Ambiophonic Principles for the Recording and Reproduction of Surround Sound for Music

Angelo Farina (1), Ralph Glasgal (2), Enrico Armelloni (1), Anders Torger (1)

(1) Industrial Eng. Dept., Università di Parma, Via delle Scienze – 43100 Parma - ITALY
E-MAIL: farina@pcfarina.eng.unipr.it - HTTP://pcfarina.eng.unipr.it

(2) Ambiophonics Institute, 4 Piermont Road, Rockleigh, New Jersey 07647, USA
E-MAIL: glasgal@ambiophonics.org - HTTP://www.ambiophonics.org

This paper discusses the psychoacoustical background and the computational issues involved in the real-time implementation of a complete Ambiophonics reproduction system based on binaural technology. Ambiophonics, which requires only two media channels, evolved from previously known approaches such as the reproduction of binaural recordings over closely spaced loudspeakers through cross-talk cancellation, and the reconstruction of hall ambience by convolution from suitable impulse responses. The equations for the design of the digital filter coefficients are derived with regard to the many possible kinds of pre-existing recordings (binaural, sphere, ORTF, M/S), and their implementation on available hardware and software platforms are described. The authors suggest psychoacoustic explanations for the perceived audible performance, and describe the first results of a comparative listening test, evaluating the realism of three periphonic surround reproduction systems: Stereo Dipole, Ambisonics and Ambiophonics.

1. INTRODUCTION

In recent years many different surround reproduction systems have been developed. Many of them, such as 5.1, are spatially limited and are not considered to be psychoacoustically valid methods for achieving a realistic reproduction of recorded music. Other paradigms that are capable, in principle, of complete periphony (reproduction of apparent acoustic sources everywhere in the space, over a complete sphere around the listener) have been proposed, but none of them has gained acceptance or become commercially available on the market, in part, because they were not compatible with the vast existing library of two channel LPs and CDs.

In this paper only potentially complete periphonic systems are considered, with the goal of providing the mathematical description and the implementation details of one of these methods, termed Ambiophonics; it will be shown that this method, used only in reproduction or used in both recording and reproduction, makes use of physical principles and digital filtering techniques, which are also found separately in other periphonic surround methods, but are here coupled together in a consistent and psychoacoustically correct form.

Most periphonic methods fall in one of two broad categories, summarized here:

A) Binaural methods: the sound field is originally recorded with some sort of dummy head microphone, and reproduced by delivering the recorded signals unaltered to the entrances of the ear canals of the listener.

B) Wavefield reconstruction methods: the system replicates the wavefronts, impinging on an array of microphones in the original space, through the use of coarse or dense arrays of loudspeakers during reproduction in a different space.

The binaural methods have the advantage of requiring the recording and transmission of just two channels, and are thus compatible with traditional two-channel stereophony, a form of monodimensional reproduction (virtual sound sources located on a line). Possible methods of replicating the recorded acoustic pressure signals at the ears of the listener include headphone reproduction (with or without head tracking) and loudspeaker reproduction. In the latter case, cross-talk cancellation is usually required, for canceling the spurious signals that go to the “wrong” ear.

The common binaural methods are very sensitive to both the shape and directional characteristics of the original microphone employed for the recording and of the particular human head of the listener [1]. If these two characteristics do not match well, the spatial illusion is poor and the reproduced sound field is judged unnatural. Furthermore, localization errors are common, particularly for sound sources near the median plane (front-back confusion, height uncertainty). High-end audiophile-level results can only be obtained when the recording mount is the actual human head of the listener, equipped with wearable binaural microphones. Thus a pure binaural technique is limited to the amateur recording of live music, and is obviously not applicable to the reproduction of the large catalogue of already existing recordings often made using a so-called stereo microphone which is usually inherently unsuitable for binaural reproduction especially via earphones.

The wavefield reconstruction methods usually require many more than two channels for transmitting the spatial information: the most basic method, 1st-order Ambisonics [2], requires four channels, carrying the pressure signal and the three particle velocity Cartesian components, all recorded at a single point in space. Although an Ambisonics system can employ a much larger number of loudspeakers for reproduction, the wavefront reconstruction capabilities are somewhat limited, and the localization of sound sources is not robust since a 1st-order system only samples the spherical space with 1st-order spherical harmonics [3].

Much more accurate wavefront reconstruction methods have been proposed: 2nd order Ambisonics requires the recording and transmission of 9 channels, and is thus already impractical. The Wave Field Synthesis method [4] goes up to some hundreds of channels, which prevent its application for recorded music distribution, being applicable primarily in the real-time recreation of performing spaces. These methods are substantially incompatible with mainstream sound recording and delivery, which nowadays is almost completely done in two-channel stereo.

The 5.1 discrete surround sound system, primarily intended for movies, is not periphonic (no height information). The addition of a center speaker does not materially alter its nature as a monodimensional system that relies on phantom imaging. The 5.1 reproduction arrangement also does little to enhance the realistic reproduction of existing two-channel material on LP or CD.

2. THE AMBIOPHONICS METHOD

The goal of the Ambiophonics reproduction method is to create a realistic listening experience starting from existing 2-channel or even 5.1 recordings. Fortunately, the recordings themselves are not usually predistorted by the stereo reproduction process. That is, the recordings do not contain crosstalk and do not know that they will be listened to via a stereo speaker triangle that engenders crosstalk, requires phantom imaging rather than binaural localization, generates comb filtering and introduces pinna/HRTF angle errors. Normal recordings typically include very limited “3D surround” information. Of course, the missing information must be recreated in some way: this is done by means of convolution with suitable room impulse responses.

The method can be basically explained as the superposition of two simultaneous periphonic reproduction systems: cross-talk cancelled reproduction over a pair of closely-spaced loudspeakers (as is usually done for binaural loudspeaker reproduction), and approximate wavefront reconstruction with an Ambisonics array, being fed with reconstructed hall ambience signals derived from the left and right direct sound disc channels convolved with a set of weakly-correlated real hall impulse responses.

Fig.s 1 and 2 show the basic scheme of the two parts of the system.

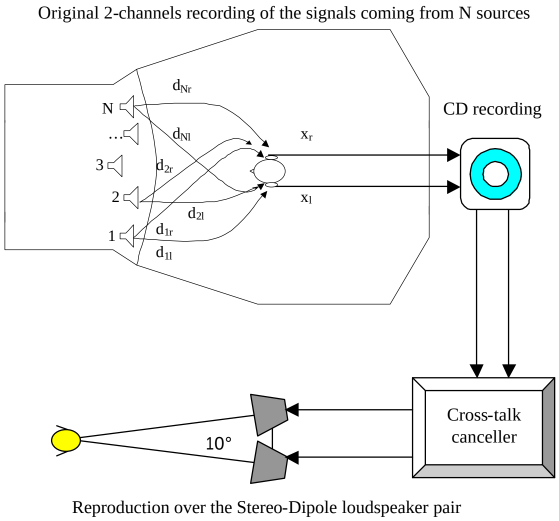

Original 2-channels recording of the signals coming from N sources
N
…
3
2
1
d_{Nr}
d_{Nl}
d_{2r}
d_{2l}
d_{1r}
d_{1l}
x_r
x_l
CD recording
Cross-talk canceller
10°
Reproduction over the Stereo-Dipole loudspeaker pair

Fig. 1 – Stereo-dipole reproduction through cross-talk canceling digital filters

The crosstalk cancellation operation is performed through the convolution of the two input signals with a set of 4 inverse filters, computed taking into account the kind of microphone employed for the recording. These inverse filters “cancel out” a great part of the microphone-dependent spatial effects (for example, the particular pinna coloration of a dummy head, in the case of binaural recordings), and thus leave each listener capable of hearing “with his own ears”. Of course, this deconvolution is easiest if the microphone did not introduce very sharp filtering curves: in fact the Ambiophonics reproduction is usually better starting from recordings made with a “pinna-less” dummy head (sphere microphone, or ORTF microphone). In principle, any kind of stereo microphone can be used, even a “virtual” one, as happens when the stereo mix is obtained by level panning many monophonic recordings of various instruments or vocalists. Thus there is almost no two-channel recording that does not benefit from being reproduced Ambiophonically.

In the following section, the mathematical details for the derivation of the cross-talk canceling inverse filters will be described.

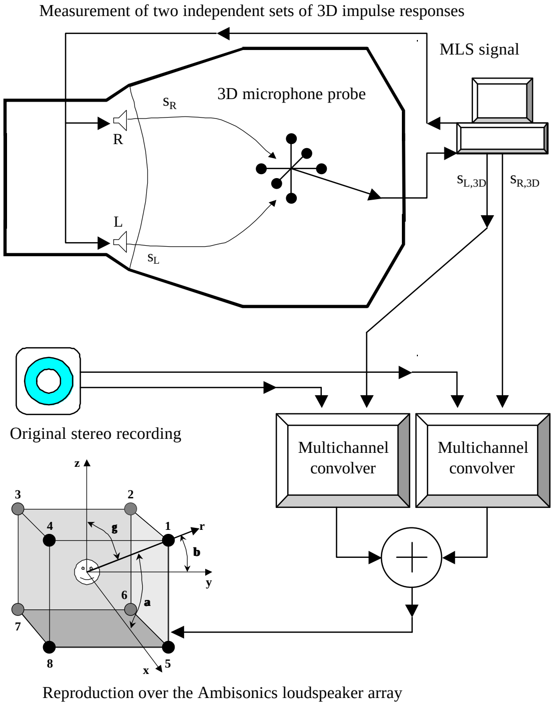

Measurement of two independent sets of 3D impulse responses
MLS signal
3D microphone probe
s_R
R
L
s_L
$s_{L,3D}$
$s_{R,3D}$
Original stereo recording
Multichannel convolver
Multichannel convolver
z
3
2
4
g
1
r
b
y
6
a
7
8
x
5
Reproduction over the Ambisonics loudspeaker array

Fig. 2 – Virtual Ambisonics reproduction by convolution with two sets of 3D impulse responses.

The surround loudspeaker array is responsible only for the reproduction of off- stage early reflections and reverberation tails. This means that the direct sound must be deleted from the impulse responses employed for convolution. In principle, these impulse responses can be obtained from Ambisonics decoding of a single B-format impulse response, synthesizing many virtual coincident (hyper)-cardioid microphones, each of them pointing towards a loudspeaker. But in practice it is preferable to consider these impulse response as an undersampled set of Wave Field Synthesis impulse responses, obtained by non coincident microphones, placed at relative positions from the main stereo microphone corresponding to the relative position of the reproduction loudspeaker from the listener. Following the WFS theory, the directivity pattern of each of these displaced microphones should depend on the directivity pattern of the loudspeaker being fed by its signal, but in practice this factor has been found to be very subtle, and can be neglected in most cases (provided that the loudspeakers employed for reproduction do not exhibit very strange directivity patterns or are in the presence of close reflecting surfaces).

In section 4 it will be shown what numerical processing is required to adapt the three-dimensional IRs, measured in a concert hall, to make them suitable for use as filters in Ambiophonics processing.

Section 5 presents available implementations of Ambiophonics, based on currently available hardware platforms, and on forthcoming software-only solutions, based on advanced convolution algorithms which have been recently implemented for real-time operation on low cost PCs.

Finally, subjective comparative tests were performed, in which it was possible to assess the preference for one of three simultaneous recording/reproduction methods: Binaural, Ambisonics and Ambiophonics. The tests were performed in a special listening room, equipped with a configurable reproduction system. At the time the tests were conducted, it was not possible to perform real-time reproduction of random recordings, but for comparative tests it was easy to pre-compute all the required signals, and leave the listener free to switch among the three systems.

The results of the subjective tests indicate that the Ambiophonics system is significantly preferred over the other two, followed by the binaural method (Stereo Dipole) and by Ambisonics. It was also confirmed that the “synthetic” Ambiophonics reproduction (in which the surround channels are derived by convolution from the main two channels) is almost indistinguishable from the “true” (directly recorded) surround obtained by processing the B-format recording. This is what makes it possible to obtain such satisfying results from existing two-channel recordings, although of course when 4 or more channels are available, it could be preferable, depending on the source material, to reproduce the frontal pair over the Stereo Dipole and employ the rear pair for convolution with the surround impulse responses.

3. CROSS-TALK CANCELLATION

The approach employed here is derived from the formulation originally developed by Kirkeby and Nelson [5], with refinement from one of the authors [6]. The following fig. 3 shows the cross-talk phenomenon in the reproduction space:

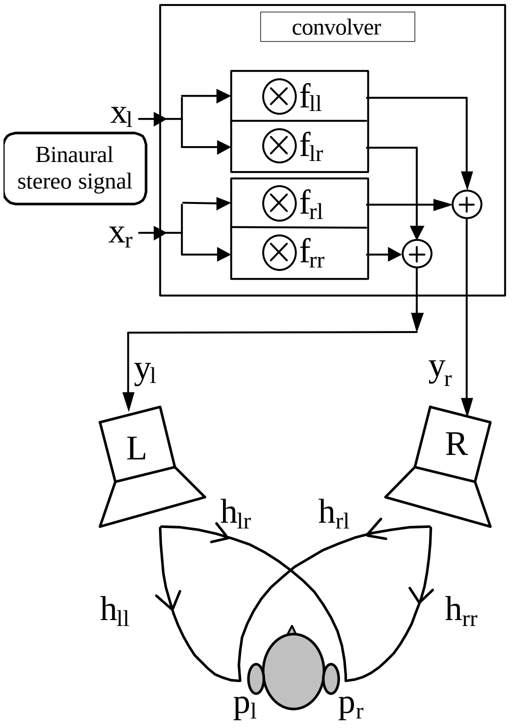


Fig. 3 – Cross-talk canceling scheme

The 4 cross-talk canceling filters f, which are convolved with the original binaural material, have to be designed so that the signals collected at the ears of the listener are identical to the original signals. Imposing that $p_l=x_l$ and $p_r=x_r$, a 4x4 linear equation system is obtained. Its solution yields:

$$\begin{cases} f_{ll} = (h_{rr}) \otimes InvDen \\ f_{lr} = (-h_{lr}) \otimes InvDen \\ f_{rl} = (-h_{rl}) \otimes InvDen \\ f_{rr} = (h_{ll}) \otimes InvDen \\ InvDen = InvFilter(h_{ll} \otimes h_{rr} - h_{lr} \otimes h_{rl}) \end{cases} \qquad (1)$$

The problem is the computation of the InvFilter (denominator), as its argument is generally a mixed-phase function. In the past, the authors attempted [7] to perform such an inversion employing the approximate methods suggested by Neely&Allen [8] and Mourjopoulos [9], but now the Kirkeby-Nelson frequency-domain regularization method is preferentially employed, due to its speed and robustness. A further improvement over the original method consists in the adoption of a frequency-dependent regularization parameter. In practice, the denominator is directly computed in the frequency domain, where the convolutions are simply multiplications, with the following formula:

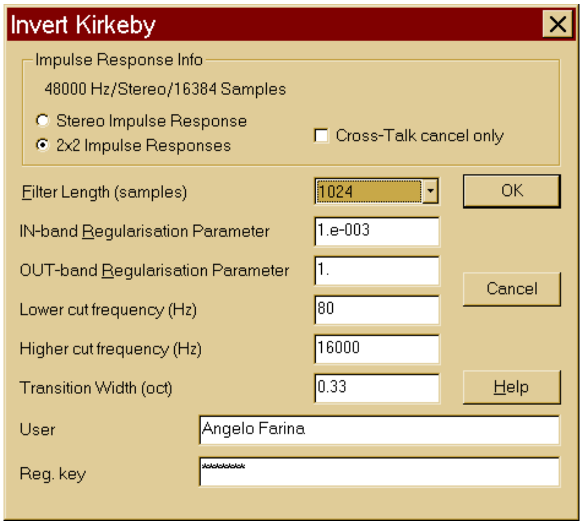
$$C(\omega) = FFT(h_{ll}) \cdot FFT(h_{rr}) - FFT(h_{lr}) \cdot FFT(h_{rl}) \qquad (2)$$

Then, the complex inverse of it is taken, adding a small, frequency-dependent regularization parameter:

$$InvDen(w) = \frac{Conj[C(w)]}{Conj[C(w)] \cdot C(w) + e(w)} \qquad (3)$$

In practice, ε(ω) is chosen with a constant, small value in the useful frequency range of the loudspeakers employed for reproduction (80 – 16k Hz in this case), and a much larger value outside the useful range. A smooth, logarithmic transition between the two values is interpolated over a transition band of 1/3 octave.

Fig. 4 shows the user interface of the software developed for computing the cross-talk canceling filters:



Invert Kirkeby

Impulse Response Info

48000 Hz/Stereo/16384 Samples

○ Stereo Impulse Response
● 2x2 Impulse Responses

□ Cross-Talk cancel only

Filter Length (samples)	1024
IN-band Regularisation Parameter	1.e-003
OUT-band Regularisation Parameter	1.
Lower cut frequency (Hz)	80
Higher cut frequency (Hz)	16000
Transition Width (oct)	0.33
User	Angelo Farina
Reg. key	******

Fig. 4 – User interface of the inverse filter module

This software tool was implemented as a plugin for CoolEdit [10], and it can directly process a stereo impulse response (assuming a symmetrical setup, so that $h_{ll}=h_{rr}$ and $h_{lr}=h_{rl}$), or a complete 2x2 impulse response set, obtained first by processing the binaural IR coming from the left loudspeaker, followed in time by the binaural IR coming from the right loudspeaker. In both cases, the output inverse filters are in the same format as the input IRs.

The computation is so fast (less than 100 ms) that it is easy to find the optimal values for the regularisation parameters by a trial and error method.

3.1 Real-time implementation of cross-talk canceling through Warped FIR filters

The filters described in the previous section are in the form of standard FIR filters. As they have to implement substantial boost and fine detail in the low frequency region, they have to be quite long (typically more than 4096 taps at 44.1 kHz). Thus it is almost impossible to implement them on standard DSP boards in the basic time-domain form.

Although frequency-domain implementation, as described below, can easily resolve this problem of running audiophile-quality cross-talk canceling filters on currently available DSP boards another possible approach is the use of Warped FIR structures. WFIR features a variable resolution in the frequency domain, and therefore is an effective variation in FIR filter design.

Let us consider the following bilinear transformation:

$$z = A_l(z) = \frac{z + l}{1 + z \cdot l} \tag{4}$$

where the parameter λ, referred to as warping coefficient can vary between –1 and 1. This transformation is the basis of the frequency warping technique. It results in a re-mapping of the complex plane, so that the z frequency plane is changed into a new ζ complex plane.

This bilinear transformation is graphically represented in fig. 5 as a function of λ.

The application of this transformation to the spectrum of an audio signal results in a stretching of the signal spectrum so that it becomes approximately logarithmic and thus more consistent with a psychoacoustics frequency scale, like the Bark scale [11].

The main advantage is that the transformed signal is more consistent with human hearing capabilities. Therefore the warped filters have higher accuracy at low frequencies, where the human ear has a higher sensitivity, and lower accuracy at high frequencies.

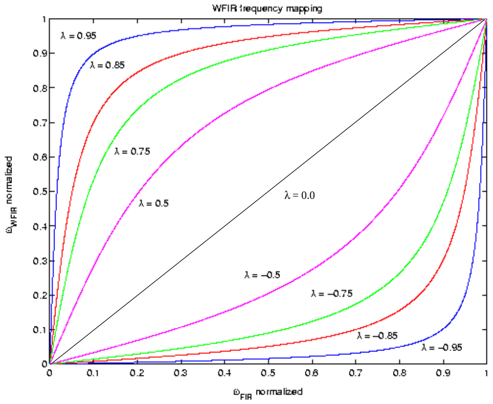


Figure 5 - Bilinear transformation of frequency with different λ values

In a classical FIR filter the frequency resolution is constant over the entire frequency range. Since human frequency resolution is about one third of an octave the equalization is unnecessarily fine at high frequencies and too coarse at low frequencies. Therefore very long FIR filters are required to obtain good results over the entire frequency range.

A warped filter based on the Bark scale provides a more efficient equalization at low frequencies. Specifically a warped FIR filter can be implemented with a number of taps ten times lower than those of a FIR filter, but still featuring the same low-frequency equalization. Its real-time implementation, however, requires more computational power. The Warped FIR structure is derived from the traditional FIR, where unit delays are replaced by the all-pass operators $D_1(z)$:

$$D_1(z) = \frac{z^{-1} - l}{1 - l \cdot z^{-1}} \tag{5}$$

Unfortunately, this structure is not suitable for real-time processing. Thus an equivalent explicit structure was developed, shown in fig. 6, which allows for efficient implementation [12].

Due to the introduction of the $D_1(z)$ all-pass block the warping produces a distortion of the complex plane. The analysis of the warped z-plane shows that the points on the unitary circle are kept on it, the points inside are kept inside, and the points outside are kept outside. Therefore an unstable system cannot become stable, while a stable system remains stable. This means that a warped FIR filter is always stable, even though it is no longer a “finite response” filter, as the network shown in fig. 6 contains loops.

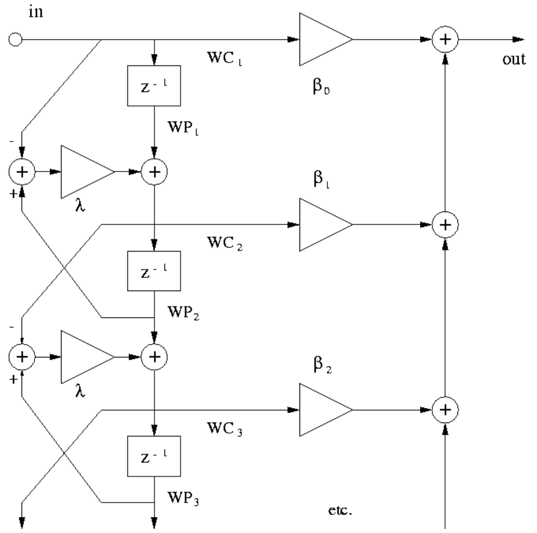


Figure 6 - Practical structure of the WFIR

It can be shown that for the points near +1 the distance from the unitary circle increases, whilst it decreases for the points near -1. Therefore the time-domain behavior of a warped signal is remarkably changed. As an example, let us consider a simple system, whose z-transformation features only a pole α on the real axis. Its expression in the z-domain, and in the time domain is respectively:

$$F(z) = \frac{1}{z-a} \rightarrow f(n) = \begin{cases} 0 & : n = 0 \\ |a|^{n-1} & : n > 0 \end{cases} \qquad (6)$$

The time constant τ is defined as the time necessary to reduce the system output to 36.7% of the maximum value, i.e. to the 1/e percentage of the maximum. Then, if α=0.99 the time constant is equal to about 100 samples. If the system is warped with λ=0.8, the above-mentioned pole (0.99) is re-mapped to 0.9135. On the other hand, a system with a high frequency pole near the Nyquist frequency, e.g. α=-0.99, would be re-mapped to -0.9989. This means that the time constant for the low frequency pole is just 12, whilst it is 900 samples for the high frequency pole.

In other words, when an impulse response is warped with a positive λ, the low frequency information is compressed in the first samples of the warped impulse response, while the high frequency components are stretched toward the last samples. Thus, a warped impulse response can be truncated after a few samples, without losing low frequency information. This property holds especially for high values of λ.

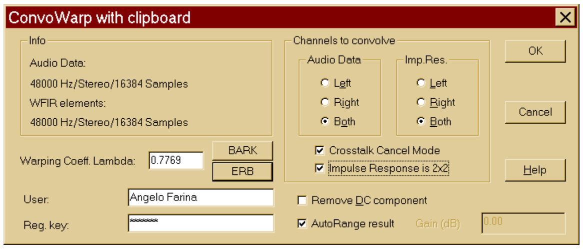
3.2 Implementation of the WFIR structure as an audio plugin and as DSP code

The WFIR structure illustrated in fig. 6 was first translated into an equivalent C-language code, suitable to operate on discrete-time samples of a sound waveform. The algorithm can be implemented with a single cycle, which is repeated as many times as the number of coefficients of the WFIR. This implementation was done independently by two of the authors [13, 14], with slightly different goals. Torger’s implementation is freely available as GNU public domain software.

The body of the cycle requires 3 multiplications and 3 sums, plus 4 memory operations (3 retrieves and one store). The algorithm requires a memory space as long as the number of filter coefficients, in order to store the partial sums of each stage.

In comparison, the traditional FIR algorithm is much cheaper, as the body of its main cycle requires only a single multiplication and addition, and two memory operations (retrieving the coefficient and the sample). The related computational cost is thus approximately given by the ratio 10/4, provided that we assign the same weight to multiplication, addition and memory operations.

Then the C code was embedded in a CoolEdit plugin, in order to mimic the behavior of the DSP code (allowing for listening tests, although not in real-time) and to pre-warp the measured impulse responses. This means that it is possible to process the measured IRs in the warped domain, deriving directly the coefficients of the cross-talk canceling inverse filters. Fig. 7 shows the user interface of the “ConvoWarp” module.



Info
Audio Data:
48000 Hz/Stereo/16384 Samples
WFIR elements:
48000 Hz/Stereo/16384 Samples

Warping Coeff. Lambda: 0.7769

Channels to convolve
Audio Data: ○ Left ○ Right ● Both
Imp.Res.: ○ Left ○ Right ● Both

☑ Crosstalk Cancel Mode
☑ Impulse Response is 2x2

User: Angelo Farina
Reg. key: ********

□ Remove DC component
☑ AutoRange result
Gain (dB) 0.00

Fig. 7 – User interface of the WFIR module

From the user point of view this module simply requires one to store the WFIR coefficients on the clipboard (in WAV format), then allows for the processing of a stereo audio signal by up to four separate WFIR filters, and thus is ideal for cross-talk canceling networks.

When the plugin is employed for the pre-warping of measured impulse response coefficients (or of pre-computed cross-talk canceling filters), a negative value of λ must be used. Furthermore, a discrete Dirac delta function is fed into the warped filter structure using the measured impulse response as the coefficients of the WFIR. This operation produces the set of pre-warped coefficients.

Both WFIR and FIR architectures have been implemented on an evaluation board equipped with an AD 21065L SHARC processor in assembly code for efficiency purposes. This DSP unit is capable of real-time processing up to approximately 900 multiply-add operations at a sampling rate of 48 kHz. This means that with the traditional FIR implementation approximately 225 taps for each of the 4 cross-talk canceling filters are allowed at maximum.

```
LCNTR=Wfilter_taps-1 , DO wmac_rr UNTIL LCE;
F12=F2*F4, F9=dm(I5,M7), F4=pm(I9,M8);
F10=F2*F5, F8=F8+F12, F9=dm(I5,M6);
F1=F9-F10, F9=dm(I5,0);
F10=F1*F7, dm(I5,M7)=F2;
wmac_rr:            F2=F9+F10;
```

Fig. 8 – Analog Devices AD20165L code for WFIR

Exploiting the parallel processing capabilities of the SHARC processor, the WFIR code was implemented with only 5 lines of code, as shown in fig. 8, and thus, in this case, the computational cost of the WFIR is exactly 5 times of that of a traditional FIR. Thus, the maximum number of taps for each WFIR cross-talk canceller is 45.

3.3 Experimental results

Experiments and listening tests were performed at ASK Industry, Italy, inside a treated listening room, equipped with a pair of professional-grade self-powered monitor loudspeakers (Dynaudio). The loudspeakers were arranged in the stereo dipole configuration (distance between the acoustic centers was 350 mm, and the listener's head was 2 m from the front of the speakers). First, the binaural impulse responses were measured, making use of a Bruel&Kjaer head and torso simulator type 4100, a PC equipped with a professional sound board (Echo Layla) and the Aurora measuring software [15]. Fig. 10 shows a typical measurement session.

Fig. 11 shows the measured impulse responses of the system, corresponding to the 4 impulse responses referred to as *h* in fig. 3.

First, a set of very long inverse FIRs was computed (2048 taps each), as shown in fig. 12. When these filters are employed (running them with the Aurora convolution plugin under CoolEdit), a good frequency response and cross-talk cancellation is obtained, as shown in fig. 13.

Despite the length of these inverse filters, the response is good only above 600 Hz: at lower frequency the response is quite uneven, although the cancellation of the cross talk remains very effective.

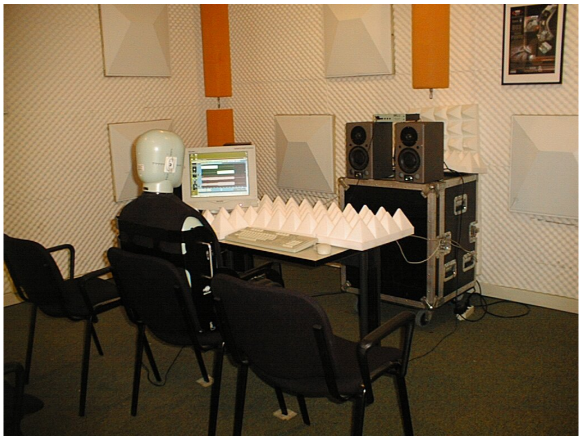
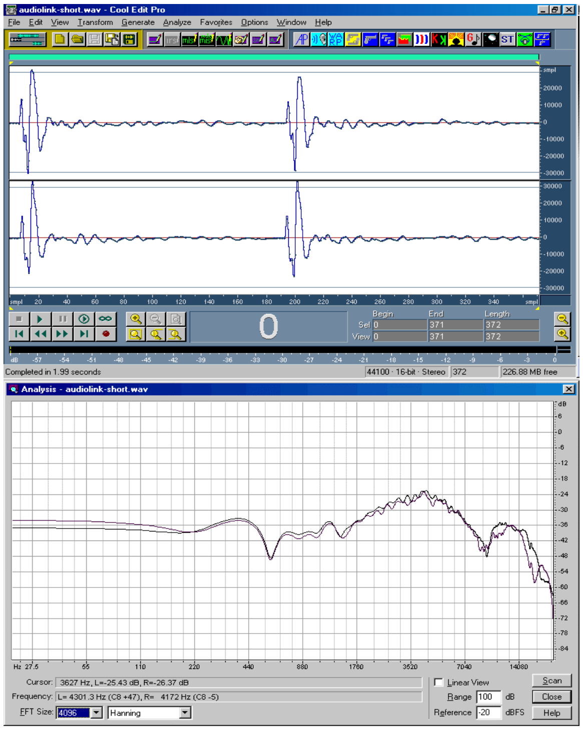


Fig. 10 – Measurements in the ASK listening room

After this, the "short" FIR and WFIR inverse filters were derived, respectively 220 and 42 taps long. Fig. 14 shows the effect of such short FIR inverse filters when applied to the system of fig. 8.

Similarly, figs. 15 and 16 show the WFIR coefficients, and the filtering effect of the WFIR structure.

From these results, it is clear that the short FIR only behaves correctly at medium/high frequency, providing poor overall response, with great problems at low frequency. On the other hand, the WFIR gives an overall flat spectrum starting from much lower frequencies, although the cross-talk cancellation is somewhat less effective, and the time response is slightly "smeared".


File Edit View Transform Generate Analyze Favorites Options Window Help
smpl 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 smpl
20000 10000 0 -10000 -20000 -30000
dB -57 -54 -51 -48 -45 -42 -39 -36 -33 -30 -27 -24 -21 -18 -15 -12 -9 -6 -3 0
dB 6 0 -6 -12 -18 -24 -30 -36 -42 -48 -54 -60 -66 -72 -78 -84
Hz 27.5 55 110 220 440 880 1760 3520 7040 14080
Scan
Close
Help

audiolink-short.wav - Cool Edit Pro

0

	Begin	End	Length
Sel	0	371	372
View	0	371	372

Completed in 1.99 seconds 44100 · 16-bit · Stereo 372 226.88 MB free

Analysis - audiolink-short.wav

Cursor: 3627 Hz, L=-25.43 dB, R=-26.37 dB

Frequency: L= 4301.3 Hz (C8 +47), R= 4172 Hz (C8 -5)

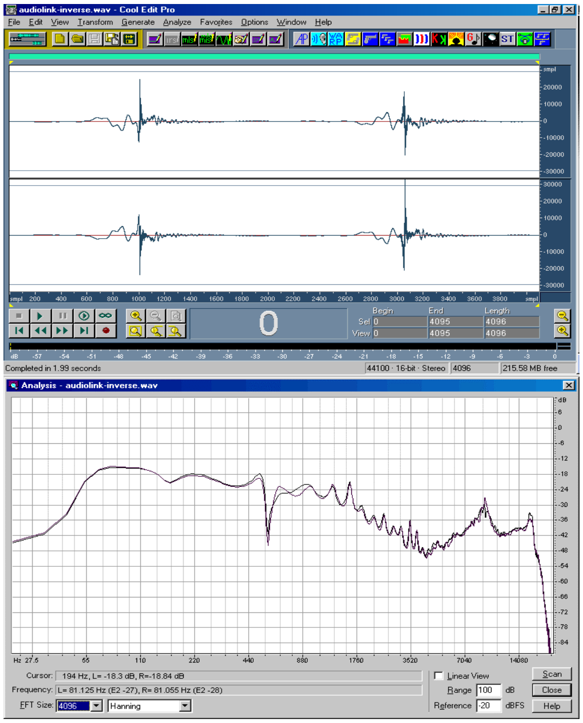
FFT Size: 4096 Hanning

□ Linear View

Range 100 dB

Reference -20 dBFS

Fig. 11 – Measured Binaural Impulse Responses and corresponding frequency response


File Edit View Transform Generate Analyze Favorites Options Window Help
smpl 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800 smpl
20000 10000 0 -10000 -20000 -30000
dB -57 -54 -51 -48 -45 -42 -39 -36 -33 -30 -27 -24 -21 -18 -15 -12 -9 -6 -3 0
dB 6 0 -6 -12 -18 -24 -30 -36 -42 -48 -54 -60 -66 -72 -78 -84
Hz 27.5 55 110 220 440 880 1760 3520 7040 14080
Scan
Close
Help

audiolink-inverse.wav - Cool Edit Pro

0

	Begin	End	Length
Sel	0	4095	4096
View	0	4095	4096

Completed in 1.99 seconds 44100 · 16-bit · Stereo 4096 215.58 MB free

Analysis - audiolink-inverse.wav

Cursor: 194 Hz, L= -18.3 dB, R=-18.84 dB

Frequency: L= 81.125 Hz (E2 -27), R= 81.055 Hz (E2 -28)

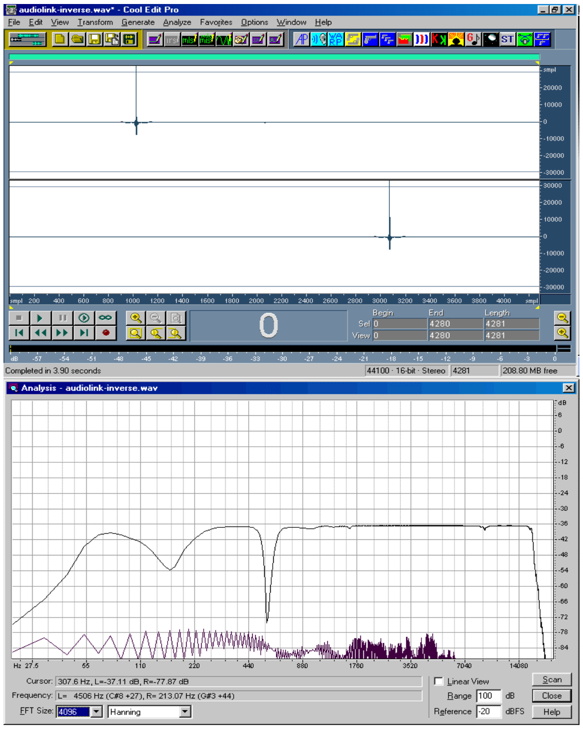
FFT Size: 4096 Hanning

□ Linear View

Range 100 dB

Reference -20 dBFS

Fig. 12 – Cross-talk cancellation with long FIR filters


File Edit View Transform Generate Analyze Favorites Options Window Help
smpl 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800 4000 smpl
20000 10000 0 -10000 -20000 -30000
dB -57 -54 -51 -48 -45 -42 -39 -36 -33 -30 -27 -24 -21 -18 -15 -12 -9 -6 -3 0
dB 6 0 -6 -12 -18 -24 -30 -36 -42 -48 -54 -60 -66 -72 -78 -84
Hz 27.5 55 110 220 440 880 1760 3520 7040 14080
Scan
Close
Help

audiolink-inverse.wav* - Cool Edit Pro

0

	Begin	End	Length
Sel	0	4280	4281
View	0	4280	4281

Completed in 3.90 seconds 44100 · 16-bit · Stereo 4281 208.80 MB free

Analysis - audiolink-inverse.wav

Cursor: 307.6 Hz, L=-37.11 dB, R=-77.87 dB

Frequency: L= 4506 Hz (C#8 +27), R= 213.07 Hz (G#3 +44)

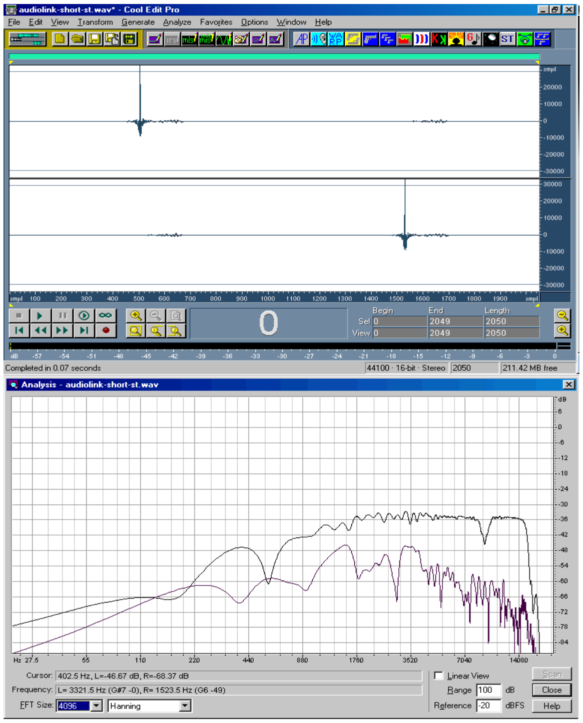
FFT Size: 4096 Hanning

□ Linear View

Range 100 dB

Reference -20 dBFS

Fig. 13 – Cross-talk cancellation with long FIR


File Edit View Transform Generate Analyze Favorites Options Window Help
smpl 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 smpl
20000 10000 0 -10000 -20000 -30000
dB -57 -54 -51 -48 -45 -42 -39 -36 -33 -30 -27 -24 -21 -18 -15 -12 -9 -6 -3 0
dB 6 0 -6 -12 -18 -24 -30 -36 -42 -48 -54 -60 -66 -72 -78 -84
Hz 27.5 55 110 220 440 880 1760 3520 7040 14080
Scan
Close
Help

audiolink-short-st.wav* - Cool Edit Pro

0

	Begin	End	Length
Sel	0	2049	2050
View	0	2049	2050

Completed in 0.07 seconds 44100 · 16-bit · Stereo 2050 211.42 MB free

Analysis - audiolink-short-st.wav

Cursor: 402.5 Hz, L=-46.67 dB, R=-68.37 dB

Frequency: L= 3321.5 Hz (G#7 -0), R= 1523.5 Hz (G6 -49)

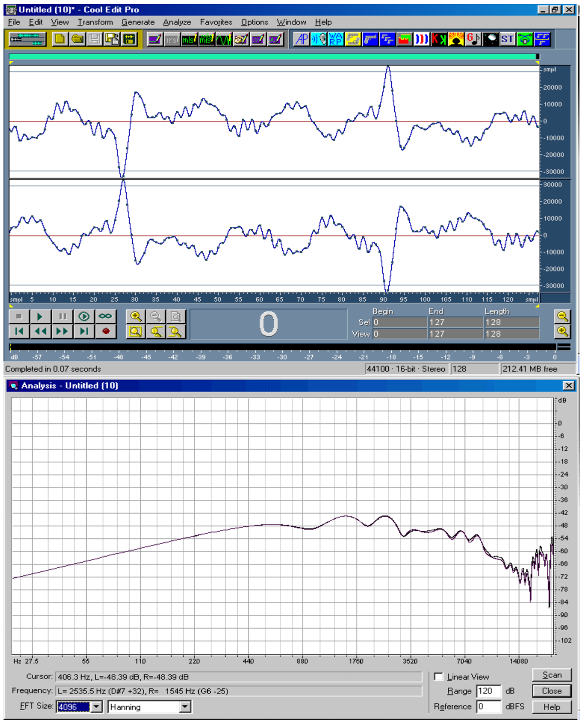
FFT Size: 4096 Hanning

□ Linear View

Range 100 dB

Reference -20 dBFS

Fig. 14 – Cross-talk cancellation with short FIR

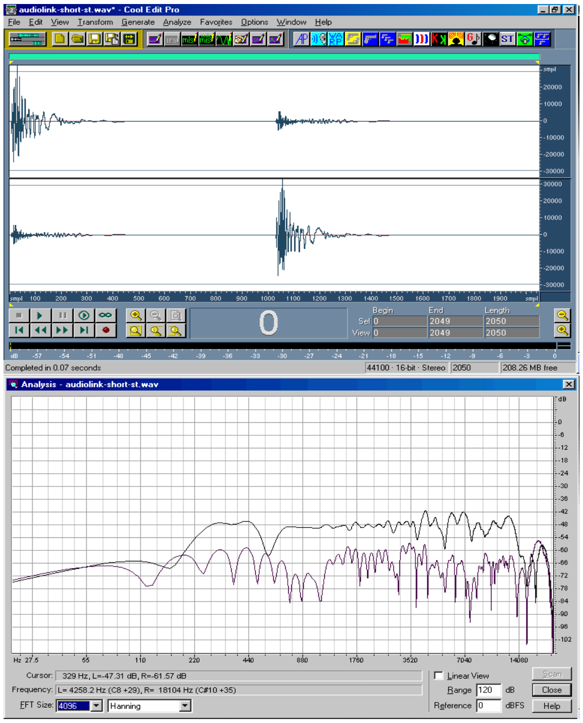

Untitled (10)* - Cool Edit Pro
File Edit View Transform Generate Analyze Favorites Options Window Help
Analysis - Untitled (10)
Hz 27.5 55 110 220 440 880 1760 3520 7040 14080
Linear View
Scan
Close
Help

Begin End Length
Sel 0 127 128
View 0 127 128

Completed in 0.07 seconds
44100 · 16-bit · Stereo 128 212.41 MB free

Cursor: 406.3 Hz, L=-48.39 dB, R=-48.39 dB
Frequency: L= 2535.5 Hz (D#7 +32), R= 1545 Hz (G6 -25)
FFT Size: 4096 Hanning
Range 120 dB
Reference 0 dBFS

Fig. 15 – WFIR coefficients


audiolink-short-st.wav* - Cool Edit Pro
File Edit View Transform Generate Analyze Favorites Options Window Help
Analysis - audiolink-short-st.wav
Hz 27.5 55 110 220 440 880 1760 3520 7040 14080
Linear View
Scan
Close
Help

Begin End Length
Sel 0 2049 2050
View 0 2049 2050

Completed in 0.07 seconds
44100 · 16-bit · Stereo 2050 208.26 MB free

Cursor: 329 Hz, L=-47.31 dB, R=-61.57 dB
Frequency: L= 4258.2 Hz (C8 +29), R= 18104 Hz (C#10 +35)
FFT Size: 4096 Hanning
Range 120 dB
Reference 0 dBFS

Fig. 16 – cross-talk cancellation with WFIR filters.

3.4 Subjective comparison

The audible performances of the two digital filtering techniques were compared in a blind subjective test. 14 normal-hearing subjects were employed, aged between 20 and 36, 6 were females. The subjects were not trained in listening tests, nor did they know anything about the research and the goals of the experiment. Each subject was comfortably seated at the "sweet spot" in front of the Stereo-Dipole loudspeaker pair. He was given control of the DSP unit through two selection buttons, which were labeled A (FIR) and B (WFIR). A CD player generated the test signals (binaural recording of natural sounds on the beach, and of music inside a car compartment). The listener was free to switch in any moment between A and B filters. He had to fill in a questionnaire containing 7 attributes, rating each of them on a 5-levels scale (insufficient, mediocre, sufficient, fair, good), for both A and B systems.

The results were analyzed using classical ANOVA [16] (performed thanks to the Excel analysis toolpack). The following table presents the statistical results (the 5% critical F-value was 4.2252, which means that values greater than it indicate that the difference between A and B is significant).

Question	Avg. A	Avg. B	Anova's F factor	Prob.
Overall appreciation	3.57	4.79	**34.47**	0.00%
Image localization	3.79	4.36	**4.38**	4.63%
Stage width	3.50	4.71	**21.72**	0.01%
Naturality	3.71	4.57	**10.88**	0.28%
Low frequency resp.	3.29	4.36	**11.56**	0.22%
Mid frequency resp.	3.79	4.07	1.60	21.7%
Hi frequency resp.	4.14	4.43	0.98	33.1%

Also the probability that A and B responses are the same was computed; the ANOVA's results can be seen in graphical form in fig. 17.

From the table above and from fig. 17, it is clear that system B (WFIR) was significantly better than system A in questions 1, 3, 4 and 5. The significance is at limit for question 2 (prob. 4.63%), and there is no substantial difference in question 6 and 7. This means that the WFIR is globally better, and, particularly, because it widens the stereo image, it is more natural, and has deeper low-frequency response. Some subjects reported also that system A is drier, whilst system B is softer (and this is certainly due to the time smearing already mentioned in the previous section).

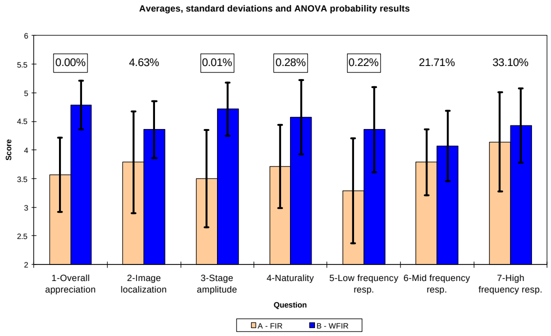


Fig. 17 – Anova of the subjective responses

3.5 Generalization to other stereo recordings

The above described procedure requires, in principle, that the same dummy head be employed, both during the recording of the stereo soundtrack, and for measuring the h impulse responses from which the cross-talk canceling filters are to be computed.

The obtained inverse filters are independent of the listener (each listener will add his own HRTF signature to the received sound), but they depend strongly on the listening setup (loudspeakers, room) and on the binaural microphone employed.

The last fact can be a problem for the reproduction of pre-recorded material available on most commercial CDs: in fact, only a small number of them were recorded with a binaural dummy head (in many cases the Neumann KU-100), and the vast majority consist of stereo recordings which can be categorized in one of two very different genders:

- coincident
- spaced

In the first case, only level differences appear between the two channels, and the recording is conceptually derived from two directional microphones coincident in space and with a certain angular divergence between their maximum sensitivity directions (Blumlein approach). In the second case, the signals come from two microphones placed at a relative distance similar to the human ears (approximately 170-180 mm), and thus they exhibit significant time misalignment between the two tracks, often with some further level difference caused by a physical obstacle between them (for example a rigid sphere) or by the directivity patterns of the microphones (which should in principle mimic the low-order HRTF spatial response). The first category also includes “virtual” stereo mixes, obtained by pure level-panning of monophonic recordings.

The spaced recordings are perfectly suited for reproduction with the cross-talk cancellation method, provided that the corresponding inverse filters are computed with the same microphone employed for the recordings. For this work, just two extreme cases were considered, namely the ORTF microphone (Schoeps MSTC64, 170 mm spaced cardioids with 110° aperture) and the sphere microphone (Schoeps KFM-360 or KFM-6). Usually recordings done with spaced microphones are easily identified as such, and often details on the microphone type and placement are specified on the CD cover.

The vast majority of released CDs fall in the category of coincident recordings, albeit most of them are really studio-made amplitude mixes of spot miked multichannel recordings. One could think that the lack of interchannel delay impedes proper cross-talk cancelled reproduction of these recordings: instead it turns out that delivering these signal through a “moderate” cross talk cancellation field yields a realistic spatial imaging (although not comparable with spaced recordings). “Moderate” cross talk cancellation refers here to the typical effect obtainable by mechanical barriers, instead of by means of digital filters.

As clearly demonstrated by R. Glasgal [17], a mechanical barrier placed between the loudspeakers in the Stereo Dipole configuration, and extending near the face of the listener, provides quite effective cross-talk cancellation at high frequency (above 1 kHz), and progressively much less cancellation towards low frequency, with no separation at all under 200 Hz. This arrangement seems to provide suitable localization cues at high frequency (where the spatial imaging is governed mainly by level difference between the two ears), and preserves the traditional cross-talk based imaging at low frequency, where there is not any phase difference encoded in the source material, and this difference has to be recreated by the diffraction around the head of the listener.

In conclusion, 4 different sets of inverse filters can be created, each of them specifically suited for a different recording technique: binaural, sphere, ORTF and coincident (M/S Blumlein). Selecting the optimal set of filters, almost any kind of recording can be reproduced successfully over a Stereo Dipole with cross-talk cancellation.

4 VIRTUAL AMBISONICS SURROUND

In most cases, the stereo recording provided on commercial CDs is obtained with good spatial information related to the position of the sound source and their direct soundfield, but very little “ambience” information is encoded in the source material. In fact, in traditional stereo reproduction, it is quite annoying to hear, superimposed on the direct sound coming from front, the discrete reflections and reverb which should arrive from the sides, above and behind the listener in a good concert hall.

So sound engineers tend to place their microphones very close to the sound source, and to shield them from “annoying” reflections and reverb coming from the back of the room. A realistic replication of a music listening experience requires that the whole three-dimensional sound space be reconstructed in the reproduction space.

This space can be obtained with the Ambisonics technique, although with very limited definition or localization of the direct sound source location. The basis for the Ambisonics method is the description of the spatial properties of the sound field by means of the B-format signal: it is a 4-channel signal, obtained from the sound pressure captured by an omnidirectional microphone (called W) and by the three first-order spherical harmonics of the sound pressure field, corresponding to the response of three figure-of-eight microphones aligned with the axes of a 3D Cartesian reference system (called X, Y and Z).

It must be clear here that the inventors of the Ambisonics technique [2,18] did not know anything about modern energy analysis of sound fields [19, 20, 3], although their very original theories were in a certain sense anticipatory of these subsequent modern developments. In practice, the spherical harmonics of the sound pressure field were easily confused with the Cartesian components of the particle velocity vector, as they are coincident, following the Euler’s equation, in the case of plane, progressive waves. In a generic sound field, though, pressure and velocity exhibit significant phase and gain mismatch, and these quantities should not be confused at present.

This is not a problem for “synthetic” B-format signals, obtained by panning a single mono track with proper gains, computed simply by using the values of the cosines of the angles between the intended direction of the sound with the three Cartesian axes; it is though a great problem for signals captured from a “true” sound intensity probe, which captures the real physical quantities (pressure and particle velocity components). The widely employed Soundfield microphone has a strange, intermediate behavior: it is close to a true sound intensity probe when the wavefront has little curvature, but deviates from it in the cases of strong curvature. However, it does not achieve the theoretical behavior of a pure cosine-weighted pressure microphone.

In a generic, reactive sound field, deriving accurate three-dimensional information from measurements done with a Soundfield microphone is not an easy task: although not corresponding to the definition of a B-format signal, it is simpler to process true pressure-velocity recordings obtained with a sound intensity probe, because the relationship between these physical quantities is mathematically known. But the signals coming out from a Soundfield mike are not so easily interpreted, because the published theory describing its behavior [18] is valid only with plane, progressive waves.

The process described here for creating a three-dimensional soundfield surrounding the listener is called “Virtual Ambisonics” because it is not based on native B-format recordings, but on B-format signals reconstructed by convolution of the original stereo recordings with B-format impulse responses.

As the B-format signal needs to be “decoded” for feeding a three dimensional loudspeaker array, and being that this decoding process is implementable as another convolution with a set of proper decoding filters [21], it is possible to connect the two convolutions into a single one: a set of three-dimensional IRs can be derived, which can be used directly as filters, applied by convolution to the original stereo recording, and then use them to drive the loudspeakers in the reproduction array. This combination can be seen as the synthesis of the impulse response obtainable by a virtual microphone, characterized by strong directivity, and pointing in the direction, relative to the listener, of the specific loudspeaker being considered, when the sound field is produced, in the original concert hall, by a sound source located on the stage. As this surround methodology need not be as accurate as the binaural one, just two positions of the sound source can be considered on the stage, as shown in fig. 2, corresponding to a generic “L” and “R” positions. Thus, for each loudspeaker in the reproduction space, two “3D Impulse Responses” are defined, named $s_{L,3D}$ and $s_{R,3D}$ respectively: the speaker feed is obtained as the sum of the results of the convolution of the two original signals with these filters.

4.1 Measurement of 3D Impulse Responses in theatres and concert halls

M.Gerzon [22] first proposed to start a systematic collection of 3D impulse responses measured in ancient theatres and concert halls, for assessing their acoustical behavior and preserving it for the posterity. His proposal found sympathetic response only very recently, with the publication of the "Charta of Ferrara" [23] and the birth of an international group of researchers who agreed on the experimental methodology for collecting these measurements [24].

Only a small number of theatres have yielded a complete three-dimensional impulse response characterization up till now. Among them, we employed for the present work the IRs measured in three Italian theatres:

- Gran Teatro La Scala in Milan
- Teatro Comunale in Ferrara
- Teatro Verdi in Trieste

The following table reports the main technical data regarding the measurement technique employed in each of these three rooms:

Room	La Scala Milano	Comunale Ferrara	Verdi Trieste
Dodechaed.	Norsonic	Look Line	Look Line
Excitation signal	MLS order 16	Log sweep 5 s	Log sweep 15 s
Microphone	7 stacked positions of a B&K ½ type 4166	3D sound intensity probe (B&K type wa0447)	Soundfield MK-V
Sound Board	MLSSA	Echo Layla	Echo Layla
Sampl. rate	60606 Hz	44100 Hz	44100 Hz

The measured three-dimensional impulse responses of these three theatres can be downloaded from: HTTP://pcangelo.eng.unipr.it/public/AES19 .

Fig.s 18, 19 and 20 show, for each theatre, a schematic plan of the room with the positions of the sound sources and of the microphone.

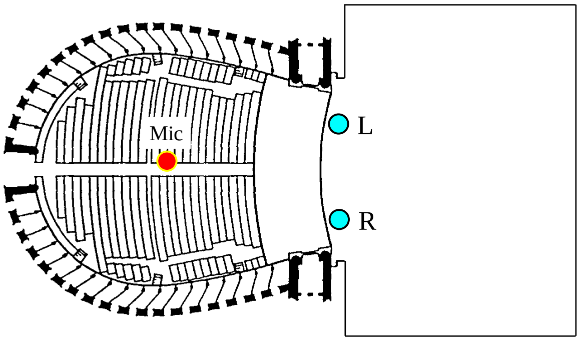


Fig. 18 – Plan of La Scala in Milan

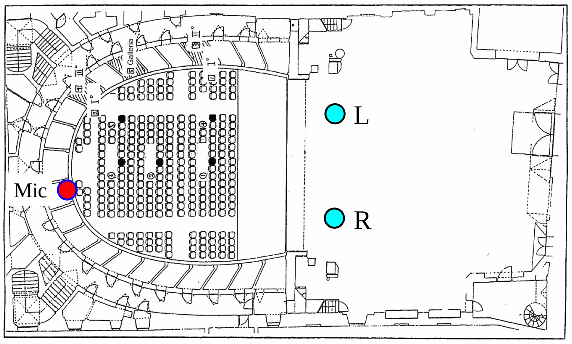


Fig. 19 – Plan of Teatro Comunale in Ferrara

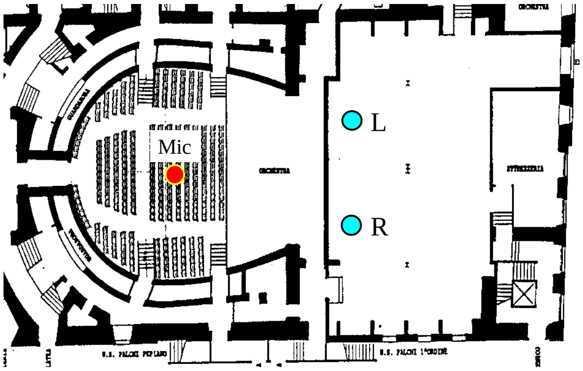


Fig. 20 – Plan of T. Verdi in Trieste

It must be noted that La Scala was in opera configuration, but the other two were in concert configuration, with a reflective orchestra shell mounted on the stage.

Some details are required regarding the three different kinds of microphones employed in these three rooms. In La Scala, a "virtual" 7-omnis microphonic array was employed [25], obtained by moving a single omnidirectional pressure microphone (B&K type 4166) into 7 close positions, and measuring a separate impulse response at each of them.

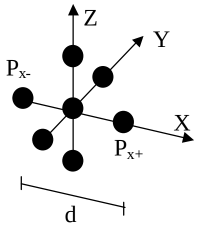


Fig. 21 – 7-omnis microphonic array

The geometry of the array is shown in fig. 21. From these 7 IRs, it is easy to extract 4 processed IRs, the first being simply the pressure response in the central microphone, and the other three being the particle velocity components along the three axes computed by means of the classic Euler's relationship, with the finite differences approximation commonly employed in sound intensity analyzers:

$$v_x(\tau)=\int_{-\infty}^{\tau}\left[\frac{p_{x+}(t)-p_{x-}(t)}{\rho\cdot d}\right]\cdot dt \qquad (7)$$

The same approach is employed for deriving pressure and velocity components from the 6 IRs measured at the Teatro Comunale in Ferrara by means of the three-dimensional sound intensity probe B&K type WA0447, which is shown in fig. 22.

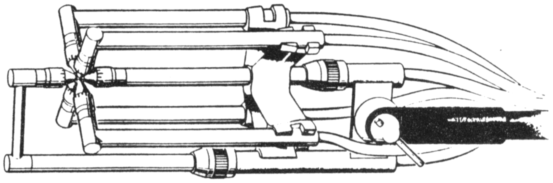


Fig. 22 – B&K type WA0447 sound intensity probe

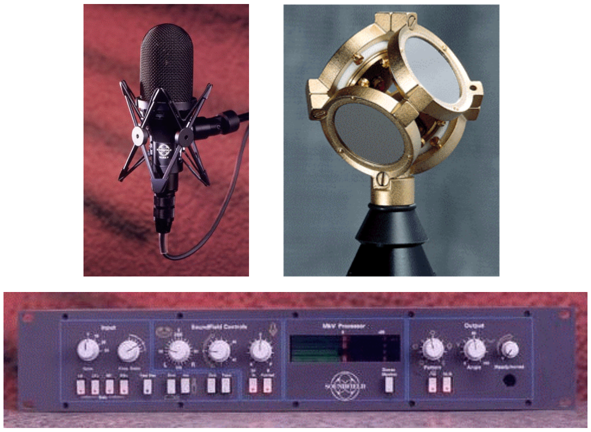


Fig. 23 – The Soundfield MK-V microphone

In this case there is no central pressure microphone, so the pressure signal has to be derived simply as the arithmetic mean of the 6 signals measured around the central virtual position. It must be noted, however, that this fact introduces some minor artifacts, as the 6 signals summed together are not perfectly coherent, and this introduces some high-frequency amplitude fluctuation, and a certain degree of smearing in the time domain.

In the third case, a standard Soundfield microphone was employed, as shown in fig. 23. This unit is equipped with a special electronic processor, which extracts the 4 signals labeled W,X,Y and Z.

As discussed earlier, in very reactive sound fields (close to the sound source, in a small, highly reverberant room) none of these three microphonic probes produces exactly the theoretical B-format signal (spherical harmonics of 0th and 1st order of the pressure field). But in the halls studied here, the sound source was very far away and the room was quite dry compared to its huge size (Italian theatres are known for their low reverberation times compared to north-European concert halls of the same size); this is demonstrated by fig. 24, which shows the measured reverberation times in the three theatres.

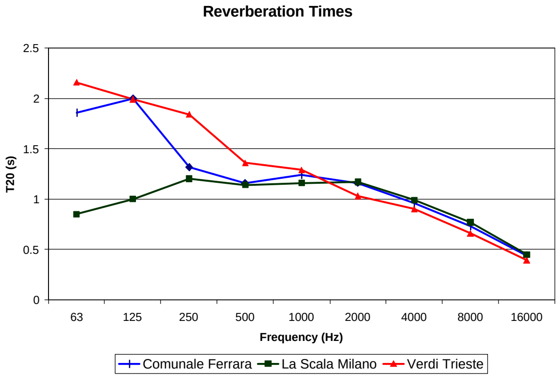


Fig. 24 – Reverberation time of the three theatres

Consequently, it can be assumed that in these three cases the measured pressure-velocity 4-channelIRs are a reasonable approximation of the theoretical B-format signals, and thus they can be processed with classic Ambisonics-like math for extracting the responses of virtual microphones, with proper directivity patterns, and pointing in any desired direction.

4.2 Derivation of directive microphone responses

The first point to clarify here is that the Soundfield microphones adhere to the old-style B-format standard, in which the W channel has a gain reduction of 3 dB compared with the other three signals XYZ. To make use of uniform notation, it is assumed here that this 3 dB gain reduction is immediately compensated for, and thus all three microphone probes result in the measurement of 4 IRs with the same absolute gain for all 4 channels.

The basis for the synthesis of a virtual microphone from the B-format signals is the fact that combining the response of an ominidirectional microphone (W) with a figure-of-eight microphone (X, for example), a cardioid response is obtained, as shown in fig. 25. If the gain of X is reduced in comparison with W, the response becomes sub-cardioid, but if the gain of X is greater than the gain of W, a hypercardioid response is obtained. This same fact is employed in the control unit of the Soundfield microphone, which allows for the recreation of the signal of two virtual microphones with selectable directivity patterns.

If the virtual microphone has to point along a generic direction described by its unitary vector $\vec{r}$, the response of the single figure-of-eight microphone X has to be replaced by a linear combination of the three signals XYZ, employing the directional cosines of $\vec{r}$ as weighting factors. Thus the response V of a generically-oriented virtual microphone can be computed as:

$$V(\vec{r}) = \frac{1}{2} \cdot \left[(2 - D) \cdot W + D \cdot \left(r_x \cdot X + r_y \cdot Y + r_z \cdot Z \right) \right] \quad (8)$$

In which the directivity factor D can assume these values:

D=0.0 ➔ omnidirectional

D=0.5 ➔ subcardioid

D=1.0 ➔ cardioid

D=1.5 ➔ hypercardioid

D=2.0 ➔ figure-of-eight

The above relationship (8) makes it easy to derive the proper impulse response corresponding to the position of each loudspeaker in the reproduction array, by post-processing the B-format IR measured in the theatre. This has to be repeated, of course, for both the B-format IRs, measured from the two source positions inside the theatre (L and R). For a reproduction array of 8 loudspeakers, for example, 16 synthetic IRs are obtained, and saved as 8 stereo waveforms (one for each loudspeaker).

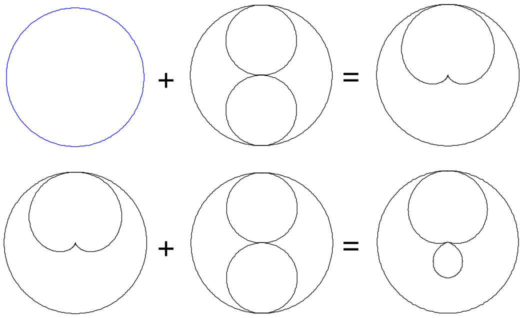


Fig. 25 – Synthesis of directive patterns

The feed for each loudspeaker can thus be derived simply convolving the stereo original recording with the stereo IR ($L \otimes IR_L$ and $R \otimes IR_R$) and summing (mixing together) the results.

By trial and error, it was found that the optimal value of the directivity factor D is approximately equal to 1.4 (hypercardioid), because this way each derived IR is much less correlated with the others. This corresponds approximately to the maximization of the field indicator r_E, as suggested in [26] by J. Daniel for optimizing the decoding of B-format recordings. This also corresponds roughly with the method suggested by Okubo [27].

4.3 Modification of the impulse responses

For pure Ambisonics reproduction, the impulse responses derived in the previous section are theoretically perfect, provided that the reproduction space is almost completely anechoic.

This was verified by a direct comparison between a live B-format recording made in the Teatro Comunale in Ferrara of a piano concert, which was then compared with the virtual reconstruction obtained by convolution of an anechoic recording of the same music piece with the impulse responses derived by the B-format measurement made with the sound source and the microphone placed exactly in the same positions as during the live performance. 13 of 18 listeners were unable to detect the difference between the live recording and the virtual one, and those who were capable of detecting a difference, were unable to reliably rank their preference for either one of the two recordings.

These tests were performed in the listening room of the University of Ferrara, making use of students of the Engineering Faculty as subjects (thus they were not sharped-ear musicians or audiophiles, and this can partially explain their inability to identify the difference between the two sound samples).

It was concluded, however, that the virtual implementation of Ambisonics by convolution is at least as good as the “live” Ambisonics recording/playback, and thus it is generally preferable, requiring the recording of just two channels instead of four, and producing a much wider dynamic range (the background noise of the XYZ channels is strongly reduced by the MLS or sine sweep measurement methods, and after convolution these channels have wider dynamic range than the corresponding channels coming from live recording).

But for the use of the Ambisonics array as a complement to the Stereo Dipole inside the Ambiophonics system, the derived three-dimensional impulse responses have to be processed: first of all, they do not need to reproduce the direct sound or the early reflections coming from the orchestra shell on the stage, because these are already included on the original stereo soundtrack, and are being reproduced with much finer detail by the frontal loudspeaker pair driven through the cross-talk canceling filters.

This means that the first part of the impulse response must be silenced (not cut away), to preserve the proper delay of the subsequent reverberant tail in relationship with the direct sound being reproduced by the stereo dipole.

In reality, the proper time alignment between the signal being reproduced through the stereo dipole and through the surround array must be checked, taking into account two other facts:

- the position of the main microphone during the original stereo recording is usually much closer to the sound source than any real listener in a concert hall, particularly looking at the place where the three-dimensional IRs were measured, as shown in fig. 18, 19 and 20.
- the cross-talk canceling filters introduce a significant delay (approximately half their length), and thus they tend to partially compensate the previous statement, causing a substantial increase of the source-receiver apparent distance.

It seems, however, that a few milliseconds of error in the delay of the reverberant tail with respect to the direct sound does not cause anything harmful, although, of course, the perceived distance from the stage is slightly changed.

The second modification required is a proper amplitude-shaping of the IRs. This is due to the fact that the reproduction space is not anechoic, and thus its reverberation (*h*) tends to add to the reverberation of the original theatre (*s*). We consider a simple exponential model of the two IRs, in the form:

$$\begin{aligned} s(\tau) &= w(\tau)\cdot \exp\left(-6.91\cdot\frac{\tau}{T_s}\right) \\ h(\tau) &= w(\tau)\cdot \exp\left(-6.91\cdot\frac{\tau}{T_h}\right) \end{aligned} \qquad (9)$$

in which T_h and T_s are respectively the reverberation time of the original theatre and of the reproduction space, and w(τ) is a white noise random process.

During the reproduction, the two IRs are convolved together, resulting in a global IR having a longer reverberation time:

$$s'(\tau) = s(\tau)\otimes h(\tau) = \int_0^{T_h} s(\tau - t)\cdot h(t)\cdot dt \qquad (10)$$

It must be noted that the convolution of two purely exponential decays is not another exponential decay, as clearly shown in fig. 26: the resulting impulse response exhibit a complex shape, with an initial part during which the amplitude of the signal increases, followed by a decay with a non-constant slope.

It can be seen from fig. 26 that the final slope of the convolved impulse response asymptotically tends to the slope of the original IR: in fact the EDT value is significantly increased (1.46 s), whilst T_{30} (1.05 s) is only slightly larger than the original theoretical value (1.0 s). In practice, the most important effect is that there is much more energy in the reverberant tail (after 1.9 s the backward-integrated curve is approximately 6 dB higher): this means that the values of most important early-to-late energy ratios have been substantially altered. The value of Center Time, for example, is 156 ms instead of the original 73 ms.

It might seem straightforward to solve the above problems by creation of a mathematically exact inverse filter: in principle, both the Mourjopoulos [9] and Kirkeby [6] theories allow for the creation of inverse filters, which can be convolved with the signal removing the effect of *h(t).* Unfortunately this is true only at the exact point where *h(t)* is measured. At all other points of the reproduction space the convolution with these inverse filters cause even more reverberant energy to be added.

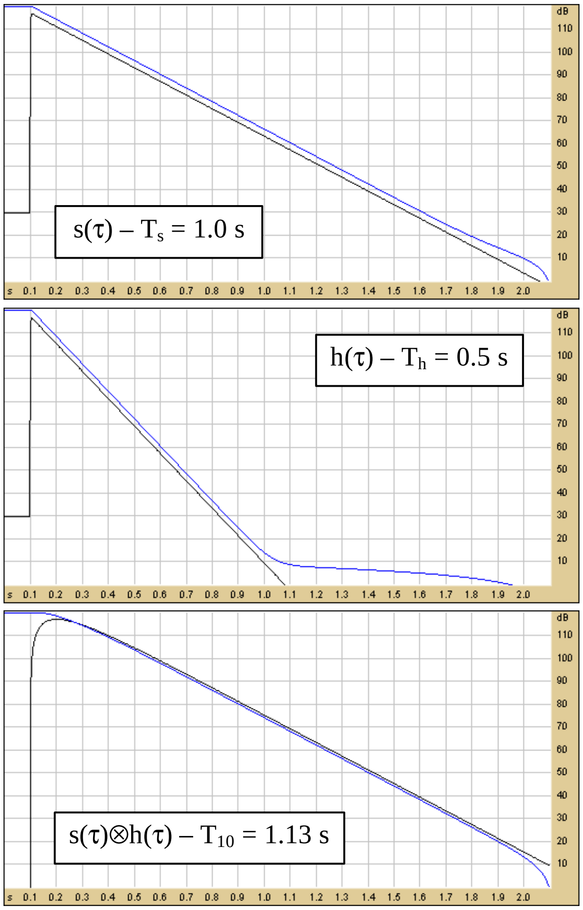
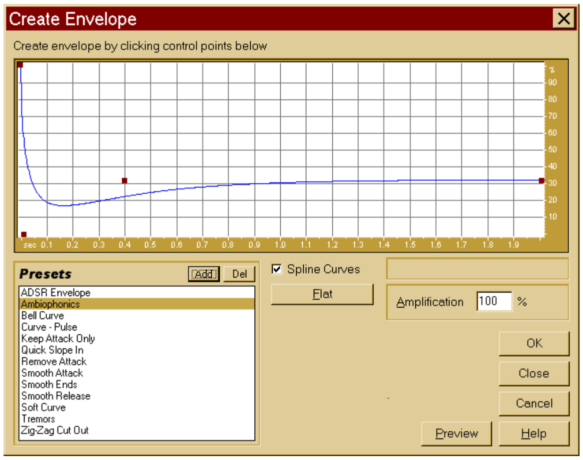


Fig. 26 – Convolution of two purely exponential decays (T_s=1s, T_h=0.5s), resulting in a non-exponential decay

The only viable solution is thus an empirical editing of the impulse response *s(t)* before this is convolved with the signal to be reproduced, so that the result of the subsequent reproduction in the listening room maximally resembles the IR of the original space. In the case of the theoretical signals shown in fig 24 and described by the equations (9), it is necessary to apply to *s(t)* an amplitude modulation described by the shape shown in fig. 27.


Create Envelope

Create envelope by clicking control points below

Presets: Ambiophonics

☑ Spline Curves

Amplification 100 %

Fig. 27 – Amplitude shaping of s(τ)

It must be clear that with real impulse responses there is not (yet) any simple theory for optimizing this amplitude shaping: furthermore, as the reverberation time is frequency-dependent, often the amplitude modification of the IRs should be made by means of a time-varying octave-band equalization, which is possible with the standard editing tools of CoolEditPro. So in practice this adjustment is actually done by trial and error, depending on the acoustical properties of the listening room: of course this process cannot solve a major problem of the reproduction space, such as focusing or resonances, it can simply ameliorate the listening experience a little. The optimal solution is always obtained by proper room treatment, employing, if required, bass traps and other sound absorbing devices, ensuring a short reverberation time with uniform spectral behavior. When the reverberation time of the reproduction space is less than 1/5 of the reverberation time of the original theatre, the effect of the listening room becomes absolutely unnoticeable, being masked by the much larger reverberation of the original space.

5. IMPLEMENTATION

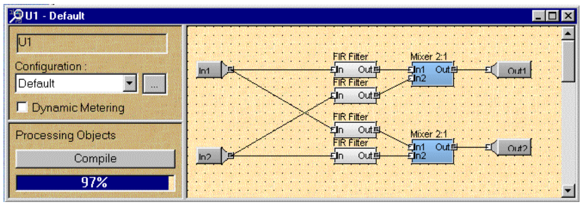
5.1 Hardware Implementation

From the theory discussed in the previous sections, it is clear how the complete Ambiophonics system can be implemented simply by means of multiple convolution of the original input signals with a number of impulse responses. A typical system can have, for example, 10 loudspeakers: two for the frontal stereo dipole, and 8 for a three-dimensional surround array.

In the most common case of a stereo (two-channel) source recording, each of these ten loudspeakers needs to be fed with the mix of the results of the convolution of the two input channels with two loudspeaker-specific impulse responses: these are cross-talk canceling filters in the case of the first two loudspeakers, and three-dimensional room IRs for the other 8 loudspeakers.

This means that, in principle, there is no need to differentiate the processing of the first two channels from the other: and in fact both the currently available software solutions do not differentiate between them in any way. But when the system is implemented by means of hardware digital convolvers, it can be useful to exploit the differences between the filters used to convolve: it was shown in section 3, for example, how a low-cost DSP board can be used for cross talk cancellation by frequency warping the FIR coefficients.

Another possible implementation makes use of a general-purpose multichannel DSP unit (Soundweb by BSS) for performing the cross-talk cancellation in real-time. This versatile machine can easily be configured for processing a cross-talk cancellation network based on FIR filters, by means of its very friendly graphical programming environment, as shown in fig. 28.



U1 - Default

U1

Configuration : Default

□ Dynamic Metering

Fig. 28 – Soundweb cross-talk cancellation network

In any case, the other 8 channels of room convolution need to be produced, again in real-time and with no processing latency, by means of hardware DSP convolvers. Nowadays the only units capable of doing this are the Lake DSP workstations and the Sony DRE-S777 processor, the JVC XP-A1010 processor having been discontinued.

Regarding the Lake DSP platform, a high-end Huron system is required for performing in real-time 16 convolutions with filters of suitable length (typically 128 kpoints at 48 kHz). The Lake's Huron system is highly versatile and easily configurable, and allows for easy substitution of the IRs; its only practical limit, apart from the cost, is the limited S/N ratio, principally due to low quality AD and DA converters; although they can be bypassed through the use of SPDIF digital interfaces, the internal processing is still done with fixed-point math, and thus the dynamic range and linearity are inherently limited compared with today's audiophile standards (24 bits resolution). It must be noted that, although the input signal can suitably be limited to 16-bits, the convolution process dramatically enlarges the dynamic range, restoring the very extended response to transients (particularly at a sudden end to the sound) which can be experienced in a real concert hall; this means that its output needs the full currently available 24 bits.

In this respect the Sony convolvers are actually state of the art, ensuring a completely uncompromised signal path and an outstanding signal purity. Their main defect is that these units were designed as two-channels units: an additional DSP board can be inserted for obtaining a four-channels system, but for obtaining 8 channels two units are needed.

The second limitation of the Sony convolvers is that they can employ only the impulse response sets contained in the special Sony CDs. One cannot load on these machines user-defined impulse responses. For at least three of the top ranked concert halls, Sony supplies multichannel impulse responses (apparently measured with substantially different microphone orientation and positions, not deriving them from a B-format IR as explained here in section 4), which can be successfully employed for Ambiophonics. The editing possibility contained in the convolvers is limited, and so the adaptation of the IRs to the reproduction space can be done only partially.

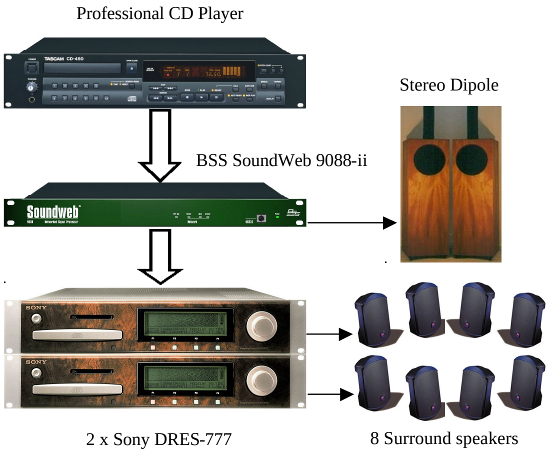


Fig. 29 – Hardware Ambiophonics system

In conclusion, an audiophile-grade complete Ambiophonics system can be built with hardware parts commonly available on the market, namely a BSS Soundweb for driving the Stereo Dipole and two 4-channels Sony DRE-S777 convolvers driving 8 surround channels, as shown in fig. 29. It must be remembered that a suitable listening room (with very little reverb and complete absence of other defects) is required, along with a good pair of loudspeakers for the Ambiopole; the other surround loudspeakers, on the other hand, are less demanding, and thus can be cheaper.

5.2 Software Implementation

At the time of writing, three possible solutions are known for obtaining Ambiophonics listening by means of a computer equipped with a high quality, multichannel sound board. These are described here only very briefly.

The first solution, which has been available now for some years, is based on the use of the CoolEditPro software. The soundtracks must be computed in advance, and stored on the hard disk inside a multichannel session of CoolEdit. Then the waveforms can be played at will. The main advantage of this approach is that, since the computation is done off-line, even a very slow computer can be employed (there is not any real-time constraint). Of course, it takes a while to convolve the original stereo waveforms with all those IRs, and to save the results as separate WAV files on the hard disk. The convolution can be done by means of any of the currently available plugins, including the one already shipped with CoolEditPro.

The second solution is based on the Ambiovolver software, developed by J.J. Lopez [28]. It is a real-time convolver, which employs the very efficient Intel FFT routines, though using the traditional select-save algorithm [29]. This solution comes in a WIN32 executable, with graphical user interface, and is capable of sustained convolution of two input channels with 10 stereo IRs, driving ten output channels. Depending on the computational power available, up to 128 kpoints long IRs can be employed. The main disadvantage of this software solution is that it causes a substantial delay between input and output (approximately twice the length of the IRs), which is not a problem for listening to pre-recorded material, but which is a serious problem in other applications (realtime acoustic display, synchronized audio-video virtual reality, etc.).

The third solution is the BruteFIR software by A. Torger [30]: it is GNU public domain software developed under Linux, which substantially does the same things as Ambiovolver, but employs a very clever convolution algorithm: the partitioned convolution scheme pioneered by Stockham [31] and refined by Soo and Pang [32].

This algorithm is substantially based on the select-save algorithm, but the IR is partitioned in many blocks of the same length, each of them being convolved separately. This reduces the overall latency to twice the length of these sub-blocks, which is significantly shorter than the whole IR lengths (typically 8 or 16 partitions of the IR are used). The typical latency is 100 - 400 ms, thus low enough for interactive use, but not low enough for synching with a zero delay video stream.

In theory, this should result in increased computational load (although significantly less than with zero-latency convolution achieved through hybrid filtering, as employed by Lake DSP and described by W.G. Gardner [33]), but in practice, since the implementation is very well suited to the memory management and parallel processing capabilities of modern processor architectures, it ends up being up to twice as efficient as the traditional unpartitioned select-save algorithm.

The extra power made free by this clever algorithm can be employed to run at a higher sampling rate (96 kHz), or to drive more channels (16 surround loudspeakers instead of 8), or even for processing source signals with more than two channels. In fact BruteFIR can be configured to process any number of inputs, and it could be advisable to start with a 6 channel recording (SACD or DVD-audio).

In this case, the best strategy for Ambiophonics reproduction may be to put the L and R channels through the cross-talk canceling filters, and mix the center channel (with proper delay corresponding to half the length of the cross-talk filters) on both the frontal loudspeakers. The surround loudspeaker array is then better driven by the surround channels of the original recording, although these usually already contain a large amount of room reflections and reverb, and thus the convolved IRs must be very short in this case; but they are still required because they ensure proper inter-loudspeaker relationships which build up the three-dimensional sound field.

Extending the theory already developed in section 4, the room IRs required in this case should have been measured with two sound source positions, in the original theatre, placed behind the microphone, at the same azimutal angular positions corresponding to the theoretical position of the "surround" loudspeakers in a 5.1 system (+/- 110°). These sound source positions should also be quite close to the microphone, so that the reverb-to-direct ratio in the measured IRs is small, and thus their convolution with the already reverberated surround channels of a 6-channel recording would not cause too much reverb in the reproduction space.

Depending on the recording, however, it could also be that the main channels L and R need to be included in the surround reproduction, being convolved with full-length room IRs as detailed in section 4. BruteFIR can easily accommodate these complex scenarios, because it can handle several input streams and do the mixing process very efficiently in the frequency domain, so that the number of forward and inverse FFTs is not increased.

It must be said that at this time the number of available 5 or 6-channel recordings is too small for allowing the evaluation of the optimal strategies for processing them Ambiophonically: nevertheless the system is potentially open to multichannel recordings, and has the potential for overcoming all the known limitations of 5.1 systems (lack of lateral sonic images, horizontal-only surround, presence of cross-talk related artifacts).

In conclusion, although at present Ambiophonics is usually obtained by means of hardware convolvers, in the near future complete systems will be built at very little cost thanks to standard PCs and multichannel sound boards. These software solutions circumvent all the limitations and the sound quality degradation typical of today's hardware solutions, and should allow also for the Ambiophonic playback of 6-channels recordings.

6. SUBJECTIVE COMPARATIVE TESTS

At the time of writing, only a very preliminary comparative test was performed among the three alternative systems: Stereo Dipole, Ambisonics and Ambiophonics.

The listening environment was a very damped listening room, equipped with a pair of Quested F-11 self-powered monitors employed for the Stereo Dipole and 8 General Music self-powered monitors for the 8-ways Ambisonics cubic rig. Fig. 30 shows a photograph of the listening room, which is actually capable of providing optimal listening conditions for just one listener standing in the center.

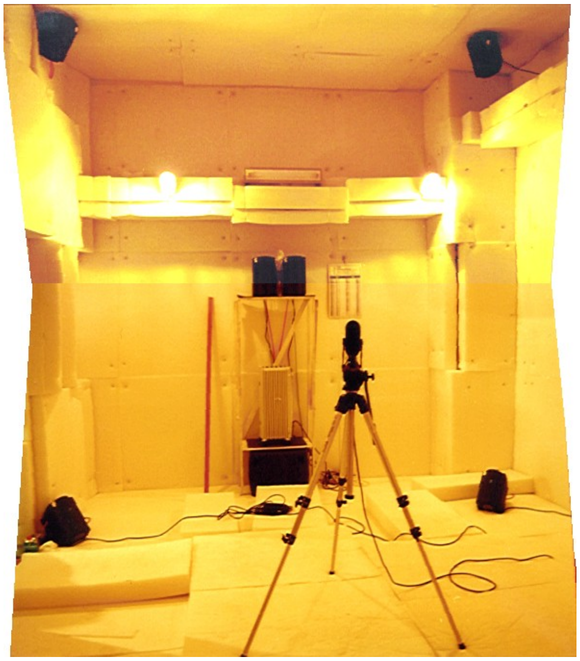


Fig. 30 – Ambiophonics Listening room

A panel of 9 students was employed for the comparative tests. Each subject was asked to rank, in order of preference, three repetitions of the same music sample, 40s long. One of them was presented through Stereo Dipole only, one through the Ambisonics rig only, and the third one employing simultaneously both systems, that is by Ambiophonics. It must be noted that all the signals were pre-computed off-line by means of the Aurora plugins [15], and CoolEditPro was employed for playing the 10-channels session through a multichannel sound board (Echo Layla). The first two types of signals (Stereo Dipole and Ambisonics) were obtained by muting the other channels, whilst the third (Ambiophonics) was played with all 10 channels unmuted.

The above methodology is somewhat misleading regarding Ambisonics, because in practice the impulse responses employed for convolution were deprived of the direct sound, and this caused the reproduction of the sound through the Ambisonics rig to be excessively diffuse, with poor localization of the sound coming from the frontal stage. On the other hand, a brief informal test conducted employing IRs containing also the direct sound shown that the difference was not very evident, and thus it was considered not be worth the increased complexity required.

Each subject had to rank 3 music pieces, each one processed with a set of IRs corresponding to a different theatre, coupled with a different set of cross-talk cancelling filters.

The following music samples were employed:

Music piece	Theatre	Cross-talk filters
Mozart, Te Deum K141, Sennheiser MKE2002 ("Mozart Sacro", n. 1)	La Scala	Binaural
Buxtehude KFM-6 (Ambiopole demo 1, n.13)	Teatro Comunale	Sphere
Mozart, Overture "Le nozze di Figaro", bars 1-50, ORTF (Denon PG 6006, n. 37)	Teatro Verdi	ORTF

This, of course, did not explore all possible combinations; furthermore, the results are certainly dependent on the source material, and it was not checked if the cross-talk filters were really the optimal ones for each recording (the coupling was based only on "a priori" knowledge of the miking technique employed for each recording).

It is planned to take the occasion of the public demonstration of the three systems planned at the 19th AES Conference for collecting a large number of qualified subjective listening tests.

Nevertheless, the results of these first low-quality subjective tests are encouraging. The following table reports the average scores obtained by the three methods, obtained assigning score 1, 2 and 3 respectively to the sound samples in their ranked order.

Method	Stereo Dipole	Ambisonics	Ambiophonics
Avg. Score	1.99	2.77	1.24

A multi-dimensional statistical test applied to the raw subjective responses demonstrated that the ranking is statistically significant (the 95% confidence interval resulted equal to 0.62, and thus both differences between average scores are significant).

Of course, the limited number of subjects and sound samples for each subject, and the fact that they were not trained, nor selected for their discriminative listening capabilities, means that the above results must be considered just a very initial confirmation of the validity of the Ambiophonics method.

CONCLUSIONS

This paper presented the theoretical background and the practical implementation of the Ambiophonics surround system. The system can be seen as the superposition of two already established surround techniques (Stereo Dipole and Ambisonics): each of them is employed for what it does best.

The superposition of the two cooperating systems produces significant advantages, which are clearly outlined by the theoretical analysis, and were confirmed by listening tests.

Although Ambiophonics until now has been a quite expensive system suited only for audiophiles in the high-end, its modern implementation in the form of freeware software for low-cost PCs opens its use to the vast majority of music lovers, provided that they take care of allocating a suitable, well treated listening environment.

ACKNOWLEDGEMENTS

David Johnston, author of CoolEditPro [10], is acknowledged for his excellent software, which was kindly made available free for this research.

This work was partially supported through a research convention between ASK Industries, Reggio Emilia, Italy and the University of Parma, co-funded by the Italian Ministry for University and Research (MURST) under the grant MURST-98 #9809323883-007.

The substantial support and liberal contribution of the Ambiophonics Institute was fundamental for the research work described here.

REFERENCES

[1] H. Moller – "Fundamentals of Binaural Technology" – *Applied Acoustics* vol. 36 (1992) pp. 171-218.

[2] M. Gerzon, "Ambisonics in Multichannel Broadcasting and Video" - *Journal of Audio Engineering Society*, Vol. 33, Number 11 pp. 859 (1985).

[3] A. Farina, E. Ugolotti, "Subjective comparison between Stereo Dipole and 3D Ambisonics surround systems for automotive applications", *16th AES Conference*, Rovaniemi (Finland) 12-14 April 1999.

[4] D. de Vries, J. Baan, "Auralization of Sound Fields by Wave Field Synthesis" - *Pre-prints of the 106th Convention*, Munich, Germany, 1999 May 8–11, # 4927.

[5] O. Kirkeby, P. A. Nelson, H. Hamada, “The "Stereo Dipole"-A Virtual Source Imaging System Using Two Closely Spaced Loudspeakers” – *JAES* vol. 46, n. 5, 1998 May, pp. 387-395.

[6] 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.

[7] A. Farina, F. Righini, ‘Software implementation of an MLS analyzer, with tools for convolution, auralization and inverse filtering’, *Pre-prints of the 103rd AES Convention*, New York, 26-29 September 1997.

[8] S.T. Neely, J.B. Allen, ‘Invertibility of a room impulse response’, *J.A.S.A.*, vol.66, pp.165-169 (1979).

[9] J.N. Mourjopoulos, “Digital Equalization of Room Acoustics”, *JAES* vol. 42, n. 11, 1994 November, pp. 884-900.

[10] D. Johnston – Cool Edit Pro v.1.2 – *Syntrillium Software*, HTTP://www.syntrillium.com, 2000.

[11] J. O. Smith and J. Abel, “The Bark Bilinear Transform”, *Proc. of IEEE ASSP Workshop*, 1995, Mohonk, New Platz, NY.

[12] M. Karjalainen, E. Piirilä and A. Järvinen, “Loudspeaker Response Equalisation Using Warped Digital Filters”, *Proc. of NorSig-96*, September 1996, Espoo, Finland, pp. 367-370.

[13] A. Bellini, E. Armelloni, G. Cibelli, E. Ugolotti, A. Farina – “Experimental validation of equalizing filter for car cockpits designed with Warping technique” - *109th AES Audio Convention*, Los Angeles, 18-22 September 2000.

[14] A. Torger, “Software Equaliser and Tools for High Resolution Digital Audio” - *NWFIIR AudioTools* http://www.ludd.luth.se/~torger/filter.html

[15] A. Farina, “Software auralization on a standard multimedia PC” – *Aurora software home page*, http://www.ramsete.com/aurora/

[16] R.R. Sokal, and F.J. Rohlf. “Biometry: The Principles and Practice of Statistics in Biological Research” - 2nd ed. New York: *W. H. Freeman*, 1995.

[17] R. Glasgal, K. Yates, “Ambiophonics - Beyond Surround Sound to Virtual Sonic Reality”, *Ambiophonics Institute*, 1995. HTTP://www.ambiophonics.org

[18] M.A. Gerzon, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound", *Preprints of the 50th Audio Engineering Society Convention*, London (1975 March).

[19] G. Schiffrer, D.Stanzial, ”Energetic Properties of Acoustic Fields”, *J. Acoust. Soc. Am.* 96, pp. 3645-3653, 1994.

[20] D. Stanzial, N. Prodi, G. Schiffrer, "Reactive intensity for general fields and energy polarization" - *J. Acoust. Soc. Am.*, 99(4): 1868-1876, April 1996.

[21] A. Farina, E. Ugolotti, “Software Implementation Of B-Format Encoding And Decoding”, *Pre-prints of the 104rd AES Convention*, Amsterdam, 15 - 20 May, 1998.

[22] M.A. Gerzon, "Recording Concert Hall Acoustics for Posterity", *J. Audio Eng. Soc.*, vol. 23, pp. 569, 571 (1975 Sept.).

[23] “Carta di Ferrara”, CIARM, http://acustica.ing.unife.it/ciarm/Carta.htm

[24] "Guidelines for acoustical measurements inside historical opera houses: procedures and validation", *CIARM*, http://acustica.ing.unife.it/ciarm/download.htm

[25] A. Farina, L. Tronchin, “3D Impulse Response measurements on S.Maria del Fiore Church, Florence, Italy” – *Proc. of ICA98*, International Conference on Acoustics, Seattle (WA) 20-26 June 1998.

[26] J. Daniel, J.B. Rault, J.D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions" - *Pre-prints of the 105th AES Convention*, S.Francisco, 26 - 29 September, 1998.

[27] H. Okubo, M. Otani, R. Ikezawa, S. Komiyama, K. Nakabayashi, “A system for measuring the directional room acoustical parameters”, *Applied Acoustics* vol. 62, pp. 203±215, Elsevier, 2001.

[28] J.J. Lopez, A. Gonzalez, “PC Based Real-Time Multichannel Convolver for Ambiophonic Reproduction”, *19th AES Conference on Surround Sound*, Schloss Elmau, Germany, 21-24 June 2001.

[29] A.V.Oppheneim, R. Schafer, “Digital Signal Processing”, *Prentice Hall*, Englewood Cliffs, NJ 1975, p. 242.

[30] A.Torger – “BruteFIR”, http://www.ludd.luth.se/~torger/brutefir.html

[31] T. G. Stockham Jr., “High-speed convolution and correlation”, *AFIPS Proc. 1966 Spring Joint Computer Conf.*, Vol 28, Spartan Books, 1966, pp. 229 - 233.

[32] J. S. Soo, K. K. Pang, “Multidelay block frequency adaptive filter”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-38, No. 2, February 1990.

[33] W.G. Gardner, “Efficient convolution without input-output delay”, *JAES* vol. 43, n. 3, 1995 March, pp. 127-136.