SUBJECTIVE COMPARISON BETWEEN STEREO DIPOLE AND 3D AMBISONIC SURROUND SYSTEMS FOR AUTOMOTIVE APPLICATIONS

ANGELO FARINA[1] AND EMANUELE UGOLOTTI[2]

[1]Industrial Engineering Dept., Università di Parma, Via delle Scienze - 43100 Parma - ITALY
tel. +39 0521 905854 - fax +39 0521 905705 - E-MAIL: farina@pcfarina.eng.unipr.it
HTTP://pcfarina.eng.unipr.it

[2]ASK Automotive Industries, via Fratelli Cervi n. 79, 42100 Reggio Emilia - Italy
tel. +39 0522 388311 - fax. +39 0522 388499 - E-MAIL: UgolottiE@askgroup.it

The paper describes the results of a subjective evaluation experiment: two methods for recording a three-dimensional sound field and for reproducing it through loudspeakers in a proper listening room are compared.
The first method is the binaural method known as Stereo Dipole, based on digital filtering of binaural recordings or binaurally synthesized sound tracks: it allows for reproduction over a pair of closely-located loudspeakers. The second method is a software implementation of the well-known Ambisonics methodology, in which a B-format recording made with a Soundfield mic, or a synthesized B-format soundtrack, is reproduced over a 3D array of 8 loudspeakers.
The subjective comparisons were made in a listening room fitted with 10 loudspeakers, and the listeners did not know at what of the two systems they were listening. Both reproduction systems were employed for blind evaluation of the sound field generated by audio systems of different cars.

INTRODUCTION

In the automotive field, sound systems are mainly compared by direct listening tests, as objective measurements hardly supply enough information regarding the sound quality. Nevertheless, when the listener moves from one car to another, his judgement is usually biased by not-acoustical effects, such as confort of the seats and knowledge of the brand and cost of the car.
Blind subjective tests can be done only having the listeners seating in a neutral room, and presenting to them recordings or auralisations coming from the different cars. Since now these comparative tests were made with the binaural technology, placing a dummy head inside the car, and recording directly the sound inside it [1]. Alternatively, the same technique was implemented by first measuring the binaural impulse responses of the sound system inside the car, and then convolving them with various music samples [2].
In both cases, anyway, the listeners had to wear headphones, and this causes some well-known defects: over-sensitivity to the background noise, front-back confusion, in-head localization, rotation of the sound field when the listener's head is moved, etc. . Most of these defects are removed if the listening tests are conducted through loudspeakers, but usually this does not allow a proper reconstruction of the spatial effects ("surround"), which are very important for the evaluation of the quality of the sound inside a car.
But now at least two realistic, three-dimensional surround reproduction systems are available: the Stereo Dipole [3,4,5] and the Ambisonics [6,7,8] methods.

The first is a derivation of the binaural technology, in which the two channels are digitally filtered prior of being sent to a couple of loudspeakers, placed at a little angular distance (typically 10 degrees). This way, the cross-talk cancellation implemented in the digital filters is much more effective, the colouring of the particular dummy head employed for the recordings inside the car is almost completely removed, and the listener enjoys a reasonably wide and robust "sweet spot".
The second technique was invented more than 20 years ago, but in the past it was implemented just as a two-dimensional surround system, encoded in two channels only (UHJ format). This allowed for a planar reproduction over 4 loudspeaker, which was certainly superior to the Quad system of those days, but really not convincing enough for absolute quality listening tests. Nowaday Ambisonics can enjoy a second youth, as digital multichannel systems are very cheap, and a standard PC has computing power well in advance for decoding a complete, three-dimensional, 4-channels B-format signal, directly recorded from a Soundfield microphone.
For these reasons, it was decided to set up a specially designed listening room, equipped with both systems. A comparative subjective experiment was started, for evaluating what of the two system was better for a large screening tests over dozens of different cars.
This paper reports about the technical problems encountered during the setting-up of the systems, and about the subjective comparisons made.
In more detail, the experimental apparatus is described, including the dummy head and Soundfield

microphones, the computer-based processing of the data, and the listening environment. An objective measurement of the transfer function of each element in the chain is presented.

A distinction is made between the Auralization technique, and the direct record/playback approach. The first is based on the measurement of proper impulse responses in the original environment of the sound system under test, thanks to a newly developed swept-sine excitation signal: the signal to be reproduced is then obtained by convolution of the original music samples with the set of impulse responses.

The second technique, which was actually employed for the subjective tests, is instead based on real-time multi-channel recording of the sound field produced in the original space (typically a car compartment, in our case) by the reproduction of the original music samples through the sound system under test.

In the following, the theory of both approaches is described in detail (for both the Stereo Dipole and the Ambisonics setup), and the advantages and disadvantages of the Auralization technique are pointed out. Finally, the first subjective results obtained in the case of the direct record/playback approach are presented.

1. AURALIZATION VS. RECORD/PLAYBACK

Nowadays, surround techniques can be broadly categorised in three ways:

a) Auralization systems
b) Recording/playback systems
c) Synthesis/spatialization systems

The third category refers to those systems where the accent is on the final reproduction space, and with the goal of making real or virtual sources to be localised all around that space: in this case, there is not an "original sound space" to be reproduced. It is well known that actually these system are the more accurate in creating the illusion of phantom sources or moving sources, as the natural acoustics of the reproducing space can be made to cooperate with the illusion, and each "virtual" sound source can be processed separately from the others.

In this study, anyway, we concentrate on systems capable of reproducing the sound field of an "original" sound space inside another, different room (the "reproduction" space). The second has usually to be almost completely anechoic, for avoiding that its acoustic behaviour superposes to the one of the "original" space.

At this point, we distinguish between Auralization systems, in which the spatial behaviour of the original sound space is first sampled in the form of a set of impulse response, and direct record/playback system, in which the spatial information is sampled together with the music signal being recorded.

The Auralization systems are based on the linearity hypothesis, and are particularly efficient for the reproduction of original sound spaces in which the sound field is generated by electroacoustic devices (loudspeakers) in fixed positions. In this case, for a fixed listening position, only a very little number of Impulse Responses has to be measured for each sound source (2 for the Stereo Dipole or other binaural approaches, 4 for the Ambisonics/B-format approach). After measuring the impulse responses, the sound field can be reconstructed by convolving any kind of original signal with the impulse responses: this convolution can nowadays be done quickly and cheaply by proper software tools.

On the other hand, when many "natural" sound sources have to be recorded, possibly moving around the listening point, the direct recording of the sound event is more straightforward, although, as it will be shown, this limits the capability of spatial reproduction in comparison with the Auralization technique.

Fig. 1 compares the two approaches in the case of the Stereo Dipole method, fig. 2 illustrates the same comparison for the 3D Ambisonics method.

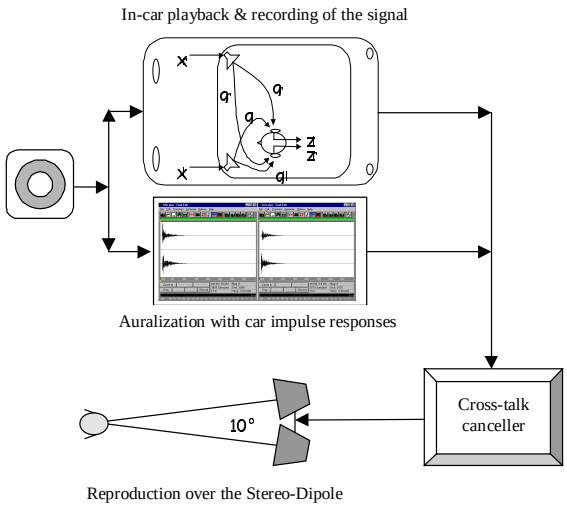

Fig. 1 - Stereo-Dipole reproduction of binaural recordings or binaural convolutions

For car acoustic applications, both methods can be employed in principle: a little number of electro-acoustic loudspeakers are placed in fixed positions, and also the listeners are almost completely fixed, so that the Auralization approach is appealing; furthermore, this approach can be implemented also starting from numerical simulations of the sound field [9], instead of experimentally-measured IRs.

But in many cases the behaviour of the loudspeakers is markedly not-linear [10], and the background noise is an important factor: thus the direct recording of what happens is much more realistic, particularly if the goal

is to reproduce the actual listening conditions during a road drive.

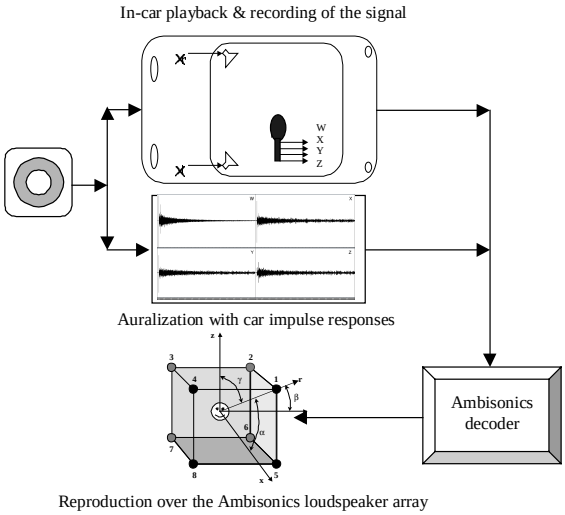

Fig. 2 - 3D-Ambisonics reproduction of B-format recordings or B-format convolutions.

In practice, the employment of the direct recording/playback methods requires a longer time when making recordings inside the original space (because it is necessary to play inside it all the sound samples which have to be evaluated), and consequently multichannel recording capability and a lot of storing space are required. A comparison between the results obtained with Auralization and direct recording is beyond the scope of this paper, although it will be a necessary future investigation.

For the first subjective tests described here, only the direct recording/playback method was employed: nevertheless, the theoretical parts describe also the processing of both surround methods (Stereo Dipole and Ambisonics) in terms of impulse responses, as this approach can in principle produce superior surround reproduction.

2. THE STEREO DIPOLE METHOD

With "Stereo Dipole" we refer to a recording/reproduction method based on the traditional binaural approach (two-channels dummy head recordings or IR measurements), in which the playback of the binaural soundtracks happens on a pair of closely-spaced loudspeakers, as shown in fig. 1. For this technique to work effectively, three points are important:

1) A good binaural microphone fitted in a realistic dummy head (or worn by a human) – the presence of shoulders and torso are very important for automotive applications!
2) An almost anechoic reproducing room, fitted with a pair of very high quality, spatially-coherent loudspeakers, properly placed in the optimal position
3) The design of proper numerical inverse filters for performing the required cross-talk cancellation.

2.1 Inverse cross-talk filters

The first two point are simply matter of choosing the proper tools and setting up them properly, which takes a lot of time and money, but is of little scientific relevance. In the next chapters, anyway, a detailed description of the hardware employed will be given.

The third point requires instead a proper theoretical explanation. The approach employed here is derived from the formulation originally developed by Kirkeby and Nelson [4,5]. The following fig. 3 shows the cross-talk phenomenon in the reproduction space:

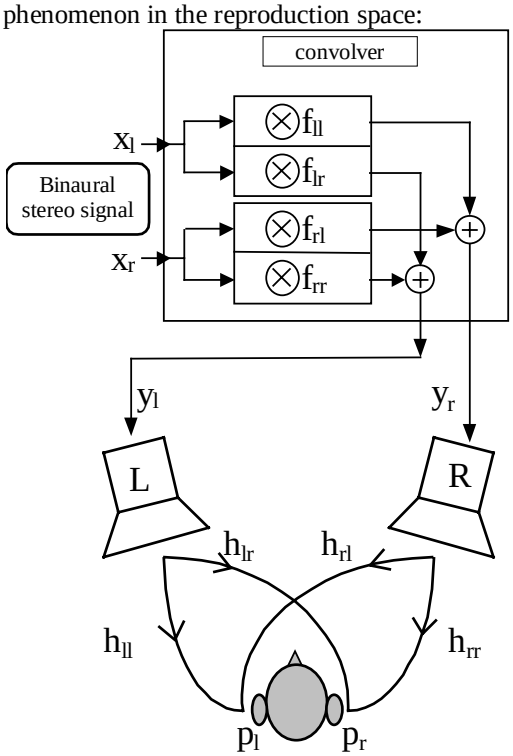

Fig. 3 – cross-talk cancelling scheme

The 4 cross-talk cancelling filters *f*, which are convolved with the original binaural material, have to be designed so that the signal 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:

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

The problem is the computation of the InvFilter (denominator), as its argument is generally a mixed-phase function. In the past, the authors attempted [11] to

perform such an inversion employing the approximate methods suggested by Neely&Allen [12] and Mourjopoulos [13], but now the Kirkeby-Nelson frequency-domain regularization method is preferentially employed, due to its speed and robustness. A further adaptation over the previously published work [14] consists in the adoption of a frequency-dependent regularisation parameter. In practice, the denominator is directly computed in the frequency domain, where the convolutions are simply multiplications, with the following formula:

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

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

$$InvDen(\omega) = \frac{Conj[C(\omega)]}{Conj[C(\omega)] \cdot C(\omega) + \varepsilon(\omega)} \quad (3)$$

In practice, $\varepsilon(\omega)$ 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's interface of the software developed for computing the cross-talk cancelling filters:

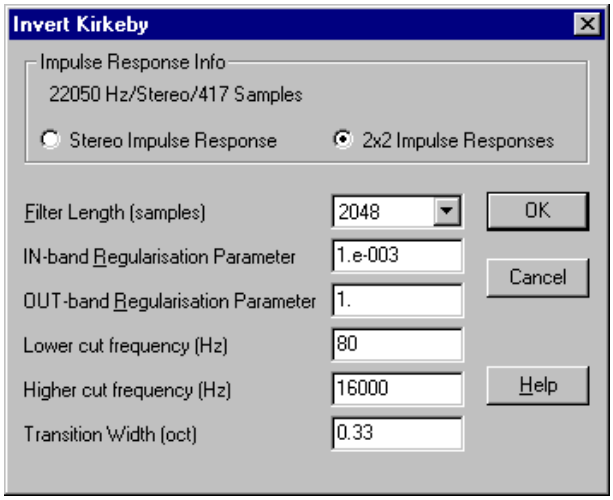

Fig. 4 – user's interface of the inverse filter module

This software tool was implemented as a CoolEdit plug-in, and it can process directly a stereo impulse response (assuming a symmetrical setup, so that $h_{ll}=h_{rr}$ and $h_{lr}=h_{rl}$), or a complete 2x2 impulse responses set, obtained placing first the binaural IR coming from the left loudspeaker, followed in time by the binaural IR coming from the right loudspeaker. In both cases, the outputted 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 an error-and-trial method.

2.2 Convolution

After the production of the inverse filter set, they have to be continuously convolved with the binaural signal being reproduced. This can be done, in real time, with the frequency-domain software implementation already developed [11,14], but now a very cheap DSP board has been programmed for such task: it is an evaluation board produced by Analog Devices, called EZ-kit, equipped with the new SHARC processor. The software can be downloaded to the board through an RS-232 port, after which it can be run outside the control of the computer.
The CoolEdit plugin, after computing the inverse filters, checks for the presence of the EZ-kit board on the serial port COM1, and if found downloads directly on it the inverse filters, along with the convolution software. The actual implementation of the SHARC convolver is in time-domain, which limits the length of the inverse filters to 400 points; a new frequency-domain version is under development, and it will be demonstrated during the 16th AES Conference. Nevertheless, 400 points are enough for Stereo-Dipole filters over a pair of remarkably flat loudspeakers in an anechoic environment, if the dummy head employed has not too much resonating effects in its pinnae and ear channels.

2.3 Improvements in case of auralization

Up to this point, there is no difference between the direct recording/playback implementation and the Auralization, as both produce a binaural signal, which can be stereo-dipoled with the above procedure. In practice, anyway, it is possible to implement a more sophisticated system when the Auralization approach is performed. Let us consider that usually the original music source is stereo, and it is played in the original space over a stereo system (although this can consist of several distinct loudspeakers). So, the binaural signal to be reproduced (x_l, x_r) can be thought as the convolution of the original stereo sound track (s_l, s_r) with the impulse responses of the original sound space g (usually measured inside the car compartment, as described in [15]):

$$\begin{aligned} x_l &= s_l \otimes g_{ll} + s_r \otimes g_{rl} \\ x_r &= s_l \otimes g_{lr} + s_r \otimes g_{rr} \end{aligned} \quad (4)$$

This suggests that the convolution of the original signal s with the filters g can be combined with the subsequent convolution with the inverse filters f. In practice, a new set of filters f' is obtained, which transforms directly the original signals s in the speaker feeds y for the stereo dipole:

$$\begin{cases} f_{ll}' = (g_{ll} \otimes h_{rr} - g_{lr} \otimes h_{rl}) \otimes InvDen \\ f_{lr}' = (g_{lr} \otimes h_{ll} - g_{ll} \otimes h_{lr}) \otimes InvDen \\ f_{rl}' = (g_{rl} \otimes h_{rr} - g_{rr} \otimes h_{rl}) \otimes InvDen \\ f_{rr}' = (g_{rr} \otimes h_{ll} - g_{rl} \otimes h_{lr}) \otimes InvDen \end{cases} \quad (5)$$

Substituting these new filters in the convolver enables a faster computation, but substantially the results remain the same. In practice, a correctly implemented Stereo Dipole can render effectively phantom source positions located anywhere in the frontal hemispace of the listener, but it is difficult to recreate virtual sources in the rear hemispace. In the case of car sound systems, usually two separate measurements are taken, one with the frontal loudspeaker system, the other feeding only the rear loudspeakers [15]. In such a case, it is possible to render separately the auralized results over two stereo dipoles, one located in the front of the listener, the second located behind him: the first is fed with the original signal, processed through the inverse filters h'_f computed with the frontal g_f and h_f IRs, the second is fed with the same original signal s, but processed through a second set of inverse filters h'_r, also computed with (5), but starting from the g_r and h_r IRs measured with the rear loudspeakers inside the car and the rear stereo dipole in the reproduction room. Only a preliminary, informal test was made on this dual-stereo-dipole system, and thus it is not yet clear if the improvement is worth the additional effort.

3 THE 3D-AMBISONICS METHOD

Ambisonics is a quite old technique [6], which was developed mainly for recording/playback and broadcasting. Its recent application for auralization opened new frontiers [7,8,16], but some basic misunderstandings still make it difficult to implement it for general-purpose surround applications. In particular, the weak points which will be addressed here are the following:

1) Confuse, unnecessarily complex theoretical formulation of both the recording/measuring side ("encoding") and particularly of the playback/rendering side ("decoding")
2) Lack of connection with the general acoustics formulation, particularly with the Sound Intensity formulation: this makes it difficult to understand why the system works
3) The 2D suboptimal implementation known as UHJ destrojed the reputation of the method. No complete 3D recordings (B-format) of musical events are available in the discography.
4) No clear separation between the three possible application fields discussed in paragraph 1) was ever made, so that the system is usually considered more a recording technique than a processing tool.

In the original formulation, a direct recording is made in the original sound space with a special microphone probe (Soundfield microphone), which samples the omnidirectional sound pressure (W) together with the three cartesian components of the air particle velocity (X, Y and Z). Please, note that the fact that the X,Y,Z signals are particle velocity components is a quite recent acquisition, in the past they were called "pressure gradients" or "figure-of-eight microphones", and the phase mismatch between pressure and particle velocity, which always happens in reactive sound fields, was completely neglected.

At the playback installation, these 4 signals are combined together in a quite simple additive matrix, deriving proper speaker feeds which are fed to a regular array of transducers, surrounding the listener: a regular cube is the simplest, straightforward reproduction array, requiring 8 identical loudspeakers. This configuration is illustrated in fig. 2. If the loudspeaker array is a perfect cube, the speaker feeds are computed as follows:

$$\begin{aligned} F_1 &= G_1 \cdot W + G_2 \cdot (+X + Y + Z) \\ F_2 &= G_1 \cdot W + G_2 \cdot (-X + Y + Z) \\ F_3 &= G_1 \cdot W + G_2 \cdot (-X - Y + Z) \\ F_4 &= G_1 \cdot W + G_2 \cdot (+X - Y + Z) \\ F_5 &= G_1 \cdot W + G_2 \cdot (+X + Y - Z) \\ F_6 &= G_1 \cdot W + G_2 \cdot (-X + Y - Z) \\ F_7 &= G_1 \cdot W + G_2 \cdot (-X - Y - Z) \\ F_8 &= G_1 \cdot W + G_2 \cdot (+X - Y - Z) \end{aligned} \quad (6)$$

In principle, the two gains G_1 and G_2 should assume different values for low and high frequency, with a smooth transition between the two formulations around 400-700 Hz. Various authors proposed different "optimal" values for the two gains, as reported in [7,8]. In this case, the simple "in-phase", frequency-independent formulation was employed, which yields:

$$G_1 = 1 \qquad G_2 = \frac{1}{\sqrt{6}} = 0.40825 \quad (7)$$

These values were obtained imposing that a single virtual source, located exactly in a corner of the cube, produces a null signal at the loudspeaker located in the opposite vertex.

3.1 Understanding the traditional formulation

It was already stated that actually the X,Y,Z signals produced by the Soundfield microphone are actually particle velocity components, and that these can exhibit a significant phase mismatch from the pressure W signal. This happens in reactive fields, that is at little distance of a point source (less than two wavelengths) or in presence of reflections (which is always the case in reverberant rooms, and also inside car compartments).

So the above formulas (6) intrinsically assume that the sound field is produced by the superposition of plane, progressive waves, with no curvature, in free field.

Furthermore, also in the reproduction space the same assumptions must hold, which is true only if the room is almost anechoic, and if the loudspeakers are far enough from the listener.
In practice, all the above assumptions are usually not met, and this causes the traditional Ambisonics method to produce less-than-optimal results: very often the pressure signal is wider than the vector sum of the velocity components, and this causes a lot of "common sound" to be radiated by all the loudspeakers of the array, and only very limited amplitude differences are introduced by the presence of a well localised sound source. In this sense, the localisation effectiveness of the Ambisonics system is quite weak, although obviously the "envelopment" effect caused by the "common sound" is always present. Only outdoor recordings of far sources (such as airplanes or trains) produce realistic playback in wide, anechoic listening rooms. The above problems can be alleviated increasing the gain G_2 in comparison with G_1, but this can cause the appearence of 180-degree-out-of-phase signal coming from the loudspeakers placed in the opposite direction of the virtual sound source, and this reduces substantially the spatial size of the area where a correct reproduction is obtained.
Furthermore, combining out-of-phase pressure and velocity signals can introduce audible artifacts, such as comb filtering and coloration.
What must be understood is that traditional Ambisonics is still a level-panning technique, as no importance is given to phase mismatch between pressure and velocity channels, and the reproduction formulas are derived in complete ignorance of the modern sound intensity theory [17].

3.2 Proposal of a modern re-implementation of Ambisonics

What follows requires that the reader has some basic knowledge of the original Sound Intensity theory [18], and of its modern re-formulation [17]. Let we consider first only steady sound fields, produced by steady sources, so that we can take time averages for any quantity. If we analyse what happens at the listening point in the original sound field, we find that some amount of energy is flowing along a particular direction, but that there is also a lot of acoustic energy which is not propagating at all, but is simply bouncing around in the environment. The first net energy flow can be measured with a 3D Sound Intensity probe, and gives us the Active Intensity (AI) vector, measured in W/m^2; the not-propagating energy cannot be measured in the same way (the so-called "reactive intensity" is really a mathematical artifact), but it can be estimated by the overall energy density (D), measured in J/m^3. D can be computed by the RMS-averaged sound pressure $\overline{p_{RMS}}$ and particle velocity $\overline{u_{RMS}}$:

$$\overline{D} = \frac{1}{2} \cdot \left[\frac{\overline{p_{RMS}}^2}{\rho \cdot c^2} + \rho \cdot \overrightarrow{u_{RMS}}^2 \right] \tag{8}$$

Also the active intensity can be computed from the pressure and particle velocity signals, but in this case a linear time average (not RMS) of their product has to be taken:

$$\overrightarrow{AI} = \overline{p \cdot \vec{u}} \tag{9}$$

It must be noted that D is a scalar, while AI is a vector, having the same direction as the $\overrightarrow{u_{RMS}}$ vector.
As the propagating energy AI is contributing to the total energy density D, the not-propagating energy E can be obtained by difference:

$$\overline{E} = \overline{D} - \left| \frac{\overrightarrow{AI}}{c} \right| \tag{10}$$

Another important quantity is the ratio between active intensity and sound density, which has the dimensions of a velocity (Stanzial calls it "sound energy speed vector"). Here its dimensionless ratio with the speed of sound c is considered, called Propagation Index Vector β:

$$\vec{\beta} = \frac{\overrightarrow{AI}}{\overline{D} \cdot c} \tag{11}$$

It is a vector having the same direction as the RMS particle velocity, with a modulus bounded within 0 (no energy propagation, diffuse field) and 1 (plane, progressive wave without any wavefront curvature and reflections). β gives a simple description of the nature of the sound field (active or reactive), although a more detailed analysis [19] requires that also the not-propagating energy is spatially analysed, deriving its three Cartesian polarisation components.
In the original Ambisonics formulation, two vector quantities quite similar to β were defined: the Makita velocity vector r_V and the energy localisation vector r_E. These quantities always refer to the reproduction space, no reference is given to their values in the original recording space: the first one is defined as the vector sum of the particle velocities produced by all the loudspeakers divided by the algebraic sum of the sound pressures produced by all the loudspeakers, normalized to 1 by multiplying for the air impedance $\rho \cdot c$:

$$r_V = \frac{\sum_{i=1}^{N} \overrightarrow{u_{RMS,i}}}{\sum_{i=1}^{N} \overline{p_{RMS,i}}} \cdot \rho \cdot c \tag{12}$$

Please note that in (12) all quantities are assumed real, the phase is not taken into account, apart for the fact that

the feed of some speakers can be "negative", and this makes its velocity vector to change versus, and its RMS pressure to become negative.
The energy localisation vector is defined in the same way, but summing the square of the velocities and of the pressures:

$$r_E = \frac{\sum_{i=1}^{N} \overrightarrow{u_{RMS,i}}^2}{\sum_{i=1}^{N} \overline{p_{RMS,i}}^2} \cdot (\rho \cdot c)^2 \qquad (13)$$

The first ratio is considered relevant for low frequency localisation, as at low frequency the waves produced by the various loudspeakers combine in modulus and phase, whilst at high frequency the second ratio is considered relevant, as the combination of the waveforms, which are considered mutually incoherent, happens in an RMS way.
Both these extremisations appears simplicistic under the light of the modern sound intensity theory, as now we have the β vector, which holds in the whole frequency range, and has much more physical meaning. Under the very restrictive conditions stated in paragraph 3, however, the energy vector r_V equals the Propagation Index Vector β.
In any case, an ideal surround system should be capable of reproducing a single plane, propagating wave coming from any direction with an unity value of the relevant vector's modulus (r_V, r_E or β).
In practice, as the number of loudspeakers is limited, the vector's modulus is always less than unity. What is really important, however, is that the value of the modulus remain constant independently on the direction of the virtual sound wave.
So called pairwise amplitude panning, widely employed in the production of today's mixes, does not satisfy this basic requirement: when the virtual sound direction is coincident with a loudspeaker, all the other are muted, so that the vector's modulus approach unity, whilst when the apparent sound has to come midway between the loudspeakers, two (or three, in the 3D case) of them are active with the same gain, and thus the vector modulus is reduced. This produces the well-known "speaker detent" artefact, which causes the sound to appear coming mainly from the loudspeakers, and not from intermediate positions.
On the other side, the original Ambisonics approach produces an always constant value of the vector's modulus, but its value is significantly lower than the minimum one produced by the pairwise panning in its worst position, as almost all the loudspeakers are always radiating.
In a recent paper [8], it was shown how maximising the value of r_E significantly ameliorates the behaviour of a 2D-Ambisonics reproduction array. Following this reference, the ratio of gains G_2 and G_1 should be doubled at low frequency and increased by multiplying by $\sqrt{2}$ at high frequency, with respect to the values reported in paragraph 3. The effects of such a modification have to be subjectively evaluated yet.
An even more modern approach should feed the loudspeakers in such a way that the β value is always constant, but equal to the minimum value obtained with pair-wise (or triplet-wise, for 3D) level panning. This means that when the sound has to come directly from a direction corresponding to a loudspeaker, also the next ones (but only them) are fed with a reduced signal, so that the not-propagating energy E is increased a bit, maintaining β constant. In practice, the accuracy is limited only from the number of loudspeakers in the reproduction array.
Also this advanced approach has to be tested yet: it requires to analyse the 4 signals WXYZ with a 3D sound intensity software, to extract the short-time averaged quantities AI, D and β, and to employ only the W signal for feeding all the loudspeakers. Their gains have to be adjusted dynamically, so that the not-propagating energy E is reproduced by all the loudspeakers with the same gain, whilst the propagating energy AI is reproduced only by a minimum number of them: this way the value of β in the reproduction room is made to approach the β value in the original space. A further improved system could even manage the not-propagating energy E in a not-isotropic way, taking into account the energy polarisation along the three Cartesian axes, and feeding consequently the various loudspeakers so that also in the reproduction space the same polarisation ellipsoid is reproduced.
The above theory is already applicable to the synthesis of arbitrary sound fields (by designing proper panning laws, as it was did for the low-frequency formulation in [20] and for the high-frequencies in [7]), but requires a lot of further research for being applicable to the reproduction of recorded original sound fields produced by multiple, moving sources acting simultaneously, which are not easily resolved from the B-format mix.
In any case, however, this means that the B-format recordings contain the whole spatial information on the original sound field, and that the limited reproduction capability of actual first-order Ambisonics systems is due only to a quite rudimental decoding scheme: there is no need of developing higher order microphones, what is needed is a more intelligent decoding scheme (which could be called, in principle, "infinite-order").

3.3 Improvements in case of auralization

Also in this case, processing the B-format impulse responses instead of the direct recordings could produce significant advantages. As each set of impulse responses refers to a different source position, it is possible to compute just once the decoding coefficients for each

set. This ensures that separate sources will be properly reproduced with the highest possible spatial separation.
Furthermore, a B-format impulse response usually enables to identify and separate the direct wave and the first, discrete reflections. Begault already developed a mathematical technique capable of resolving even closely spaced or overlapped reflections [21]: this means that each single reflection can be considered as a separate (virtual) source, being separately decoded with proper coefficients. This process can be seen as "panning" correspondingly each reflection by means of the optimised panning laws suggested in the previous paragraph.
Bringing this process to the limit, it is possible to compute decoding impulse responses, which route the W signal of each source to the N loudspeakers in the reproduction array; each of them can be also convolved with an inverse filter, for the equalisation of each loudspeaker. This approach has to be tested yet, although the software tools for implementing it are already available [7].
It can be observed that the result of such a speaker-by-speaker convolution process are substantially similar to the surround part of the Ambiophonics system developed by Glasgal [22], although in that case the impulse responses are synthesised, instead of being computed from the ones measured in the original space. For completeness of information, it must be said that in Ambiophonics the surround impulse responses are without the direct wave, as the “stage sound” is reproduced through a separate Stereo Dipole. In a future development of this research, a similar approach will be attempted, feeding simultaneously the Stereo Dipole and the Ambisonics array.

4 HARDWARE IMPLEMENTATION

4.1 Listening room

A 6x5x4 meters room, completely covered with 100-mm-thick polyester fibre sheets, and mounted at saw-tooth, was equipped with 10 General Music 2-ways, self-powered monitors. 8 of them were mounted in the vertexes of a regular cube (4x4x4 m) the last two were placed in front of the listening position, in the Stereo Dipole configuration. For the comparative tests, only a single listener was allowed to seat exactly in the middle of the reproduction array, facing the stereo-dipole pair.
Fig. 5 shows a photograph taken inside the listening room: the Soundfield microphone is placed at the listening position; the Stereo Dipole and the 4 frontal loudspeakers are clearly visible.

Fig. 5 – Listening room

The following pictures show the impulse response and the frequency response produced by one of the loudspeakers inside the room. The sound field is almost anechoic, as the reverberation time is below 0.1 s in any octave band between 63 Hz and 16 kHz.

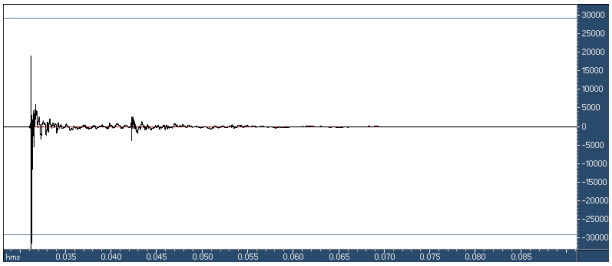

Fig. 6 - Impulse response inside the listening room

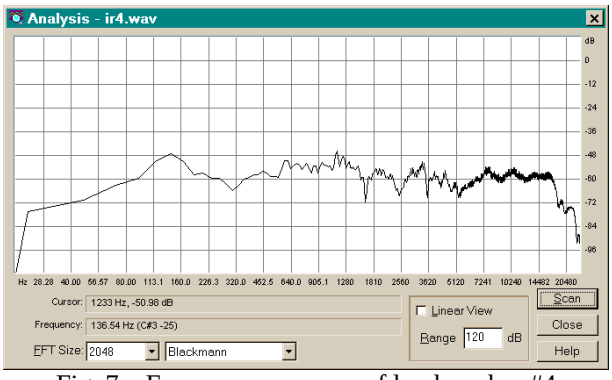

Fig. 7 – Frequency response of loudspeaker #4

4.2 Microphones

Two microphonic systems were employed for the experiment: a binaural dummy head (Ambassador) and

a B-format probe (Soundfield MKV). The following images show the microphones mounted on proper torso simulators, inside a car.

Fig. 8 – Ambassador binaural dummy head

Fig. 9 – Soundfield microphone

The frequency response of the Soundfield microphone is nominally flat, although it can be compensated together with the loudspeaker response, as the measurement of their frequency response was made employing the Soundfield itself: thus creating an inverse filter would compensate for both transducers. In this case, anyway, no equalisation was attempted on the Ambisonics chain.

Instead, the frequency response of the Ambassador dummy head is remarkably uneven, as demonstrated by fig. 10. It must be noted that this dummy head has internal microphones, placed at the end of a realistic ear channel, because it was developed for testing hearing aids. But, being the same dummy head employed also during the measurement of the h impulse responses produced by the Stereo Dipole loudspeakers, the overall frequency response of both loudspeakers and microphones is completely compensated for by the inverse filters f, computed as depicted in paragraph 2.1. For this reason, none of the equalization procedures discussed in [23] is required in this case.

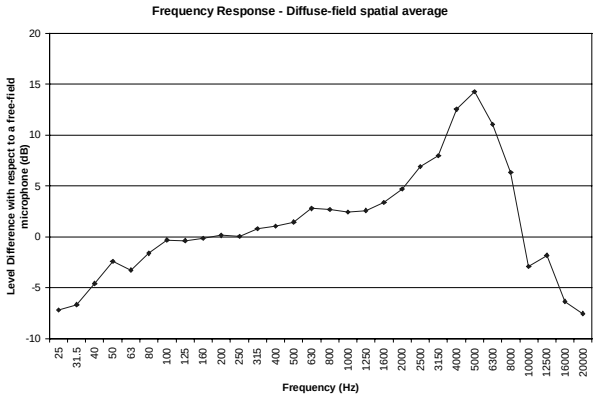

Fig. 10 - Frequency response of the Ambassador dummy head

4.3 Digital recording/processing equipment

All the processing of audio waveforms was done in digital domain, making use of a low-cost PC (Pentium II-400) fitted with an high-quality multi-channel sound board (Event Layla). The system is capable of simultaneous record & playback of 8 analog inputs and 10 analog outputs, with 20 bit resolution. The CoolEditPro multi-channel wave editor program was employed for all the tasks, as it acts as host program also for the specialised plug-ins developed for generating the test signals, for deconvolving the impulse response, for computing the Stereo Dipole filters and for convolving the original signals with them.

The Ambisonics processing was obtained simply mixing down the 4 input signals (WXYZ) with proper gains, without the need of additional software tools: a CoolEdit macro was recorded for automating the mix process.

The impulse response measurement inside the listening room was made with a new type of excitation signal, constituted by an exponentially-sweeping sine wave. A dedicated plug-in was developed, as shown in fig. 11, which generates the test signal and also pre-loads in the Windows clipboard the proper inverse filter: this is simply the time reversal of the excitation signal, with an amplitude shaped accordingly to the inverse of the spectral energy content of it. Fig. 12 illustrates a very short excitation signal and its inverse filter.

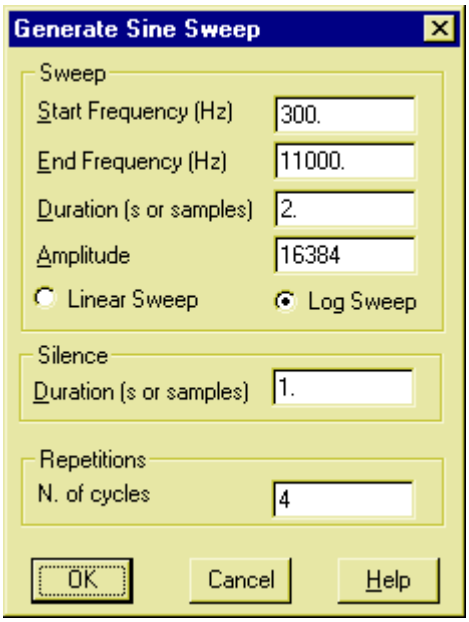

Fig. 11 – generation of sine sweeps

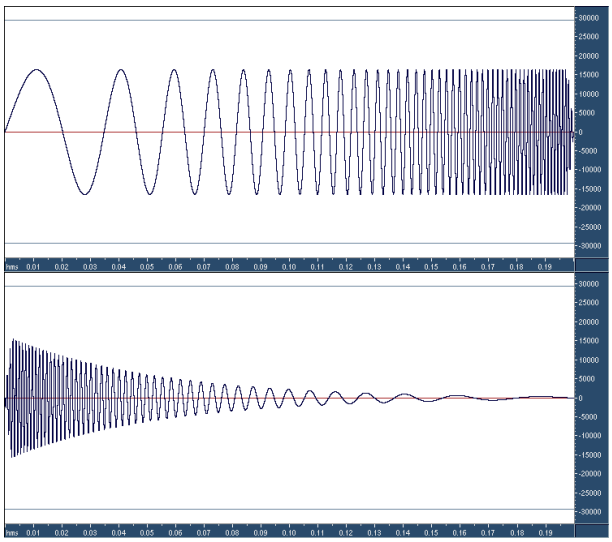

Fig. 12 – test signal (above) and inverse filter (below)

Thanks to the synchronous Rec/Play capabilities of CoolEditPro, the response of the system can be sampled simultaneously with the emission of the test signals: some repetitions are made, for ensuring that the system reached the steady state, and typically the response to the second or third repetition is analysed.

For recovering the system's impulse response, the inverse filter is convolved with the recorded system's response. This method revealed to be substantially superior to the Maximum Length Sequence (MLS) method previously employed [11,15]: the S/N ratio is better (by around 20 dB), and the measurement is almost immune from non-linearity and time variance. Furthermore, by properly setting the frequency limits for the sine sweep, it is avoided to damage the transducers applying too much energy outside their rated response limits.

5 SUBJECTIVE COMPARISONS

13 listeners had to fill-up a questionnaire when listening to each of 6 different sound samples. These were 3 pairs taken in three different cars. Each pair was constituted by the same music piece, recorded simultaneously with the dummy head (processed with the stereo dipole), and with the Soundfield microphone (processed through the 8 other loudspeakers). The listener did not know what of the two methods was in use for each sample, being the presentation order of them randomly shuffled for each listener. Usually, it resulted difficult to understand if the sound was coming from the 2 loudspeakers of the stereo dipole or from the 8 of the Ambisonics system, and this means that actually both systems were capable of relocating the sonic images far from the speakers: this is already a great result, as so-called "surround" systems actually being sold (5.1 systems for home theatre applications) usually produce sonic images only very close to the 5 main speakers.

The questionnaire was made of two sections: the first investigated the objective characteristics of the sound field (localization capability, robustness of the spatial effect, frequency response). The second section was about subjective quality items, such as listening fatigue, naturality, transparence.

The compilation of the questionnaires was completely automated, thanks to a specially-written computer program, which presented to the listeners the various sound samples (allowing to re-listen at will at any of them, in any order), and through which the responses were collected. Each question was presented as a couple of counter-posed attributes, and the listener had to place a mark between them, along a discrete scale with only 5 steps. Fig. 13 shows the subjective-testing software with the set of 6 questions.

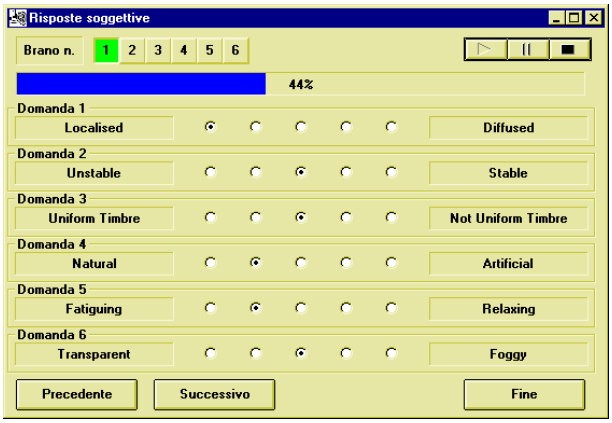

Fig. 13 - Software questionnaire

6 CONCLUSIONS

The collection of the subjective response was still uncompleted at the moment of writing, and thus the statistical analysis was limited and very simple.
The following table reports the average scores (and their Std. Deviation) obtained by the two systems:

Question	St.Dipole score	Ambisonics score
Localised(1) - Diffused(5)	2.1 (1.2)	3.2 (1.4)
Unstable(1) - Stable(5)	2.7 (1.6)	3.8 (1.8)
Unif.Timbre(1) - not unif. Timbre (5)	2.6 (1.0)	1.8 (1.9)
Natural(1) - Artificial(5)	3.2 (1.6)	2.4 (2.0)
Fatiguing(1) - Relaxing(5)	2.0 (1.4)	4.0 (1.8)
Transparent(1) - Foggy(5)	2.4 (1.8)	2.9 (2.1)

Nevertheless, the results of the first subjective responses has shown that the Stereo Dipole has limited capability of reproducing the low-frequencies, and makes the sweet spot very narrow at high frequency. Instead, the Ambisonics system has a much wider effective frequency range, and the sweet spot is always very large.
From the subjective point of view, it resulted that, although the Stereo Dipole gives superior spatial definition and localisation capability, the listening quickly becomes fatiguing, and this causes usually a judgement of lower naturality than the Ambisonics system. It must be remembered, anyway, that the original sound field, coming from a car sound system including the rear loudspeaker, was very confused and highly enveloping (this means a low value of the β coefficient). The Ambisonics array was able to recreate a very similar acoustic experience, although it is not capable of recreating soundfields with higher values of β.
In fact, the Ambisonics system revealed its limits in an informal test, in which the sound had to appear coming exactly from one of the 4 loudspeakers located on the floor, as the B-format recording was made directly inside the listening room, feeding anechoic speech only to the selected loudspeaker itself. Instead, during the Ambisonics reproduction a lot of sound was coming out from almost all the other loudspeakers, and the localisation was very weak, if not completely absent.
For these reasons, the conclusions reported above do not have to be considered absolute. It could happen that, when reproducing a very different sound field (such as a concert hall), the above subjective judgements change a lot, and the Stereo Dipole reveals superior.

Thus, the research will prosecute in two directions: the comparison between the two systems will be extended to other sound fields, in particular to opera houses [24]. The Ambisonics method, which actually seems superior for reproducing the car acoustics, will be employed for a comparative test between the same 9 cars already evaluated during the past year through the binaural (stereo dipole) technology.
In the meanwhile, all the cars which have to be subjected to measurements for further comparison, will be tested by playing the standard CD sample (which includes MLS signal, sine sweep and 5 different music pieces), and recording their response simultaneously with both the dummy head and the Soundfield microphone.

REFERENCES

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

[2] A. Farina, E. Ugolotti, “Subjective comparison of different car audio systems by the auralization technique”, Pre-prints of the 103rd AES Convention, New York, 26-29 September 1997.

[3] O. Kirkeby, P. A. Nelson, and H. Hamada – “Virtual Source Imaging Using the Stereo Dipole””, Pre-prints of the 103rd AES Convention, New York, 26-29 September 1997.

[4] 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.

[5] O. Kirkeby and P. A. Nelson – “Digital Filter Design for Virtual Source Imaging Systems”, Pre-prints of the 104th AES Convention, Amsterdam, 15 - 20 May, 1998.

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

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

[8] 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.

[9] A. Farina, E. Ugolotti - "Numerical model of the sound field inside cars for the creation of virtual

audible reconstructions"- DAFX-98 Conference, November 19-21, 1998 Barcelona, Spain.

[10] Bellini, G. Cibelli, E. Ugolotti, A. Farina, C. Morandi - "Non-linear digital audio processor for dedicated loudspeaker systems" - Proc. of International IEEE Conference on Consumer Electronics, Los Angeles, June 1998.

[11] 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.

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

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

[14] A. Farina, E. Ugolotti - "Spatial Equalization of sound systems in cars" - Proc. of 15th AES Conference "Audio, Acoustics & Small Spaces", Copenhagen, Denmark, 31/10-2/11 1998.

[15] A. Farina, E. Ugolotti, "Automatic Measurement System For Car Audio Application", *Pre-prints of the 104rd AES Convention*, Amsterdam, 15 - 20 May, 1998.

[16] J.M- Jot, S. Wardle - "Approaches to binaural synthesis" - *Pre-prints of the 105th AES Convention*, S.Francisco, 26 - 29 September, 1998.

[17] G. Schiffrer and D.Stanzial, "Energetic Properties of Acoustic Fields", J. Acoust. Soc. Am. 96, pp. 3645-3653, 1994.

[18] Fahy F.J. - *Sound intensity* - Elsevier Applied Science, London, 1989

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

[20] M. Poletti - "The design of encoding functions for stereophonic and polyphonic sound systems" - J.A.E.S., 44(11):948-963, November 1996.

[21] D. Begault, J. Abel - "Studying room acoustics using a monopole-dipole microphone array" - proc. of 16th ICA/ASA conf., Seattle, 20-26 June 1998, pp. 369.

[22] R. Glasgal - "Ambiophonics: The science of home music theater design" - HTTP://www.ambiophonics.org.

[23] V. Larcher, G. Vandernoot, J.M. Jot - "Equalization methods in binaural technology" - *Pre-prints of the 105th AES Convention*, S.Francisco, 26 - 29 September, 1998.

[24] P. Fausti, A. Farina, R. Pompoli - "Measurements in opera houses: comparison between different techniques and equipment" - Proc. of ICA98 - International Conference on Acoustics, Seattle (WA), 26-30 june 1998.

7 ACKNOWLEGDMENTS

This work was co-funded by the Italian University and Scientific Research Ministry (grant MURST-98 #9809323883-007) and by ASK Automotive Industries.

Substantial help in setting up the playback experiments came from colleagues at the University of Ferrara – Engineering Dept. (R.Pompoli, P.Fausti, N.Prodi).

Some suggestions and helpful hints came from O. Warusfel and colleagues at IRCAM - Paris