DxM}} \tag{7}$$

This over-determined system doesn't admit an exact solution, but it is possible to find an approximated solution with the Least Squares method, employing a regularization technique for avoiding instabilities and excessive signal boost [3]. The block diagram of the least-squares method is shown in figure 3:

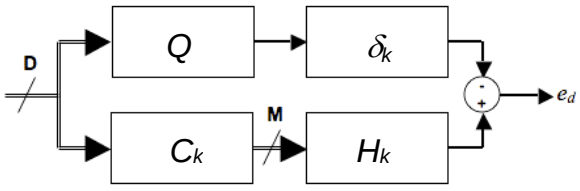


Figure 3: scheme of the Least Squared method with a delay in the upper branch.

In this scheme we observe the delay block $\delta_k$ required for producing causal filters, and the resulting total modelling error $e_d$, which is being minimized by the least-squares approach
Albeit various theories have been proposed for defining the optimal value of the causalisation delay $n_0$, we did take the easy approach, setting $n_0=N/2$. Choosing $N/2$ samples is a safe choice, which creates inverse filters

with their "main peak" close to their centre, and going smoothly to zero at both ends.
Furthermore, a regularization parameter is required in the denominator of the matrix computation formula, to avoid excessive emphasis at frequencies where the signal is very low.
So the solution formula, which was first proposed in Kirkeby et al. [4], becomes:

$$[H_k]_{MxV} = \frac{[C_k]^*_{MxD} \cdot [Q]_{DxV} \cdot e^{-j\pi k}}{[C_k]^*_{MxD} \cdot [C_k]_{DxM} + \beta_k \cdot [I]_{MxM}} \quad (8)$$

As shown in the image below, the regularization parameter $\beta$ should depend on frequency [5]. A common choice for the spectral shape of the regularization parameter is to specify it as a small, constant value inside the frequency range where the probe is designed to work optimally, and as much larger values at very low and very high frequencies, where conditioning problems are prone to cause numerical instability of the solution.

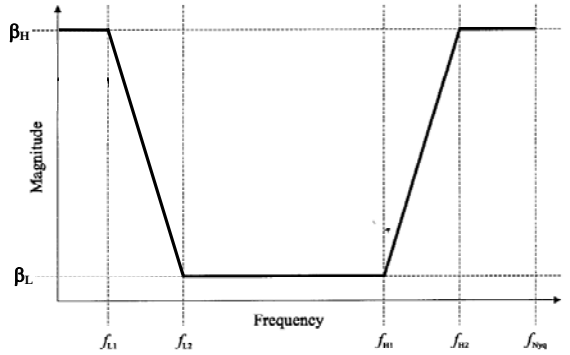


Figure 4: frequency-dependent regularization parameter.

### 1.3 Sets of virtual microphones

In this paper we focus on two possible sets of virtual microphones, for processing the same 32-channels raw signals ("A-format") coming form the first commercially-available spherical microphone array, the Eigenmike™, as shown in figure 5.

Figure 5: The Eigenmike™ microphone array.

- A set of 16 virtual microphones having directivity patterns given by the spherical harmonic functions of order 0, 1, 2 and 3, as shown in figure 6.
- A set of 32 virtual microphones having 4th-order cardioid patterns, pointing in the same directions as the 32 capsules of the Eigenmike, as shown in fig. 7, 8 and 9.

It must be noted that the first set of virtual microphones produces what is normally known as an High Order Ambisonics signal (HOA), a.k.a. B-format.

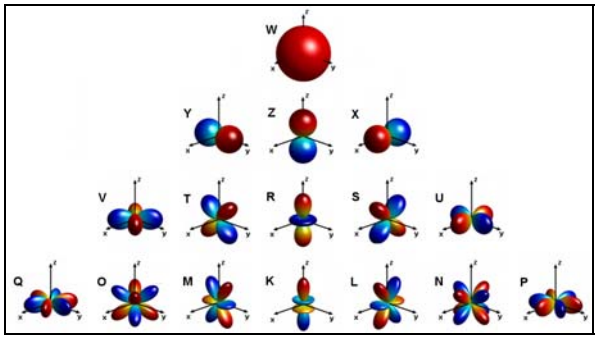


Figure 6: polar patterns of the 16 virtual microphones for HOA.

The second set of 32 virtual microphones, instead, is a first, rough approximation to Spatial PCM Sampling, and hence the resulting signal is named SPS or, simply, P-format.
The chosen polar pattern for these 32 virtual microphones is a 4th-order cardioid, defined by:

$$Q_n(\varphi) = [0.5 + 0.5 \cdot \cos(\varphi)]^4 \quad (9)$$

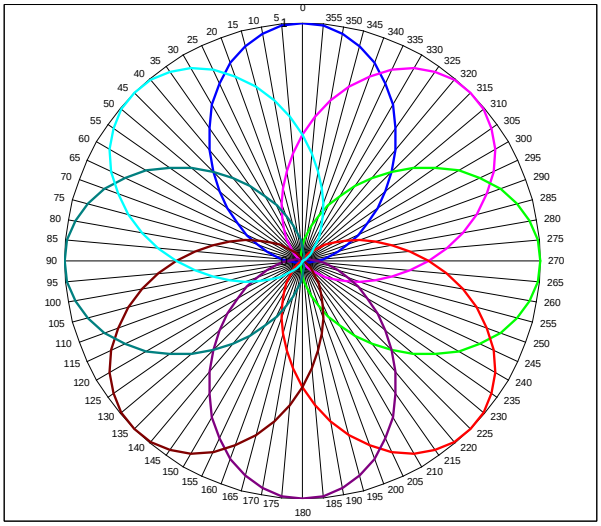

Figure 7: polar patterns of 8 adjacent 4th-order cardioids (theoretical).

They provide just the correct amount of overlap between adjacent microphones, and do no exhibit spurious side or rear lobes.

The 32 4th-order cardioids are pointed exactly in the same directions of the 32 capsules fitted in the Eigenmike™.

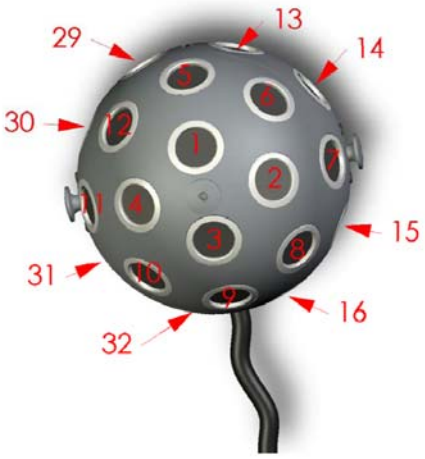


Figure 8: positions of the 32 capsules of the Eigenmike™.

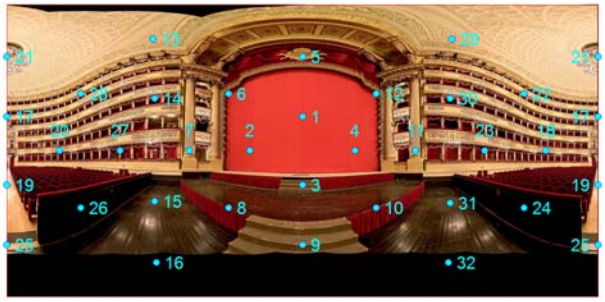

Figure 9: position of the axis of the 32 virtual microphones plotted over a 360°x180° panoramic image (courtesy Teatro alla Scala, Milan).

The number of virtual microphones being synthesized, in these two cases, is hence quite large (16 or 32). Typically, each filter is at least 2048 samples long (at 48 kHz sampling rate). Each virtual microphone, thus, requires summing the results of the convolutions of 32 input channels with 32 FIR filters. And for getting all the required 16 or 32 virtual microphone outputs, we need to convolve-and-sum over a matrix of 32x16, or 32x32, FIR filters, each of 2048 samples.

For performing these massive multichannel filtering operations, a special VST plugin was developed, called X-volver, and running either on Mac or Win32 platforms; this plugin is freely available in [6]. Fig. 10 shows the X-volver plugin being used inside Audio Mulch, a multichannel VST host program: a 32x32 filter matrix is being employed for converting the signal coming form the 32-capsules spherical microphone array to the 32 SPS signals.

A modern laptop, equipped with at least an Intel i5 processor, can easily perform such filtering in realtime, during the recording.

Nevertheless, we usually prefer to always record the "raw" 32-channels coming from the capsules, for being able subsequently to reprocess them with different sets of filters, or for deriving directly other types of virtual microphones.

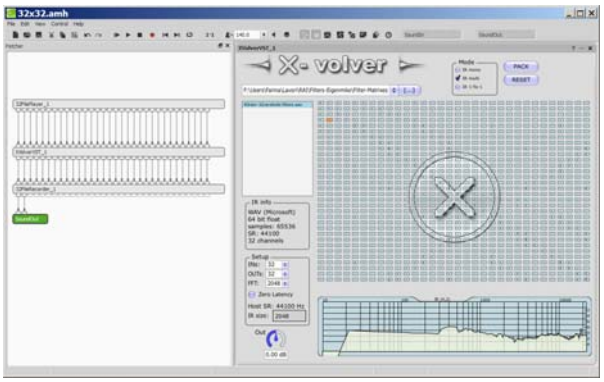


Figure 10: Graphical User's Interface of X-volver, inside the Audio Mulch host program.

### 1.4 Experimental verification

For evaluating the real behaviour of these virtual microphones, the Eigenmike was installed over an automated turntable inside an anechoic room, and a set of impulse responses was measured on the horizontal plane with 5° steps, employing a coherent point source loudspeaker (Tannoy dual concentric monitor).

The following figure 11 shows the real polar patterns measured for the 8 virtual microphones lying on the horizontal plane (n. 2, 7 27, 20, 18, 23, 11, 4) in different octave bands.

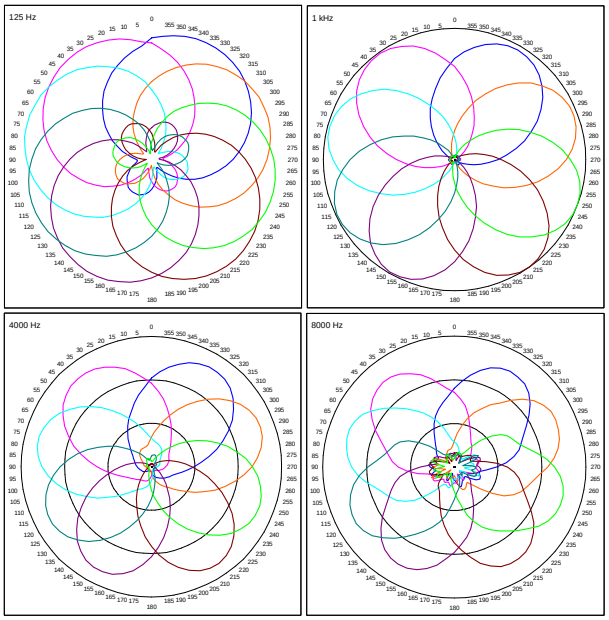


Figure 11: polar patterns of 8 adjacent 4th-order cardioids (experimental).

## 2 SPATIAL PCM SAMPLING

Spatial PCM Sampling (SPS) is the equivalent (in space), to the representation of a waveform (in time) as a sequence of impulses of proper amplitude (PCM, pulse code modulation).
Conversely, it can be seen as High Order Ambisonics is the equivalent (in space) as the Fourier analysis (representation of a complex waveform as the summation of a number of sinusoids and cosinusoids, each with proper gain).
The 32 superdirective virtual microphones described in chapter 1.3 perform an approximate spatial PCM sampling, as each of them can be thought as having a directivity pattern approximating a "spatial Dirac's Delta function".
Fig. 12 compares the standard PCM representation of a waveform in time with the "spatial PCM" representation of a directivity balloon in space.

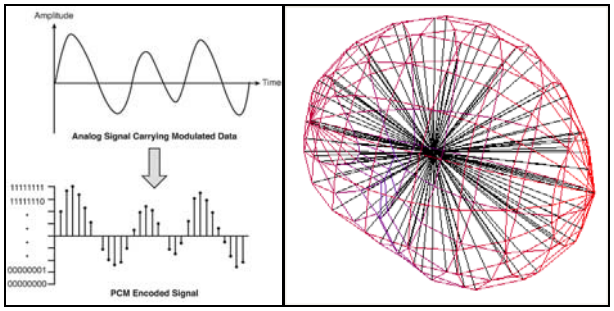


Figure 12: PCM sampling of a waveform in time (left) and of a balloon in space (right).

Fig. 13, instead, shows the reconstruction of a waveform (in time domain) or of a spatial directivity balloon by means of the Fourier principle, that is, the superposition of a number of sinusoids (in time) or of spherical harmonics (in space), each with proper gain.

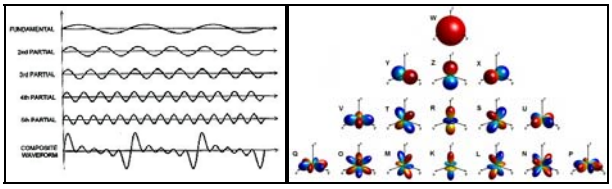

Figure 13: Fourier Analysis (left) and Spherical Harmonics (right).

### 2.1 SPS encoding

A formal definition of SPS requires to define the "sampling rule". At the time of writing, SPS has been attempted only up to 32 samples, located as in figures 8 and 9. The following table defines the "standard" order of the 32 virtual microphones, defining azimuth and elevation of each of them.

| Mic # | *Az*[°] | *El*[°] | Mic # | *Az*[°] | *El*[°] |
|---|---|---|---|---|---|
| 1 | 0 | 21 | 17 | 180 | 21 |
| 2 | 32 | 0 | 18 | 212 | 0 |
| 3 | 0 | -21 | 19 | 180 | -21 |
| 4 | 328 | 0 | 20 | 148 | 0 |
| 5 | 0 | 58 | 21 | 180 | 58 |
| 6 | 45 | 35 | 22 | 225 | 35 |
| 7 | 69 | 0 | 23 | 249 | 0 |
| 8 | 45 | -35 | 24 | 225 | -35 |
| 9 | 0 | -58 | 25 | 180 | -58 |
| 10 | 315 | -35 | 26 | 135 | -35 |
| 11 | 291 | 0 | 27 | 111 | 0 |
| 12 | 315 | 35 | 28 | 135 | 35 |
| 13 | 91 | 69 | 29 | 269 | 69 |
| 14 | 90 | 32 | 30 | 270 | 32 |
| 15 | 90 | -31 | 31 | 270 | -32 |
| 16 | 89 | -69 | 32 | 271 | -69 |

Table 2: angular coordinates of the 32 virtual microphones.

Knowing the angular coordinates $Az_m$ and $El_m$ of the 32 virtual microphones makes it easy to compute the encoding formulas, which are useful when a mono soundtrack must be encoded as a 32-channels P-format signal, appearing to come from a direction defined by the angles $Az_{in}$ and $El_{in}$.
For each virtual microphone $m$, the angle between the arriving sound and the microphone axis, $\varphi_m$, must first be found by means of the Haversine formula:

$$\varphi_m = 2 \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{El_m - El_{in}}{2}\right) + \cos(El_m)\cdot\cos(El_{in})\cdot\sin^2\left(\frac{Az_m - Az_{in}}{2}\right)}\right) \quad (10)$$

Then the gain for the encoded channel m is given by eq. 9.

### 2.2 SPS processing

Once the SPS signals have been obtained (either by recording or by synthesis), it is possible to manipulate them quite easily, performing standard operations such as rotation, stretching, zooming, etc.
Some of these transformations are easier in the SPS domain, whilst others, such as rotation, are more easy and accurate in the "spatial frequency" domain, operating on the spherical harmonics signals.
The most basic and simple transformation is obtained by changing the gains of the SPS components. This processing can be thought as "spatial equalization", boosting the signal from regions where the arriving sound is too weak, and reducing the signal form regions where it comes too loud. But this can also provide the opposite effect, that is, making a weak sound coming from a very precise direction to emerge above the general confusion.
A significant drawback of current implementation of the SPS technology is that the spatial sampling is not really uniform. As clearly shown in figures 8 and 9, the geometrical locations of the virtual microphones are not

perfectly regular, and hence "integer" rotations are not applicable easily. So in practice the only possible rotations are those corresponding to permutations of the faces of a dodecahedron, which is the basic geometry of our 32-virtual-microphones, as they are aimed at the vertexes and at the center faces of a dodecahedron.
It must also be noted that, whilst the set of spherical harmonics employed as the basis of the HOA approach form a perfectly orthogonal basis, the set of 32 4th-order cardioids currently employed are NOT a perfect orthogonal basis. Hence, a lossless analysis and resynthesis of the original sound field is theoretically possible with HOA, but not with SPS.
The fact that SPS is currently employing a set of spatial functions which are not a perfect orthogonal set can be disturbing for mathematically-oriented people.
But, coming back to the comparison of PCM sampling of a waveform, all we know that each "pulse" is not really independent from previous and subsequent ones, and some kind of "time smearing" always occur to the sound being PCM sampled and reproduced.
And also in HOA it is well known that the real performance of current microphone arrays when deriving high-order spherical harmonics is questionable, either in terms of signal-to-noise ratio and in terms of effective polar patterns being captured. Also in this case significant cross-talk always occurs, and the claimed mathematical independence and orthogonality of these signals remains just a dream…

### 2.3 SPS decoding

Finally, it is possible to employ the SPS signals for deriving the speaker feeds, to be employed in a playback system. The math for designing these "decoding" filters is substantially identical to the math employed for creating the virtual microphone filters, employed fore "encoding" the raw signals coming form the capsules into the SPS (P-format) signals.
The SPS signals can be reproduced employing a suitable loudspeaker rig. This approach shares with Ambisonics the capability of rendering the signals over a generic loudspeaker array, in principle composed of an arbitrary number of transducer, and in arbitrary positions, as the SPS signals being transferred are not "speaker feeds", such as in 5.1, 7.1, 10.2, 22.2, etc. Instead, the 32 signals of the SPS signal are a "spatial kernel", codifying the whole spatial information, exactly as the Ambisonics signals. With the difference that the SPS signals are "PCM encoded", whilst the Ambisonics signals are in the domain of "spatial frequency".
So let's assume that we have a suitable listening room, equipped with a reasonable number of loudspeakers, more-or-less uniformly covering the whole sphere, as shown in fig. 14.

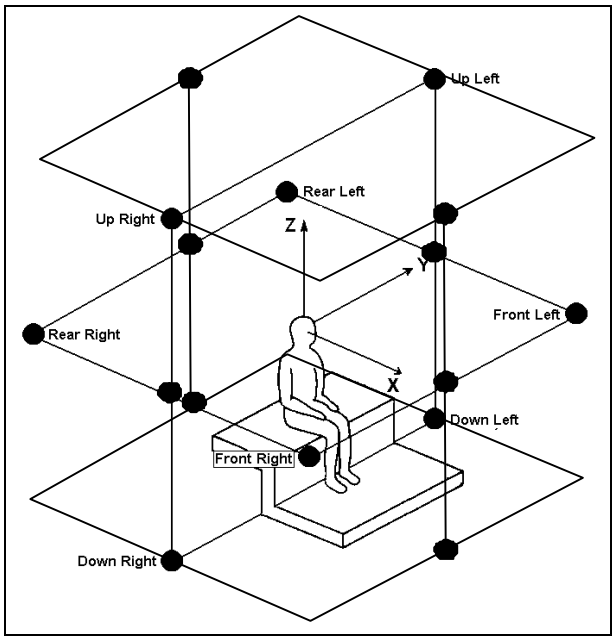

Figure 14: loudspeaker array with 16 loudspeakers and a listener at the centre.

In our approach, there is no requirement for the loudspeakers to be equidistant from the listener, so they can be conveniently placed along the walls and in the corners of the room.
For feeding our 16-loudspeakers array with our 32-channels SPS signals, we need to create a "decoding matrix" of 32x16 FIR filters, with substantially the same mathematical approach employed for deriving the "encoding matrix" of 32x32 FIR filters, already described in chapter 1.
In practice the 32 SPS signals *{y}* must be convolved with the matrix of filters *[f]*, yielding the required speaker feeds *{s}*:

$$s_r(t) = \sum_{i=1}^{32} y_i(t) * f_{i,r}(t) \quad \Rightarrow \quad \{s\} = \{y\} * [f] \tag{11}$$

For determining the filters *[f]*, we start form a set of measurements of the loudspeaker's impulse responses, performed placing our 32-capsules microphone array at the centre of the listening room (in the "sweep spot position", where the head of the listener should be). Let's call *[k]* the matrix of these measured impulse responses.
The conditions to be imposed for finding the values of *[f]* are that the signals captured by the microphone array, if placed in the centre of the listening room, are identical to the "original" SPS signals {y}:

$$\{y_{out}\} = \{s\} * [k] = \{y\} * [f] * [k] \tag{12}$$

Of course, the recovered signals $\{y_{out}\}$ will never be really identical to the original ones {y}, some error will always occur, as shown in fig. 15.

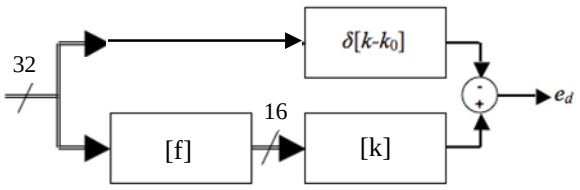


Figure 15: block diagram of the playback system.

As we did for computing the encoding filters *[h]*, we now set up a least-squares approach for finding the matrix of decoding filters *[f]*, operating in frequency domain and employing a frequency-dependent regularization parameter $\beta$, and setting up a "modelization delay" $\delta$ of *N/2* samples:

$$[F]_{16x32} = \frac{[K]^*_{16x32} \cdot e^{-j\pi k}}{[K]^*_{16x32} \times [K]_{32x16} + \beta_f \cdot [I]_{32x32}} \qquad (13)$$

Again, the frequency dependence of $\beta$ is as shown in fig. 4, with frequency limits generally narrower than those used for encoding (typically loudspeakers have a more limited usable frequency range than microphones). The creation of a pseudo-inverse of the reproduction matrix *[k]* is much more difficult than the inversion of the microphone matrix *[c]*: the inversion is optimal only if the loudspeakers are all identical, placed on a perfect sphere, as shown in fig. 16. This is the playback system employed by Nelson and Fazi [7] at ISVR, in Southampton, UK.

Figure 16: ISVR's spherical playback system.

In our case, we employ a much worst playback system, as shown in fig. 17 (panoramic image): as the room is not really anechoic, and the loudspeakers are not located at the same distance form the centre, the matrix becomes more tricky to invert, and the resulting filters need to be much longer, typically 4096 or even 8192 samples.

Figure 17: Listening room of Casa della Musica, University of Parma, ITALY.

Due to the acoustical and geometrical deficiencies of such a listening room, the matrix of inverse filters has to do a difficult task. A "brute force" approach for automatic computation of the decoding matrix revealed to be problematic, and some constraining and simplification had to be hand-coded.
The room was originally designed for Ambisonics 2D and 3D playback, and for this task the regular location of loudspeakers makes the system to work reasonably well.
Se we ended up performing a side-by-side comparison between SPS and HOA.

### 2.4 "brute force" SPS decoding

In this case, the theory exposed at chapter 2.3, and in particular eq. 13, was employed processing the matrix of impulse responses measured inside the listening room shown in fig. 17.
The resulting filter matrix is shown in figure 18.

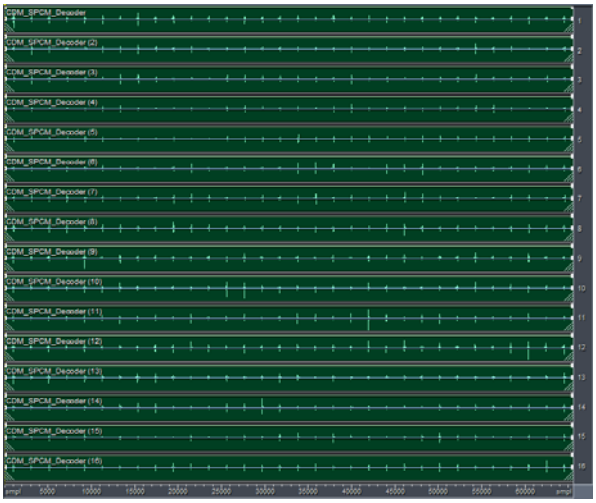

Figure 18: matrix of "brute force" SPS decoding filters.

### 2.5 "Manually tuned" SPS decoding

Observing the filters in fig. 18 it is evident how each loudspeaker is being fed with significant contribution from ALL the SPS components, and this is definitely wrong. For optimal decoding, the loudspeaker rig should be employed for creating a set of 32 "virtual loudspeakers", one for each "virtual microphone" of the SPS signal, and then the feeding should be 1-to-1.
Typically any virtual loudspeaker can be created by feeding at most the three surrounding real loudspeakers, and employing suitable "vector panning" algorithms, such as VBAP, for ensuring that the sound appear to come from the position of the virtual loudspeaker.
The following figure shows the superposition of the positions of the 16 "real" loudspeakers" and the 32 "virtual" loudspeakers to be created.

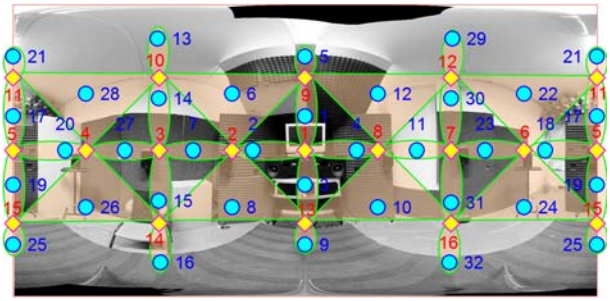

Figure 19: real and virtual loudspeakers
in the listening room.

The circles are the virtual loudspeakers, the romboids are the real ones, and the shaded areas indicate the fact that the sound of each virtual loudspeaker is being created by feeding just a small number of real loudspeakers (1, 2 or 3).

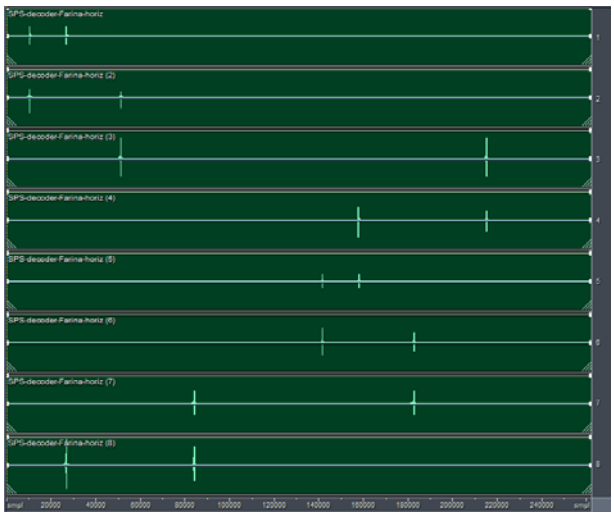

Figure 20: matrix of "manually tuned"
SPS decoding filters (for the first 8 loudspeakers).

It can be seen that 16 virtual loudspeakers can be created by just a pair of real loudspeakers, 8 require a triplet of real loudspeakers, and 8 (the ones close to the North and South poles) can be reasonably emulated by feeding just one real loudspeaker.
So, forcing that each SPS component only feeds, through a suitable FIR filter, the "right" real loudspeakers, and imposing the remaining cells of the matrix to be zero, a new decoding matrix was computed, as shown in figure 20.

## 3 SPS VS HOA

We compared side-by-side the usage of the HOA method and of SPS, starting with the same signals captured by a spherical microphone array, and playing back the recording inside our listening room equipped with 16 loudspeakers.
For HOA, the 3rd order Ambisonics decoder developed by Menzel Digenis as VST plugin was employed [8].
A third playback method, known as 3DVMS, was also employed for comparison: this is obtained by simply computing a set of filters synthesizing a 3rd order cardioid for each loudspeaker, aiming the cardioid at the same azimuth and elevation of each loudspeaker. In this case there is not an encoding stage, an intermediate format (HOA or SPS) and finally a decoding stage. Instead, the raw Eigenmike signals are directly filtered and sent to the loudspeakers.
The evaluation of these different methods was based on two procedures:

a) The polar patterns of the virtual microphones resulting by the combination of the encoding and decoding processes were experimentally evaluated (by processing the Eigenmike recording performed on the turntable inside the anechoic room).
b) A formal blind listening test was performed, with 5 subjects and two sound recordings of human voices performed in different environments, and evaluating perceptual qualities such as localizability of the talkers, timbric neutrality, absence of artefacts and response to transients.

### 3.1 Resulting virtual microphones

The following figures show the polar patterns of the virtual microphones feeding the 8 loudspeakers located at 0° elevation (horizontal plane). The loudspeakers are located at the vertexes of a regular octagon, so the theoretically optimal virtual microphones should look as ultradirective cardioids point at 0°, 45°, 90°, etc. – of course without any side or rear lobes.

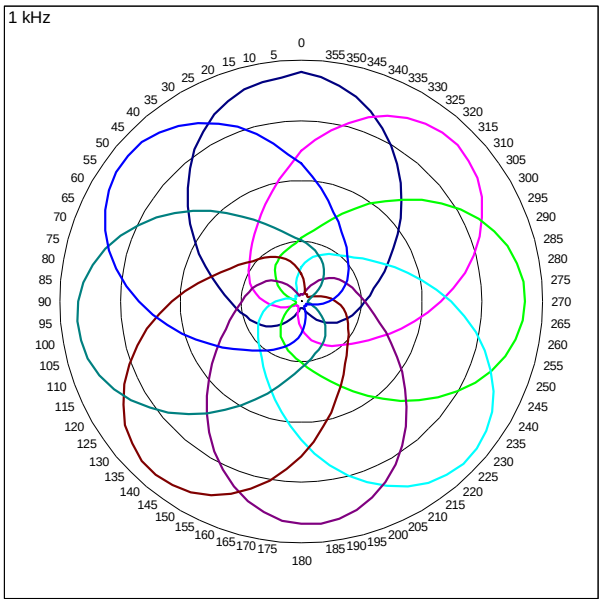


Figure 21: 3rd order HOA encoding/decoding.

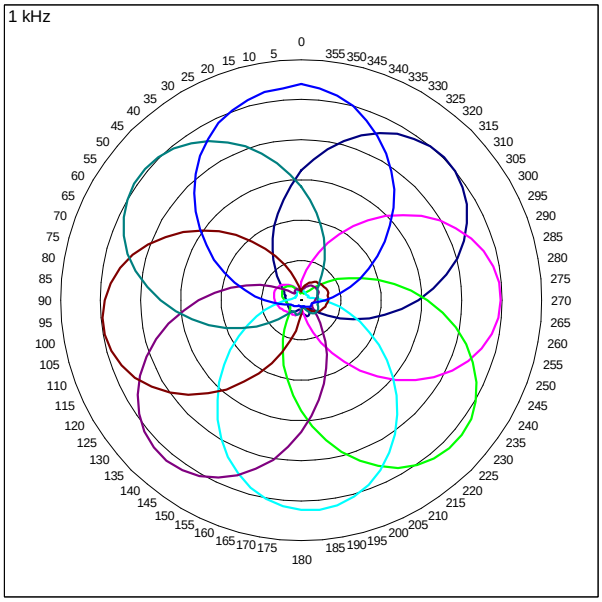


Figure 22: 3DVMS direct feeding.

Comparing HOA with 3DVMS; it can be seen that the Digenis' 3rd-order decoder employed here (with its standard settings) produces quite broad cardioids, but there are substantially no side or rear lobes.
On the other side, the 3DVMS direct synthesis of 3rd-order cardioids produces nice polar patterns in the frontal area, but significant rear lobes.

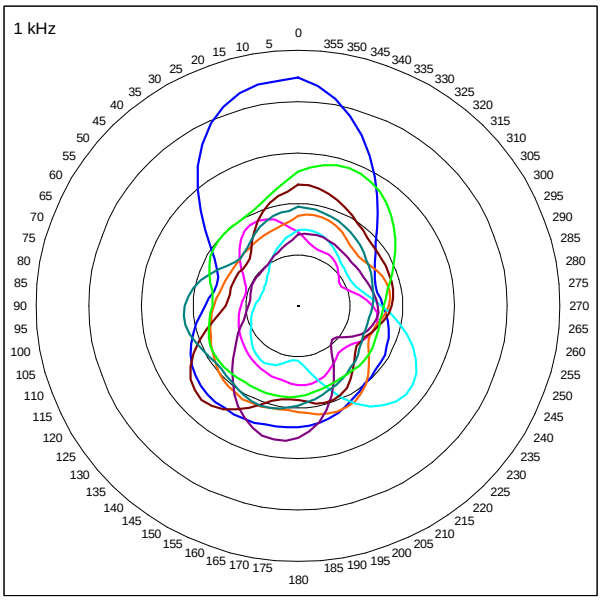


Figure 23: SPS encoding/decoding – "brute force" decoding filters.

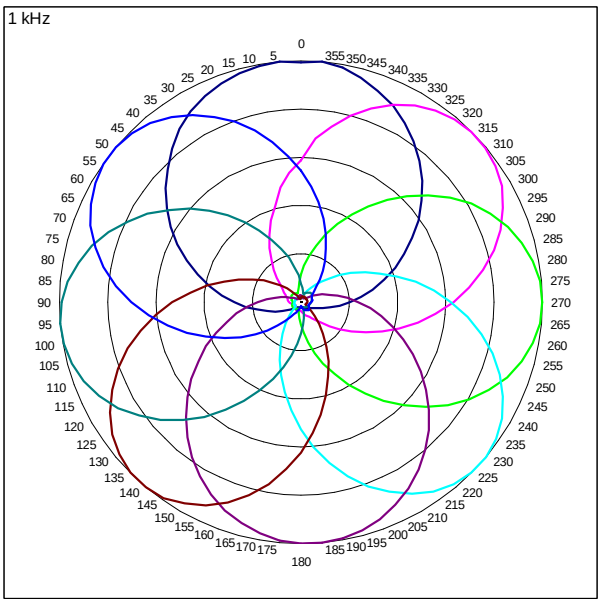


Figure 24: SPS encoding/decoding – manually-tuned decoding filters.

From the last two figures, it is evident that the "brute force" approach failed, with evident analogy with the failure of the 2nd-order Ambisonics decoder described in chapter 1.1.
A closer comparative analysis of these two cases reveals a common fact: in both cases the number of constraining equations is equal to the umber of unknowns. So the inversion problem is well defined, but not over-constrained.

In practice, we have seen that these inversion problems are better solved when severely over-constrained: for example, the virtual microphones are working so well because they are derived by a measurement set containing over 600 different directions, for synthesizing just 32 filters…
One possible solution for getting reasonable results with the "brute force" approach, indeed, could be to perform a larger number of measurements inside the listening room, for example rotating the Eigenmike in small angular steps, providing a better spatial resolution along both azimuth and elevation.

### 3.2 Resulting virtual microphones

The listening tests were actually performed BEFORE the polar patterns shown in the previous chapter had been measured. However, the results are perfectly consistent.
The best evaluation was given to the direct 3DVMS method, in which a single stage of filtering is employed, feeding each loudspeaker with a signal coming from a virtual microphone with well-controlled directivity and proper aiming.
The SPS decoding based on manually-tuned filters did perform juts after, with small degradation of just one perceptual parameter, that is the temporal response to transients.
High Order Ambisonics was judged worst, mostly for the "colour" of the sound, due to the fact that in the HOA processing there is nothing taking care of the actual response of the loudspeakers, but also the spatial separation of simultaneous talkers was not so effective due to the larger cross-talk between individual speaker feeds.
Finally, the general judgement for Ambisonics was to be "soft" in any sense: spectrally, with attenuated low and high ends. Spatially soft, with "enlarged" sources and spatially-smeared localization. And temporally soft, with smudged attacks of transients.
Of course, these evaluations are probably related to some drawbacks of the particular 3rd-order decoder employed. The author of this software is not distributing it anymore, and it is known that much better HOA decoders do exist, for example the Ambdec decoder by Fons Adriaensen [9], which is planned to be inserted in a future, more extended comparative listening test.
Finally, the "brute force" filters for SPS decoding were really awful: the sound was coming from everywhere, and the interaction between loudspeakers made the sound field to be terribly unstable for small movements of the listener.

## 4 CONCLUSIONS

This paper has described the first attempt to create SPS signals (spatial PCM sampling), to manipulate them and to render them over a three-dimensional loudspeaker system.
SPS can be thought an alternative approach to High Order Ambisonics. It shares the same concept of encoding the spatial information in a small number of channels, each representing some "spatially-dependent" filter. The encoded signals can be processed, and later played back over a loudspeaker system with arbitrary geometry and number of loudspeakers.
So, both systems enable to transfer the spatial audio information in a format which is independent both on the geometry of the microphone array which captured the sound and of the loudspeaker array which will play it back.
Despite the fact that the first attempt of employing the SPS concept had to be constrained by some significant limitations in both the capture and rendering systems, a side-by-side comparison with HOA revealed some strong advantages for SPS: better spatial resolution, more "clean" and "unprocessed" sound.
Of course the method should be perfected: a different set of encoding functions can be employed, a more uniform coverage of the spherical surface can be achieved, and better hardware can be built and employed at both sides of the recording/playback chain.
And we missed the simplicity of performing rotations in the HOA domain, so we definitely need to develop a "fractional rotation" module for SPS, the spatial equivalent of a "fractional delay" for a time-domain PCM signal.
The comparison with Ambisonics was probably a bit biased by the fact that the Ambisonics decoder employed is definitely suboptimal, and employing a better Ambisonics decoder it is certainly possible to obtain signals corresponding to more directive virtual microphones.
In principle, increasing the Ambisonics order properly (for example to 4th or even 5th order, in the case of the Eigenmike), it is possible to use the Ambisonics technology for obtaining exactly the same virtual microphones as we did obtain with the SPS technology. The problem is that there are currently no 4th-order or 5th-order Ambisonics decoders available.
As both HOA and SPS employ perfectly linear filtering techniques, in principle both approaches can be employed for getting exactly the same signals. So the choice between the two approaches has to be made weighting the operational advantages and disadvantages of both, and this paper demonstrated that the SPS technique is already viable, requiring reasonable computational performance and providing very good results.

**REFERENCES**


[1] 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

[2] http://www.muse.demon.co.uk/ref/speakers.html

[3] Angelo Farina, Andrea Capra, Lorenzo Chiesi, Leonardo Scopece - “A Spherical Microphone Array For Synthesizing Virtual Directive Microphones In Live Broadcasting And In Post Production” - 40th AES Conference "Spatial Audio - Sense the Sound of Space", Tokyo, Japan, 8-10 October 2010

[4] Kirkeby, O., Nelson, P.A., Hamada, H., Orduna-Bustamante, F., “Fast deconvolution of multichannel systems using regularization”, IEEE Transactions on Speech and Audio, 6, (1998).

[5] O.Kirkeby, P.A. Nelson, P. Rubak, A. Farina , Design of Cross-talk Cancellation Networks by using Fast Deconvolution, 106th AES Convention, Munich, 8-11 may 1999.

[6] http://pcfarina.eng.unipr.it/Public/Xvolver/

[7] Poletti, M., Fazi, F.M. and Nelson, P.A. “Sound-field reproduction systems using fixed-directivity loudspeakers”, Journal of the Acoustical Society of America, 127, (6), 3590-3601, (2010).

[8] Menzel Digenis, Ambisonics Decoder VST Plugin (http://www.kvraudio.com/product/ambisonic_3rd_order_decoder_by_digenis)

[9] Fons Adriaensen, Ambdec decoder for Linux (http://kokkinizita.linuxaudio.org/linuxaudio/adec-pict.html)