Audio Coding and 3D Sound Simulation

Gaël Richard, Ariane Le Doré, Cédric Sibade, Jérome Boudy, Philip Lockwood,
MATRA Communication, rue JP Timbaud, 78392 Bois d'Arcy, France

Ulrich Horbach, Matthias Rosenthal
STUDER Professional Audio AG, Althardstr. 30, CH-8105 Regensdorf, Switzerland

Abstract

This paper presents recent advances in Audio Coding and 3D Sound Rendering. The techniques described are in lines with MPEG-4 requirements and have already been proposed to the MPEG community. This paper is divided in two parts. On the one hand, a scaleable audio decoder based on an ITU core coder is described. The scheme proposed is particularly attractive since it allows for fine granularity scalability with high coding quality. On the other hand, original and efficient 3D sound rendering techniques are described. These algorithms combine high quality sound rendering with real time implementation capabilities.

1 Introduction

The purpose of this paper is to propose innovative techniques for scaleable audio compression and 3D sound simulation developed within the European ACTS project AC 105 called EMPHASIS. These algorithms have already been proposed to MPEG (*Moving Picture Expert Group*) for the future MPEG-4 standard. This standard will cover all aspects of audio/video coding and rendering and will include several degrees of interactivity. Concerning Audio compression, the current MPEG 4 *Verification Models (VM)* includes some of the following key functionalities:

- Variety of objects (speech, music, various sampling frequencies)
- Bitrate scalability (by means of an editable bitstream)
- Complexity scalability (by means of an editable bitstream)
- Random Access
- Compression efficiency

However, the main concept of scalability that is to say the capability for the multiplexer to shorten/simplify the bitstream in order to fit the object to the bandwidth available and to the decoder processing power is still missing. This document proposes, in a first section, a solution integrating any speech or music standard as a core algorithm in a bitstream structure providing enhancement layers permitting to transmit and decode objects at various bit-rates, sampling frequencies and complexity. Our solution consists in decomposing the audio signal on frequency bands and in encoding the base band with a standard algorithm (MPEG-4 VM, ITU standard or GSM). The base band forms the base layer of the bit stream. An enhancement layer re-encoding the error of the standard algorithm can be added if necessary and then the higher bands are encoded into separate tags of the bitstream by mean of Vector Quantization (VQ).

Concurrently, the MPEG-4 standard will include means for 3D Sound rendering. More precisely, the "Structured Audio" part of the SNHC VM 3.0 *(Synthetic and Natural Hybrid Coding Verification Model)* consists of very comprehensive means of generating and manipulating both synthetic and natural sound signals (see [1]). Little attention has been paid, however, to spatialization techniques, which in fact demand most of the real time processing power.
The aim of the second section is thus to propose newly developed schemes which offer great advantages in terms of computational efficiency and flexibility where stereo and multichannel reproduction are both supported.

In summary, the paper is organised as follows: the next section describes an innovative technique for scaleable audio compression. Section 3 will propose efficient schemes for 3D sound rendering. Finally, section 4 will suggest some conclusions and future directions of research.

2 A generic scaleable audio coder

2.1 Main principle

In order to meet MPEG-4 requirements, a scaleable audio codec using existing ITU standards is developed (see [2]). The basic idea is to decompose the audio signal into a base band (typically 0-4kHz, 0-8kHz, 0-16kHz or 0-32kHz) coded using an ITU standard and several higher bands coded using Vector Quantization. The subband decomposition is obtained thanks to a filter band module. This original scheme results in a

very good coding quality (thanks to the ITU standard) but also in a variable coding rate (from 6.3 kbit/s to 80 kbit/s with a medium quality full band reproduction at 38kbit/s and in a variable decoder complexity. At the decoder side, the bitstream can be either completely decoded in order to obtain the highest possible quality or be only partially decoded when low bit rates becomes mandatory. Moreover, it must be emphasized that this scheme allows for very small steps enhancements (on the order of 1.8 kbit/s).

2.2 Subband decomposition

The filter bank module is designed to achieve either a linear bandwidth decomposition, or a non linear bandwidth decomposition simulating Bark bands (see for example [11]). The Bark scale filter bank structure turned out to be much more efficient for coding, because of a better bit allocation and consequently a better signal quantization of signal information.
It is extremely important that the decomposition itself does not degrade the input signal. To that end, a perfect reconstruction filter bank was chosen. The current procedure uses 36 taps wavelet filters which are known to be smooth and regular with less coefficients than QMF filters.

The linear decomposition is in fact a regular tree, which provides subbands with 1 kHz bandwidth (except for the subband 0-4 kHz), and as a consequence, a constant delay for each subband, except for the first one (see Figure (top)). The Bark decomposition is non linear and more complex. Because an exact Bark scale is not mandatory for speech coding, an approximate Bark scale was adopted which reduced the complexity of the decomposition (see Figure 1 (bottom)).
It must be emphasised that if the Bark decomposition is more complex, it leads to better results than the linear decomposition (higher audio quality at same bit rates). This is not a surprising result since the Bark scale is defined from human perception criteria.

The configuration shown below uses a standard encoder for the 0-4kKz band, another possibility (as depicted with the doted lines) is to use a wide band encoder and to add enhancement layers up to 48 kHz. Yet another configuration allows the use of MPEG-2 AAC for the full bandwidth for bit-rates greater than 64kb/s. One may also notice from Figure 1 that these schemes accept different sampling frequencies. More precisely, the following sampling frequencies are accepted: 8 kHz, 16 kHz, 32 kHz and 64 kHz. In order to get a generic audio encoder/decoder which supports any input and output sampling frequencies, the following rules are adopted:

- At the encoder side: the input signal is up-sampled (if needed) to the closest higher internal sampling frequency. The possible values for the internal sampling frequency are 8, 16, 32 or 64 kHz.
- At the decoder side: the output signal is either downsampled to the original input sampling frequency or another as needed by the system, e.g. the internal frequency of the audio compositor.

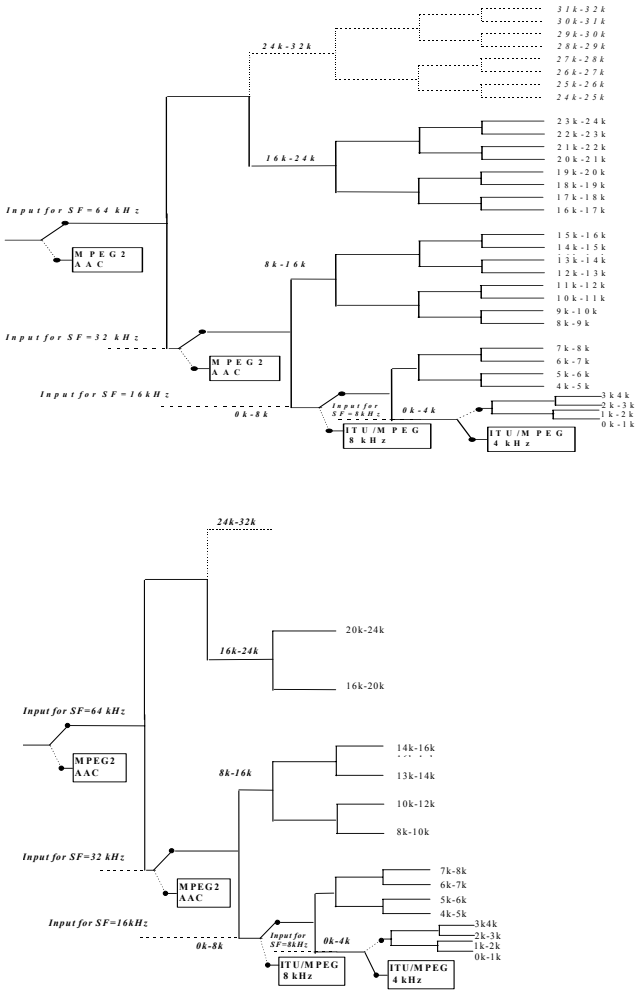

Figure 1: Linear analysis filter bank decomposition (top); pseudo Bark filter bank decomposition (bottom)

For example, an input signal sampled at 10 kHz will be up-sampled to 16 kHz before entering the filter bank. Once received, the audio signal can be played at any predefined sampling frequency.

2.3 Base layer coding (ITU, MPEG-4)

To allow a large degree of scalability, only ITU standards defined for a 8 kHz sampling frequency were considered. If some standards achieve high quality low bit-rate coding for speech signals (G. 723, G. 728, G. 729, MPEG-4 VM II), their performance on other audio signals (such as music for example) are poor. This result is not asthonishing since these standards exploits speech signal characteristics to improve the bit rate. Its already common use on Internet and its low bit rate (6.3 kbit/s) makes the G. 723 the most appropriate standard to code the base band. However, despite its high quality for speech signals, its performance is rather poor on music signals. It is thus essential for such signals to be

able to improve the quality by improving the bit rates and this even in the base band. The idea is to use a common technique based on Vector Quantization (VQ) to efficiently code the error between the original and G. 723 coded signals. In the base band, the smaller enhancement step corresponds to a raise of about 6 kbit/s leading to an average of 12.3 kbit/s.

2.4 Upper bands coding

The upper bands are directly coded using Vector Quantization (VQ). This technique is known to be very efficient for coding with high scalability capabilities. With such a technique, it is possible to distribute the bits for each band according to some criteria. The most straightforward criterion is a measure of energy. In that case, the frequency bands having low energy are not transmitted.
The same idea can be used to transmit signal at constant bit rates but variable quality. More precisely, it is possible to choose a number of bits, related to the desired bit rate, and to allocate bits to subbands according to their energy until there is no more bits available. Only the most "significant" subbands would be coded for a given bit rate. Each coded subband of 1 kHz correspond to bit rate of about 1.8 kbit/s for each enhancement layer. In practice, the gain in quality is almost insignificant above 3 enhancements layers which gives a maximum bitrate equals to 5.4 kbit/s per subband.

2.5 Bit stream structure

The main idea of the proposed bit stream structure is to allow variable decoding complexity and quality. For example, if the desired bit rate at the decoder is lower than the maximum proposed, it is possible to give up some of this information. However, it is necessary to drop the least important information to still have the highest quality for a given bit rate.
The following scheme is then adopted. The bit stream contains on the one hand the bit stream provided by the ITU standards and on the other hand those provided by the VQ. Figure 4 shows a simplified diagram of the bit stream structure for N subbands and P possible enhancements for each subband.

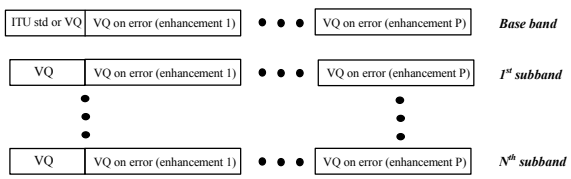

Figure 2: Bit stream structure

It must be emphasised that such a bit stream allows for a total decoding scalability with a unique coding scheme. In other words, it is possible from a given bit stream to either decode only the ITU standard part leading to a 4kHz bandwidth signal or to decode a full band/ high quality version of this audio signal. This scalability appears to be especially important for services over the Internet where the user will appreciate to choose the quality and thus the bit rate of the transmitted audio signals. It is also important to say that in the current scheme, the enhancement can be performed either within an already coded band (for example using VQ on the error signal) or by adding new subbands.

3 3D Sound Rendering

In a real environment, human perception is highly sensitive to the room characteristics and to sound sources location. The aim of the 3D audio composition tool is to give the impression to the user that he is immersed in a real environment and that the sound is in fact coming from a precise point in space. As a preliminary step towards this goal, a 3D room effect renderer is built and briefly described below. It is based on the one hand on the computation of early echoes in a square room and on the other hand on the simulation of late reverberation. The sound is finally spatialized using a "stereo panpot". Because this technique is too computionaly intensive to include sophisticated tools such as "Head Related Transfer Functions", simplifying assumptions are needed. Based on such assumptions, a new efficient scheme is proposed in section 3.2.

3.1 Simple procedure for room effects and spatialisation rendering

The aim of the 3D audio composition tool is to simulate the sound field perceived by a listener in a room containing several sound sources. To simulate the so-called room effect, it is necessary to take into account the effects of boundaries (walls, floor and ceiling) on the sound propagation. However, previous studies showed that the room effect can be decomposed into three contributions: a direct sound , followed by the early reflections (typically reflections of first and second order) and finally by late reverberation.

By taking advantage of the analogy between sound and light, a simple model can be built based on optical geometry theory. The basic principle is to compute the virtual sound sources that are created by reflection of the sound wave on an obstacle. This principle is depicted in a 2D space in Figure 3 for first order reflections. To achieve natural sounding simulation, it is necessary to compute a fairly high number of reflections (typically over three hundreds reflections).

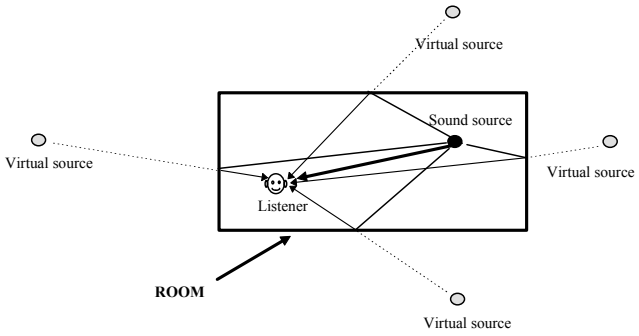

Figure 3: First order reflections in a room. The wider arrow symbolises the direct sound

Typically, the estimation of early echoes leads to a filter impulse response which characterises the effect of the room for given positions of a listener and a sound source. Figure 4 shows such a filter impulse response for a rather large room with highly reverberant walls. The first peak corresponds to the direct sound and each following peak represents an echo. The amplitude of each peak corresponds to the intensity of the corresponding reflection. This intensity is a function of the distance between the corresponding virtual source and the listener position and of the reflection coefficients at room boundaries.

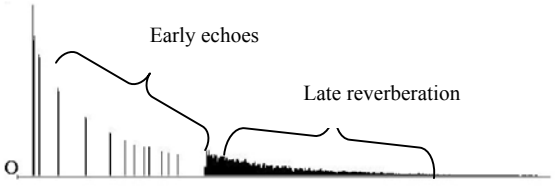

Figure 4: Schematic filter impulse response of a room. The first peak represents the direct sound.

To simulate, the late part of the impulse response of the room, a reverberant filter is applied to the sum of all the echoed sources. Its structure (cf. [3]) is based on matrix combination of comb filters, that provides a dense exponential decrease response. From parameters such as volume, delays and reverb, can be easily computed and hence the reverberation filter is simulated.

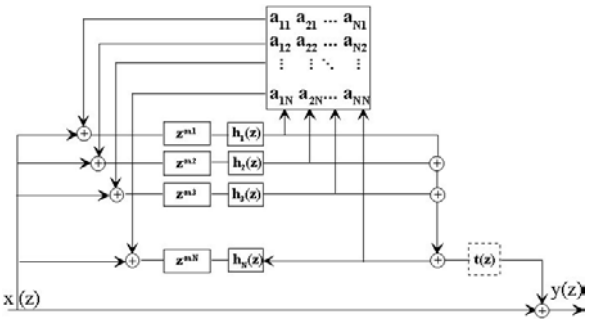

Figure 5 : Scheme of the reverberant filter (after [3])

Finally, to place a sound source at a particular position in space, a simple technique, in term of computation time, the "stereo panpot" is used. This technique means *stereophonic panoramic potentiometer*. From a monophonic signal, a two channels signal is calculated, by applying two different coefficients on the input signal: the only parameter of this method is the source angle of incidence with respect to from the listener.

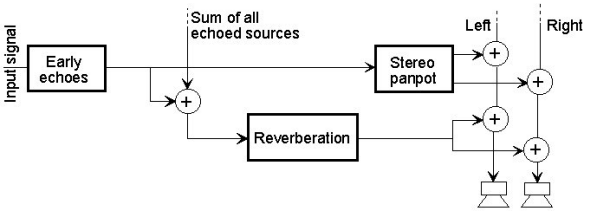

Figure 6: General structure

All these techniques have the advantage to be based on a physical description of the room and thus have the capability to provide very faithful sound field representation. For instance, it is possible to extend it to any type of room where each wall can be covered by a different material (wood, tiles, carpet,….). However, it is clear if one wants to design a real time algorithm, the room geometry must be simple and the number of computed echoes must be kept rather low. In our implementation, real time simulation is achieved on a Pentium 200 for up to three moving sound sources in a rectangular room where all echoes of first and second orders are computed. It is however possible to reduce the computational load by reducing the number of echoes to be calculated. In our MPEG paper [12], we describe a scaleable implementation of this technique.

Nevertheless, this high computational load motivates our research towards simpler scheme which would provide high quality 3D sound field rendering. Such a scheme is proposed below after a brief description of what would be a complete sound spatialization system.

3.2 Head Related Transfer Functions and Speaker decoder

In MPEG-4 document 1437 [4] Jens Spille gave a brief introduction into spatial audio techniques, which apply in general a pair of head related transfer functions (HRTFs) to the sound signal, corresponding to its angle of incidence with respect to the listener. In order to simulate the distance, the sound object must be placed into an environment with reflective walls, which create a number of echoes coming from different directions. With increasing distance, the density and relative energy of early echoes are rising, respectively. In order to give a realistic impression, the echoes itself must be treated as sound sources, hence filtered with pairs of HRTFs according to their positions in space.
The resulting well-known scheme is shown in Fig. 7. The temporal distribution of the direct path and the echoes, respectively, is represented by a tapped delay line, with total length of typically 3000-4000 samples (100msec), and N=5..30 taps. For each tap, a pair of

HRTF filters is applied, in order to simulate the left and right ear signals that would occur in a real environment, dependent on the angles of appearance of the auditory events. The individual time delays and the filters must be updated smoothly, if the objects change their positions in space. The same procedure applies for all of the audio objects to be spatialized (in the figure only one monaural signal is shown). Their binaural signals are summed up by means of a 2-channel summing bus, such that we obtain a "binaural representation" of the scene. The resulting stereo signal must in general be post-processed. In the case of headphone reproduction, appropriate equalization filters are needed. For loudspeaker reproduction a so-called "speaker decoder" is inserted, which contains a cross-cancellation circuit (see [4]).

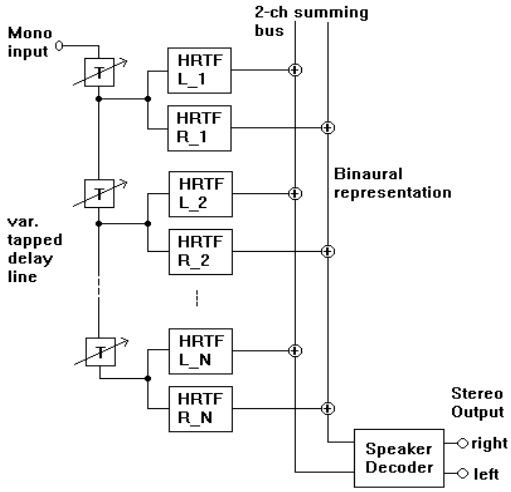

Figure 7: Spatialization scheme using head related transfer functions (HRTF).

The following drawbacks of the above described algorithm are obvious:

1. The required amount of processing power is very high. For example, if we implement the HRTFs as 50-taps FIR filters, each pair will cost around 3.2 MIPS (sampling frequency 32kHz). Assuming N=5 (we calculate 5 echoes only), as much as 16 MIPS (one standard DSP) are required for the calculation of only one channel. Thus, the spatialization of more than a few objects seems not reasonable.
2. Only a stereo output is available, multichannel formats are not supported easily.
3. The scheme is not very flexible in terms of scaleable complexity.
4. Non-point sources with directional characteristics, as described in VRML 2.0 [5], cannot be spatialized in a straightforward way.

3.3 Simplified spatialization scheme

The new scheme is depicted in Fig. 8.

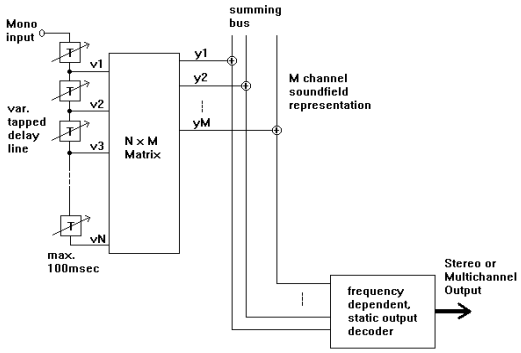

Figure 8: Simplified spatialization scheme

Each channel consists of a tapped delay line (representing the first echoes), the output signals (v_1 .. v_N) of which are fed into an NxM matrix with real-valued coefficients, in order to create an M-channel soundfield representation, which is, like before, the overall sum over all channels. Compared to the standard approach, the HRTF-related filtering is now performed statically (that means not dependent on the movement of the objects or the scene) with a bank of filters located at the output. Only one single output decoder is therefore needed in a system of arbitrary complexity. Any output format (mono, stereo, multichannel) can be supported, by adapting the decoder only. The main idea behind this approach is the following. The M-channel soundfield representation provides a coding scheme such that the original direction of the signals can be recovered in the decoder within a given resolution which depends on the number M. The higher M, the higher will be the angular resolution. For example, if a stereo zone of +/- 30° needs to be covered with a 10° resolution, 6+1 channels are required. The decoder contains 6 pairs of HRTFs in this case. An example of such a coded representation is the so-called "Ambisonic B-Format" (see [6]). Here the matrix coefficients (columns) are w, sin mϖ, cos mϖ , m=1..(M-1)/2, w stands for the level of the signal, ϖ the angle of incidence. For conventional stereo, 2 channels are sufficient. We then obtain the signals of an MS-microphone w, sin ϖ. The corresponding output decoder is shown in Fig. 9.

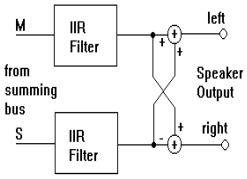

Figure 9:Output Decoder for Stereo

The filters which are applied here for the M (mid) and S (side) signals, respectively, can be derived from simplified HRTFs, and realized by cost effective IIR (infinite impulse response) designs [7]. For a more detailed description, you may read [10].

3.4 Complexity Reduction and Scalability

The following parameters and methods are subject to individual, per channel, and overall complexity scalability.

1. Number of echoes N to be calculated (per channel)
2. Angular resolution M (overall or per channel)
3. Number of output channels. Mono, 2-channel Stereo, and the newly introduced 3/2 format [8] are possible candidates. Only the output decoder needs to be modified.
4. Directional characteristics of the sound sources. Non-point sources must be represented by a number of point sources located at different directions. This can be accomplished by applying several "direct paths", with short or zero delays in between, but different matrix coefficients.
5. Accuracy of the HRTF model (overall, complexity of the decoder). Can be reduced for example by principal components analysis [7], or adapting the parameters of biquad filter sections. Note that for surround reproduction with loudspeakers, a very coarse head model, e.g. sphere or ellipsoid, may be advantageous, because of the lower coloration associated (see [8]).
6. Optimisation of the preprocessor. Mainly, the echo calculation scheme can be simplified. Rather than using well-known algorithms like ray-tracing or the mirror-image method, which need a lot of processing power, and depend on a comprehensive geometric description of the scene, which may not be available in any case, stochastic generation methods based on perceptual parameters may be applied. For an introduction, see [9]. This is very useful, for example, if only a sense of distance needs to be provided, without unwanted side effects (coloration, unnatural sound, if the room model is very simple). These stochastic methods exploit basic principles (dependencies of the density, the number and amplitude of echo lines on the room size, the distance to the listener, absorption factors, decrease of the amplitude by 1/r with distance r, etc.) rather than calculating them individually.

4 Conclusion

In the first part of this paper, a scaleable audio and speech coder was presented whose philosophy and underlying concepts strictly follow the MPEG-4 requirements. This proposal does not aim to replace existing verification models but rather to provide additional scalability to the previously proposed models. As a matter of fact, it provides a total scalability at the decoder side thanks to its bit stream structure. It must also be noted that its complexity is compatible with hardware constraints. Nevertheless, this codec still needs improvements in order to increase the subjective audio quality along with higher compression efficiency. In particular, current work is devoted to include prediction modules within bands and to continue our initial efforts toward a generic codec that would better exploit the AAC standard. It is important to emphasise that any improvement can be tested without changing the general codec structure thanks to its true genericity.

In the second part of this paper, issues on 3D sound rendering were discussed and two techniques implemented on PC were described. The realism of the 3D sound images was very satisfying and showed that it is important to have a closer look at the spatialization tools provided by the SNHC verification model. As a matter of fact, it may be appropriate to develop a methodology for comparing and evaluating 3D sound algorithm by means of listening tests.

5 Acknowledgements

This work is partly supported by the European Community within the project ACTS - AC 105 (EMPHASIS).

References

[1] C. Horne: *SNHC Verification Model 3.0,* MPEG97/ N1545, Feb. 97, Sevilla

[

2] G. Richard, P. Bonnard, Ariane Le Doré, *"A solution for a scaleable Audio and Speech coder based on Core Coders",* input document MPEG97/ M1997, April. 97, Bristol

[3] J.M. Jot, *"Etude et réalisation d'un spatialisateur de sons par modèles physique et perceptifs",* (in french) Télécom Paris 92 E 019.

[4] J. Spille: *"Human Sound Localization",* Input Doc. MPEG96/ 1437, Nov. 1996

[5] INTEL Architecture Labs: *"Audio Models in VRML: A Proposal"*
http://developer.intel.com/ial/rsx/vrmlnode.htm

[6] J. S. Bamford: *"Ambisonic Sound for Us",* 99th Convention of the Audio Engineering Society (AES), New York 1995, preprint 4138

[7] M.J.Walsh, D.J.Furlong*: "Improved Spectral Stereo Head Model",* AES New York 1995, preprint 4128

[8] G. Theile: *"On the Performance of Two-Channel and Multi-Channel Stereophony"* 88th AES Convention, Montreux 1990, preprint 2887

[9] M. Gerzon: *"The Design of Distance Panpots",* 92th AES Convention, Vienna 1992, preprint 3308

[10] M. Rosenthal: *"Implementation of Audio Compositing Functions: Software Considerations",* Input Document M1932, Bristol 1997

[11] E. Zwicker and E. Feldtkeller. *"Psychoacoustique, l'oreille récepteur d'information"* (in french). Masson Edts. Paris, 1981.

[12] P. Bonnard, G. Richard, C. Sibade, F. Rigoulet, *"An implementation of graceful-degradation concept in a 3D audio compositor"* input document MPEG97/ M1998, April. 97, Bristol