SPATIAL SOUND RECORDING WITH DENSE MICROPHONE ARRAYS

ANGELO FARINA[1], SIMONE CAMPANINI[1], LORENZO CHIESI[1], ALBERTO AMENDOLA[1], LORENZO EBRI[1]

[1] *Industrial Engineering Department, University of Parma, Italy*
farina@unipr.it

Multichannel recordings are usually performed by means of microphone arrays. In many cases "sparse" and discrete microphone arrays are used, where each microphone is employed for capturing one of the channels, which in turn is routed to one loudspeaker.
However, also the usage of "dense" microphone arrays has a long history, dating back to the first MS-matrixed microphones setups and passing through the whole Ambisonics saga.
A dense microphone array is employed differently from a sparse array: each channel is obtained by a combination of the signals coming from all the capsules, by means of different matrixing and filtering approaches. And similarly, each loudspeaker feed results from a re-matrixing of all the transmitted channels.
This paper is the third of a series: in the previous two [1,2] a numerical method for computing a matrix of FIR filter was employed for processing the microphone signals (encoding, [1]) and for computing the speaker feeds (decoding, [2]). In this third paper, the same numerical approach is extended to intermediate processing (rotation, zooming, stretching, spatial equalization, etc.): hence we have now a general meta-theory, providing a unique framework capable of processing the signals for any kind of dense microphone array, providing any kind of intermediate manipulation, and finally projecting the signal to every kind of loudspeaker arrays. The same framework can operate according to different standards and formats, including A-format (raw signals), B-format (High Order Ambisonics signals), G-format (speaker feeds) and P-format (Spatial PCM Sampling signals), and can be used for converting freely among them.
Experimental results are presented, including "traditional" tetrahedral probes, a commercial spherical microphone array, and two newly-developed massive microphone arrays developed by the authors, a cylindrical and a planar array, both incorporating 32 high-quality condenser microphones and a panoramic video camera.

INTRODUCTION

The use of microphone arrays is the basis for making audio recordings (or acoustical measurements) capable of capturing information about the spatial distribution of sound impinging onto a listener.
In a **sparse** microphone array the microphones are placed at large distances, with the goal to sample the sound field at points where the sound is significantly different.
On the other hand, in a **dense** array, the microphones are usually small and close each other, so that the minimum distance between two microphones is significantly smaller than the wavelength for any allowed direction of arrival of the wavefront.
The intended usage of the signals captured by the two types of microphone arrays is completely different: in a sparse array, the signal coming from each capsule has to be kept well separated from all the others, and is tailored to become the feed for a single loudspeaker in the playback system. The only adjustments possible for the sound engineer is to apply to each signal some amount of gain, some spectral equalization, and perhaps some delay. But each signal is always processed independently from the others. The only exception to this rule is when a multichannel signal is downmixed to stereo or mono.
In a dense microphone array, instead, the whole set of signals are treated as an unique entity, and the processing required at each stage (encoding, manipulation, decoding) involves generally complex matrixing operations, so that each output channel includes some amount of information captured by all the microphones. Furthermore, heavy filtering is usually required, instead of the simple gain and delay adjustments employed for sparse arrays.
This means that the dense array technology requires more effort. All the microphones must be reliable and high quality (a noisy capsule affects the behaviour of the whole system), the processing algorithm requires more computational power (as each output signal requires to apply specific filtering to all the input channels, and each filter is generally computationally very heavy), and the risk to get artefacts is vastly larger.
These factors explain the comparatively larger success encountered, until now, by sparse microphone arrays in comparison with dense microphone arrays.

GOALS

In this paper a single, unified meta-theory is proposed for describing the processing performed in any kind of dense microphone array, at any stage of processing (encoding, manipulation, decoding), and operating within any currently employed or foreseeable standard (WFS, HOA, SPS, etc.). The approach is very accurate whenever the processing required is perfectly linear and time invariant, as it is the case for many classical approaches.

Whenever a not linear, time variant approach is employed, such as the "steering" methods based on analysis of the acoustical three-dimensional scene and consequent continuous change of the filters (as in the Dirac and Harpex-B algorithms [3,4]), the approach presented in this paper is still useful for representing each temporary "state" of the time-variant processing system.

1 THE VIRTUAL MICROPHONE CONCEPT

The approach proposed in this paper for unifying the treatment of any existing and future dense microphone array is the concept of "virtual microphone".

Whatever signal is present at the beginning, in the middle, or at the end of a sound processing system, such a signal must always be interpreted as the signal coming from a microphone, which was placed in a particular position with a particular aiming and with a specific polar pattern in a specific room.

So we can always think to "virtual microphones" when we listen to:

- the signals coming from the real capsules (in this case, the virtual microphone coincides with a real one),
- the intermediate signals created when "encoding" the signals onto the delivery medium (for example, the B-format WXYZ signals employed by 1st-order Ambisonics microphones),
- the modified signals resulting from any kind of intermediate processing and manipulation,
- the output signals being created when "decoding" the signals for feeding loudspeakers in a playback system.

Every signal is always a virtual microphone signal !

After having defined this general "virtual microphone" concept, let's manage the whole multichannel stream as a linear vector of signals, so that we can use vectorial calculus for performing math operations efficiently and employing a compact notation. At every stage we have just a vector of virtual microphone signals, and passing form each stage to the next just means to apply a vector processing, which we assume to be linear and time-invariant (a matrix of FIR filters).

1.1 Multichannel digital signal processing

Given a vector of input virtual microphone signals, a set of digital filters is employed for creating the vector of output virtual microphone signals.

Let's consider, for example, the case of encoding V intermediate signals (B-format) from the M raw signals (A-format) coming from the capsules.

So we need a bank (a matrix) of $M \times V$ filters. For maximum generality, stability and better computational performances, we prefer to employ long FIR filters.

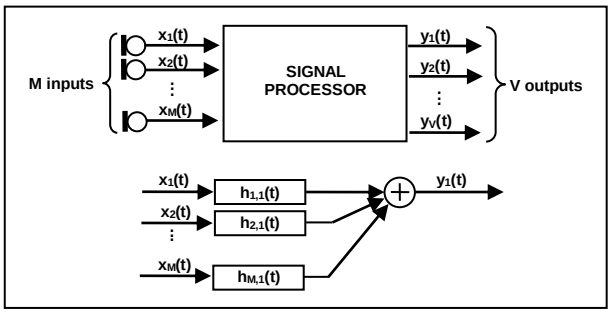

Figure 1: Scheme of the signal processing

Assuming $\boldsymbol{x_m}$ as the input signals of M microphones, $\boldsymbol{y_v}$ as the output signals of V virtual microphones and $\boldsymbol{h_{m,v}}$ the matrix of filters, as shown in fig. 1, the processed signals can be expressed as:

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

Where * denotes convolution, and hence each virtual microphone signal y_v is obtained summing the results of the convolutions of the M inputs x_m with a set of M proper FIR filters $h_{m,v}$.

By changing the filtering coefficients h, it is possible to synthesize virtual microphones having arbitrary location, aiming and directivity pattern.

The processing filters h can be computed theoretically, based on the solution of the wave equation, assuming that the microphones are ideal and identical.

But we advocate instead the use of numerically-computed filters, based on real-world measurements.

Equation (1) does not describe just the encoding process, it can also describe any kind of "intermediate manipulation" performed on the signals for modifying the spatial information (stretching, rotating, zooming, etc.), and also the final "decoding" process, which is done for generating the feeds for the loudspeakers.

At each processing stage, the number of channels can change: whenever the number is not reduced, the whole spatial information is preserved, and with proper techniques, it is possible, in theory, to retrieve the original signals (as the processing is perfectly linear, and hence reversible). However, whenever the channel count is reduced, the transformation is inherently lossy, and hence some spatial information is lost.

1.2 Computing the FIR filters' coefficients

Instead of dealing with (often unnecessarily) complex mathematical theories for computing the filtering coefficients *h*, the authors proposed a novel approach, not requiring any theory [1]: the set of filters *h* is derived directly from a set of measurements.
The original formulation was for the encoding stage, and is recalled very shortly here just for making it available for the reader.
The characterization of the array is based on a matrix of measured anechoic impulse responses ***c***, obtained with the sound source placed at a large number *D* of positions all around the probe, as shown in Figure 2.

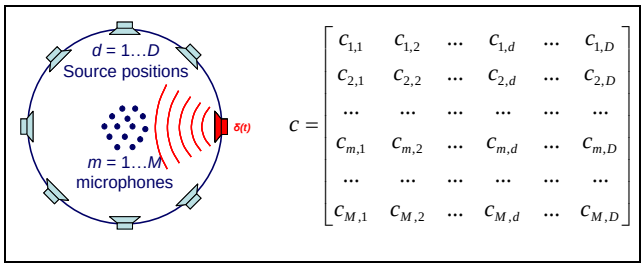

Figure 2: impulse response measurements from D source positions to the M microphones

The matrix ***c*** has to be numerically inverted, imposing that the unknown filters, applied to the measured data, for any direction *d*, match as closely as possible the response of a prescribed virtual microphone $\boldsymbol{p}_d$. This method also inherently corrects for transducer deviations and acoustical artefacts (shielding, diffraction, reflection, etc.).
Going to frequency domain this becomes:

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

Now we can search for the unknown coefficients ***H***.
A direct inversion of the system is unfeasible: an approximate inversion with regularization and delay is required, to ensure causality and to avoid excessive emphasis at frequencies where the signal is very low.
We employ the numerical solution originally developed by Kirkeby and Nelson [5]:

$$H_{m,k} = \left[C^{T}_{m.d.k} \cdot C_{m.d,k} + \beta_k \cdot I_{m,m}\right]^{-1} \cdot \left[C^{T}_{m.d.k} \cdot P_{d,k} \cdot e^{-j\pi k}\right] \quad (3)$$

Where $[\ ]^T$ means conjugate transpose, $[\ ]^{-1}$ means pseudo-inversion, and · means dot product.
According with the modification proposed in [6], the regularization parameter β is made dependent on frequency index *k*, so that the inversion is accurate in the central frequency range and more "robust" at extreme frequencies, as shown in fig. 3.

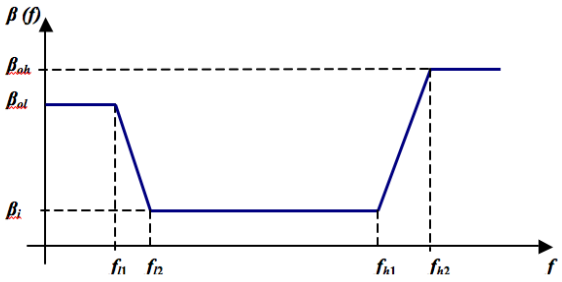

Figure 3: frequency dependence of the regularization parameter β_k

The same approach can be used also for decoding, as shown in [2]: in this case, each filtered signal becomes the feed for a loudspeaker in the playback system, and the matrix C of measured impulse responses is measured in the playback room over the loudspeakers.
The novelty presented here is the usage of the same processing scheme (eq. 1) with a set of filtering coefficients computed numerically according to eq. 3 also for performing any possible manipulation of the signals at an intermediate stage. Hence, we can design a matrix of FIR filters for performing rotation, or for expanding a certain angular region and compressing the remaining. Also "spatial equalization" can be performed, reducing the gain from directions where disturbing sources are active, and boosting the gain from direction where the useful signals are arriving.
In the end, we have a sequence of subsequent processing stages: encoding, several manipulations, decoding: each stage is always implemented by means of matrix convolution of the signals with a matrix of FIR filters (eq. 1), and their coefficients are not derived theoretically, they are instead obtained by the numerical inversion process described by eq. 3.

2 THE MICROPHONE ARRAYS

The experiments described in this paper were performed using four dense microphone arrays (figs. 4, 5, 6):

- A DPA-4 tetrahedral 1st-order Ambisonics microphone probe
- An Eigenmike™ microphone array produced by MH acoustics [7].
- A Cylindrical microphone array equipped with 32 capsules, which was built for this research
- A Planar microphone array equipped with 32 capsules

The DPA-4 is a classical tetrahedral assembly of 4 high-quality, medium-diaphragm microphones. The usage of four 16mm DPA-4011 capsules results in outstanding sound quality, with very low noise and perfectly flat frequency response. Their size, indeed, causes the diaphragms to lye over the surface of a theoretical sphere having a radius of 18mm, which is a bit too large for a 1st-order Ambisonics probe, limiting the upper frequency to approximately 4 kHz for keeping control

of the polar patterns of the virtual microphones.
The Eigenmike™ is a sphere of aluminium (the radius is 42 mm) with 32 high quality capsules placed on its surface; microphones, pre-amplifiers and A/D converters are packed inside the sphere and all the signals are delivered to the audio interface through a digital CAT-6 cable, employing the A-net protocol.
Fig. 5 shows an exploded view of the new cylindrical microphone array: this prototype was built employing a disassembled Eigenmike™, keeping the same capsules and electronic circuits, which were mounted in a solid aluminium-methacrylate body, featuring a diameter of 110mm and a length of 352mm.
The capsules are mounted at uniformly-spaced azimuth angles ($\Delta\theta = 11.25°$) and stacked along the Z-axis at 6 levels, along a capsule holder section having a length of 100mm.
The optimal geometry of the capsule holder was obtained by running a number of numerical simulations, as explained in the following chapter 2.1. The chosen geometry optimizes the performances of the resulting virtual microphones over a frequency range of 100 Hz to 12 kHz.
This cylindrical microphone array also includes an advanced optical system, as discussed in chapter 2.2.
Finally, fig. 6 shows the new Planar microphone array: again, this was built reassembling microphones and electronics of an Eigenmike™, and fitting everything inside a ruggedized flat box, equipped with a surface-mounted wide-angle 3 M-pixels IP camera.
Also in this case, the positions for the 32 capsules have been optimized by means of numerical simulations, as described in chapter 2.1.

Figure 4: the three "panoramic" microphone arrays: DPA-4, Eigenmike™, new Cylindrical array

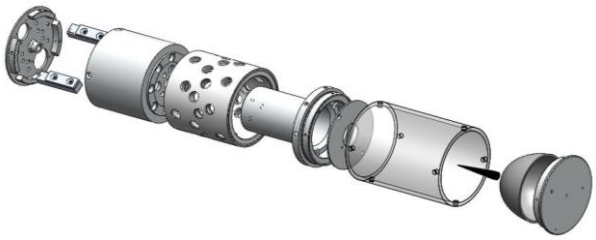

Figure 5: construction of the Cylindrical microphone

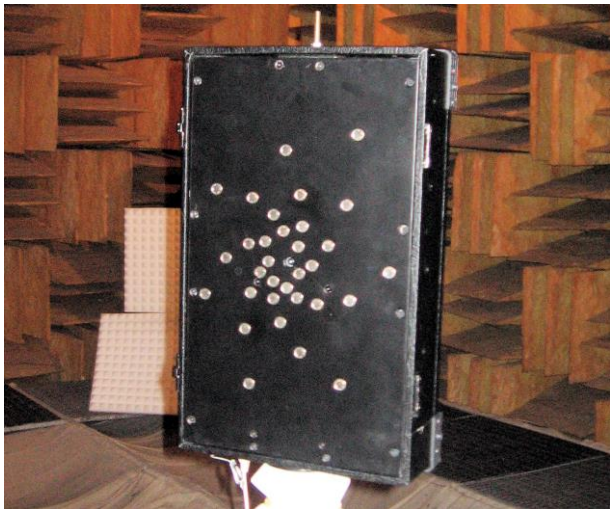

Figure 6: the new Planar array

2.1 Geometry optimization by numerical modelling

The "theory-less" approach described in chapter 1.1 is suitable not only for experimental measurements, but also for numerical simulations. This means that it was possible to create sets of simulated impulse responses for the new microphone arrays (Cylindrical, Planar) before they were actually built.
Repeating the simulation with slightly different geometries, it was possible to hand-pick the best one, which provides virtual microphones with wide and flat frequency response and smooth polar patterns, immune from artefacts up to very high frequency.
The simulations were performed with Comsol 4.2, imposing a plane wave incoming from the left face of an "air cube" surrounding the microphone array. Profiting of the axial symmetry of the cylinder, a single simulation allows for the analysis of all azimuth angles for the given elevation. Hence the construction of the complete directivity balloons is fast.
After repeating the simulation for dozens of proposed geometries, the one shown in fig. 5 was chosen, providing the more reasonable behavior in the useful frequency range.
The same comparison was performed for choosing the microphone positions for the Planar array, and the resulting optimized geometry is shown in fig. 6.

2.2 Auxiliary optical system

The two new microphone arrays (Cylindrical, Planar) were built around high quality optical systems.

The one installed at the center of the Planar array is very simple: a surface-mounted Sony Ipela CH-201 IP camera was employed, providing a video stream with 3-Mpixel resolution (2048x1536) at 15 Fps. The wide-angle lens cover a rectangular field of view measuring approximately 80°x60°, which corresponds to the solid angle where this microphone array works optimally. Hence the images or videos captured through this camera are perfect as "background" for the acoustical scene being recorded.

The optical system embedded inside the new Cylindrical array, instead, is much more complex and requires some explanation. A 5-Mpixel Wision WS-M8P31-38B camera module was employed, equipped with a short-range 6mm C-mount lens and a hyperbolic mirror manufactured by the company 0-360.com. The latter was equipped with a black needle for removing unwanted reflections inside the methacrylate tube.

Fig. 7 shows the optical scheme and the mechanical structure of the panoramic vision system.

The image captured by the lens is the typical "donut" image shown in fig. 8.

For getting a usable picture, the video stream coming from the camera needs to be intercepted and "unwrapped", before being recorded or used as background. This kind of video processing is computationally heavy, and possibly risky, when operated on the same portable computer already dealing with massive audio sampling and realtime audio processing.

For minimizing the computational load, the video-unwrapping algorithm was coded at low level in C++ inside a modified G-streamer "filter".

This made it possible to feed the high-level software with a rectangular panoramic video stream, having a 2:1 aspect ratio, and covering an angle of 360°x180°, at a reduced resolution of 1280x640 or 960x480 (depending on the available screen resolution), as shown in fig. 9.

Also in this case the image being captured corresponds perfectly to the usable acoustical panorama of the Cylindrical array, which is optimized for the synthesis of virtual microphones pointing at every azimuth, but with elevation limited in the +/- 45° range.

The black stripes at bottom and top, indeed, allow for aiming virtual microphones in the whole elevation range of +/- 90°, when this unwrapped image or video is employed as background for the Eigenmike™.

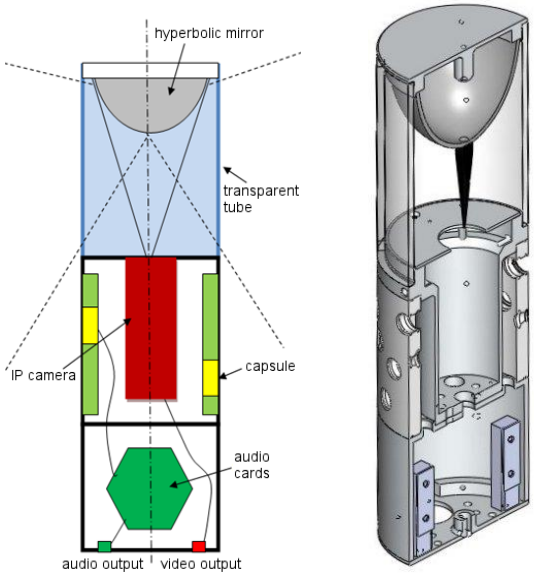

Figure 7: scheme of panoramic vision system and cross-section of the cylindrical body

Figure 8: "donut" wrapped image

Figure 9: unwrapped image

2.3 Experimental characterization of the arrays

After the four microphone arrays were purchased (DPA-4, Eigenmike™) or built (Cylindrical, Planar), they were subject to extensive measurements inside a large anechoic room (kindly provided by ASK Industries).

A two-axes turntable was designed and built, providing minimal acoustical interference and very good precision, thanks to precision step motors and high-quality mechanical components, as shown in fig. 10.

The turntable contains the required control electronics board, and is powered and controlled through a single CAT5 cable, carrying Power Over Ethernet. A set of Matlab routines were developed, allowing for complete automatization of the whole measurements procedure. This made it possible to cover almost uniformly the spherical surface, with a pattern of measurement points which is both well distributed and fast to sample, following a sensible measurement path, as shown in fig. 11 (this minimizes rotation times, and simultaneously avoid that the microphone cable makes too much contortions).

A "point source" loudspeaker was employed (Tannoy dual-concentric studio monitor), for ensuring that each measurement samples just one very precise direction-of-arrival of the wavefront. Thanks to the size of the anechoic room, it was possible to keep a distance of 5.0m between the acoustic centers of the sound source and of the microphone array, as shown in fig. 12.

The Exponential Sine Sweep (ESS) method was employed, in order to obtain M Impulse Responses for each direction of arrival of the sound.

The number of directions, D, was equal to 122 for the DPA-4 probe, and to 362 for the other three, denser microphone arrays.

The ESS method was chosen due to its capability of removing unwanted artefacts due to nonlinearities in the loudspeaker, and because it provides significantly better S/N than other methods based on periodic signals, such as MLS or linear sine sweep (TDS), as one of the author already discovered [8].

The raw results of the measurement are the impulse responses $\boldsymbol{c}$ of each capsule of the array (m=1..M) to the sound arriving by every direction (d=1..D).

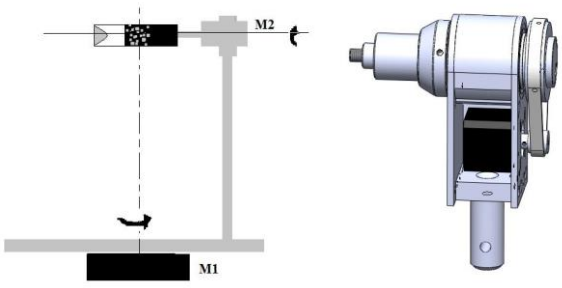

Figure 10: the two-axes turntable

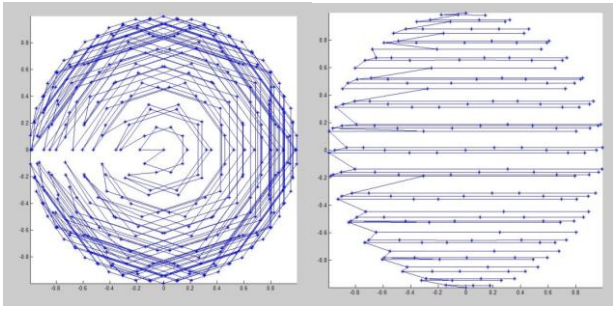

Figure 11: the measurement pattern

Figure 12: the equipment in the anechoic room

2.4 Synthesis and test of virtual microphones

In order to derive the matrix of filters, a Matlab script was produced. This script employs 2048 samples of each impulse response and it needs as inputs the number of virtual microphones to synthesize, their location (usually assumed at the nominal center of the array, but it can also be specified a small offset, if wanted) their directivity, their azimuth and elevation. From these inputs, according with the theory and the procedure described in paragraph 1.1, it is possible to get the set of processing filters matrix h.

The convolution of the M signals coming from the capsules of the array with these FIR filters should output the signals of V virtual microphones with the desired characteristics.

In figs. 13-16 it can be seen how, thanks to the "tricks" played by the experimentally-optimized FIR filters, it is possible to synthesize virtual microphones having directivity patterns significantly sharper than the maximum theoretically possible for a given geometry (for example, 1st order for the DPA-4, 4th order for the Eigenmike™ spherical array, etc.).

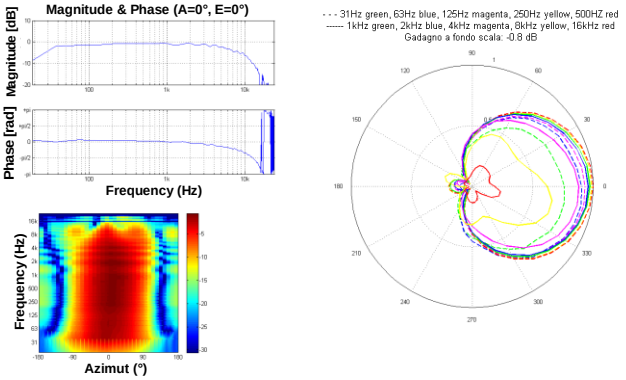

Figure 13: 2nd-order cardioid (DPA-4)

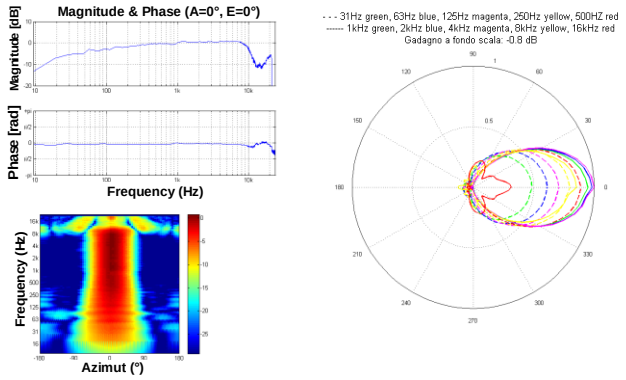

Figure 14: 6th order cardioid (Eigenmike™)

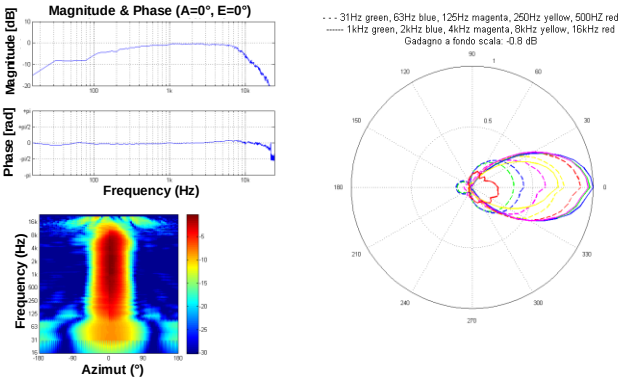

Figure 15: 8th order cardioid (Cylindrical)

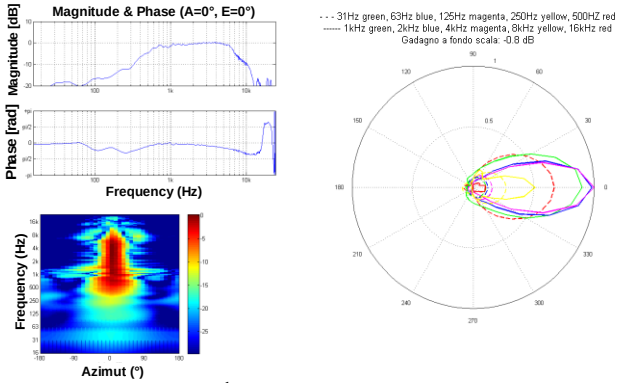

Figure 16: 16th order cardioid (planar array)

2.5 Manipulation

At this stage of the research, we did attempt just very basic manipulation tasks, such as rotation around the vertical axis, conversion between different formats and removal of noise by reducing the gain at unwanted directions. In all cases, the known matrix $\boldsymbol{c}$ is always based on the original set of measurements performed over each microphone array, as described in the previous subchapter, but after applying to them the set of FIR filters corresponding to the processing already done (encoding all the preceding manipulation stages).

In case of rotation, for example, we derived a set of 360 filtering matrixes for the Spatial PCM Sampled (SPS) signals: each of them performs rotation around the Z axis at 1° increments. Switching continuously the filtering matrix creates the illusion of a sound field rotating continuously around the listener.

This solved the main drawback encountered for the SPS approach, that is the difficulty of performing rotations by an arbitrarily small angle, as reported in the conclusions of [2].

These rotation matrixes perform equivalently as fractional delays for a sampled PCM waveform.

On the other hand, spatial equalization resulted to be very easy in the SPS domain, as the unwanted noise signals are reduced simply lowering the gain of the corresponding directional components.

2.6 Decoding

Regarding decoding, the known matrix $\boldsymbol{c}$ needs now to include the measured transfer functions between each loudspeaker and the microphone array, as described in [2].

Eq. 3 yields, in this case, a matrix of FIR filters which creates the speaker feeds. The measurements where performed inside a listening room equipped with 16 loudspeakers, as shown in fig.17.

The same sound system can also be operated with more traditional decoding methods, such as HOA and VBAP, allowing for comparisons with the new matrix-based approach. Also in this case the numerical approach provided advantages, as the loudspeaker locations did not correspond exactly with a regular polyhedron.

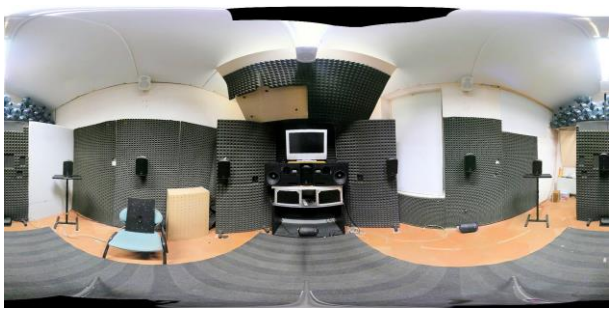

Figure 17: Listening room of Casa della Musica, University of Parma, ITALY.

CONCLUSIONS

A single mathematical framework (eq. 1) has been proposed for representing all the digital signal processing required by a dense microphone array. Encoding, manipulation and decoding of the spatial information are all obtained with a matrix of FIR filters.
Not-linear and time-variant effects can be accommodated simply changing the filtering coefficients when needed, an operation which is easy and occurs without artefacts employing modern and efficient convolution engines.
The approach is formally identical whatever "format" is employed for storing and transferring the spatial information, as the content of each channel is always interpreted as the signal pertinent to a given "virtual microphone", and this holds for any possible sound format (A-format, B-format, P-format, G-format, etc.)
Although the filtering coefficients for each of the above stages and formats can be derived theoretically, the proposed general approach is to compute them numerically, solving eq. 3. And, whenever possible, the "known" information to be entered in eq. 3 should be **measured**, so that the numerical inversion procedure automatically compensates also for acoustical problems (shielding, diffractions, reflections) and for the unavoidable deviations of transducers.
This makes our approach "theory-less", but of course this poses great demand on the capability of performing accurate, unbiased electroacoustic measurements: the authors did devote a lot of effort developing and perfecting reliable measurement methods [8, 9], which are required for being able to substitute the theoretical knowledge with experimental results.
The proposed approach has been applied to 4 different dense microphone arrays, and the performances obtained with our purely-numerical processing method have been favorably compared with traditional processing tools based on theoretical solutions.
It resulted possible to synthesize virtual microphones more directive than what was theoretically expected (for example, a 2nd-order cardioid from a 1 st-order tetrahedral probe), to perform "fractional" rotations and zooming, and to decode to ill-conditioned loudspeaker systems.
In conclusion, the new, unified approach can deal perfectly with existing dense microphone arrays, and allows for the usage of new types of arrays, having "strange" geometries, for which a theoretical approach is unfeasible.
These new microphone arrays provide an advantageous trade-off between the extension of their effective panorama and the possibility to create virtual microphones with very sharp and smooth polar patterns.

REFERENCES

[1] Farina, Angelo; Capra, Andrea; Chiesi, Lorenzo; Scopece, Leonardo, "A Spherical Microphone Array for Synthesizing Virtual Directive Microphones in Live Broadcasting and in Post Production", *AES Conference:40th International Conference: Spatial Audio: Sense the Sound of Space*, Paper Number 3-1, October 2010.

[2] Farina, Angelo; Amendola, Alberto; Chiesi, Lorenzo; Capra, Andrea; Campanini, Simone, "Spatial PCM Sampling: A New Method for Sound Recording and Playback", *52nd AES Conference: International Conference: Sound Field Control - Engineering and Perception*, Paper Number 7-2, September 2013.

[3] Pulkki, Ville, "Spatial Sound Reproduction with Directional Audio Coding", *JAES* Volume 55 Issue 6 pp. 503-516, June 2007.

[4] Berge, S. Barrett, N., "High Angular Resolution Planewave Expansion" *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris, May 6-7, 2010.

[5] Kirkeby, O., Nelson, P.A., "Digital Filter Design for Inversion Problems in Sound Reproduction", *JAES Volume 47, Issue 7/8* pp. 583-595; July 1999.

[6] Kirkeby, .; Rubak, P.; Nelson, P.A.; Farina, A., "Design of Cross-Talk Cancellation Networks by Using Fast Deconvolution", *AES 106th Convention*, Paper Number 4916, Munich, May 1999.

[7] Jens Meyer, Gary W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield", *ICASSP 2002*, pp. 1781-1784.

[8] A.Farina – "Simultaneous measurement of impulse response and distortion with a swept-sine technique", *110th AES Convention*, Paris, February 2000.

[9] Farina, A., "Advancements in impulse response measurements by sine sweeps", *122nd AES Convention*, Vienna, Austria, May 2007.