Audio Engineering Society

Convention Paper

Presented at the 122nd Convention
2007 May 5–8 Vienna, Austria

New Enhancements to Immersive Sound Field Rendition (ISR) System

Chandresh Dubey[1], Raghuram Annadana[1], Deepen Sinha[1] and Anibal Ferreira[1, 2]

[1] ATC Labs, New Jersey, USA

[2] University of Porto, Portugal

Correspondence should be addressed to Chandresh Dubey (chandresh@atc-labs.com)

ABSTRACT

Consumer audio applications such as satellite radio broadcasts, multi-channel audio streaming and playback systems coupled with the need to meet stringent bandwidth requirements are eliciting newer challenges in parametric multi-channel audio coding schemes. This paper describes the continuation of our research concerning the *Immersive Soundfield Rendition* (ISR) system and the different enhancements in various algorithmic components. The need to maintain a constant bit rate for many applications requires a rate control mechanism. The various strategies utilized in the rate control mechanism are presented. In addition, an innovative phase compensated down-mixing scheme has been incorporated in the ISR system so as to generate a high quality carrier signal. Enhancements have been made to the blind up-mixing scheme and to considerable gains have been made in terms of acoustic diversity. The various enhancements of the ISR system and its performance are detailed. Audio demonstrations are available at http://www.atc-labs.com/isr.

1. INTRODUCTION

Parametric multi-channel audio coding at low bit rates (e.g., 0-12 kbps overhead) has numerous emerging applications. These include multi-channel satellite broadcast systems and audio streaming and plans are underway to incorporate such schemes in Satellite Digital Audio Radio (SDAR) and other digital broadcast systems. We describe enhancements to our recently introduced *Immersive Sound-field Rendition* (ISR) [1] capable of operating in the range of 0-12 kbps range. Several new algorithm components have been incorporated:

- A new phase compensated down-mixing scheme has been incorporated. The scheme compensates phases in the ODFT domain.
- A new blind up-mixing scheme has also been incorporated leading to further gains in terms of image quality of the signals.

- ❖ Acoustic diversity has been further improved upon by detecting harmonic patterns [4] in surround and the carrier signal.

- ❖ The need for constant bit rate requires the use of a rate control mechanism. The various strategies used for bit allocation include flattening the multi channel temporal envelope on the time-frequency grid. In addition, a new correlation coding [2] has also been incorporated to lower bit rates.

The organization of the rest of the paper is as follows. A brief overview of the ISR system is presented in section 2. Section 3 discusses the various enhancements to the ISR system viz. the rate loop implementation, phase compensated downmixing and blind upmixing, in addition to a scheme for acoustic diversity creation. Some preliminary results are presented in section 4 and is followed by conclusions in section 5.

2. OVERVIEW OF PREVIOUS WORK

Immersive Sound-field Rendition (ISR) System [1] is an innovative scheme for very low bit rate multichannel parametric audio coding. As shown in Figure 1 a conventional stereo encoder can be upgraded to behave as a multi-channel encoder with the aid of an associated low overhead ISR bitstream. A key idea in ISR coding is that the spatial localization cues [5, 6] of the original multichannel audio are accurately reproduced with the aid of a multi-band temporal envelope generated on a detailed time-frequency grid. The multi-band temporal envelope is efficiently encoded and is applied to the down-mixed stereo carrier in the ISR decoder. Coding of multi-band temporal envelope, Multiband Temporal Amplitude Coding (MBTAC) [2, 3], involves an initial *utility filter bank* (*UFB*) analysis. *UFB* is an over-sampled complex modulated filterbank with up to 16 times over-sampling. The subbands of the *UFB* are successively grouped in a perceptually motivated manner followed by quantization and coding. The bit requirement for *MBTAC* is further reduced by exploiting the correlation in frequency of the multiple time domain envelopes.

A second key idea in the ISR coding scheme is a technique to create acoustic diversity between front and surround decoded channels [1] utilizing techniques for accurate tone and harmonic analysis [4, 11]. The Acoustic Diversity creation techniques allows for the generation of a different surround carrier from the primary (front) carrier over which the surround envelope is applied. This is quite useful when the acoustic characteristics of the surround channels are substantially different from the front channel; for example if an instrument or vocal component is present only in the front channel, its leakage to the surround channel can be minimized with the help of this technique.

Four modes of operation were proposed and discussed in the original ISR system

- ❖ Detailed multi-channel reproduction with 14 – 17 kbps as overhead
- ❖ High quality multi-channel reproduction with 8 – 12 kbps as overhead
- ❖ Realistic multi-channel reproduction with 4 – 6 kbps as overhead
- ❖ Blind upmixing with a 0-2 kbps overhead (blind/ near blind upmixing)

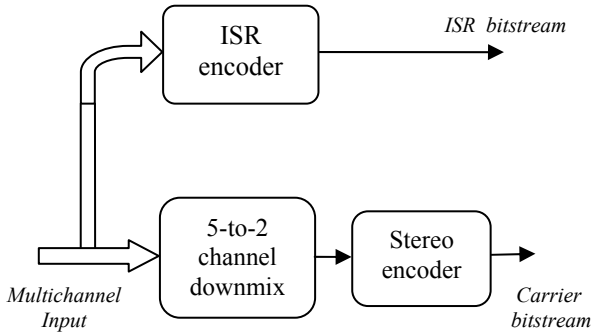

Figure 1: Architecture of ISR Encoder

3. ENHANCEMENTS TO ISR

Low bit overhead modes of operation of the ISR system such as the Realistic multichannel reproduction (4 – 7 kbps) and blind upmixing requires a significantly different processing in order to create a better perception of multi-channel audio, with improved stability of the audio image and spaciousness. The targeted areas for improvements have been to create a better downmixing and upmixing scheme, in addition to an improved acoustic diversity creation approach.

3.1. Phase Compensated Five-to-Two Channel Down-mix

Experiments with ISR System have proven that an accurate synthesis of multi-channel audio is dependent on the carrier down-mix. A new phase compensated

downmixing scheme has been incorporated for this purpose. Downmixing is performed in the ODFT domain.

As shown in Figure 2, a relative phase alignment of the Left and Right channel pairs along with the center channel is performed on ODFT partitions. Temporal phase smoothing is followed by a scheme for adaptively mixing the channels to obtain a downmixed signal. This scheme has led to significant gains in terms of stereo image quality of the synthesized front and surround pairs.

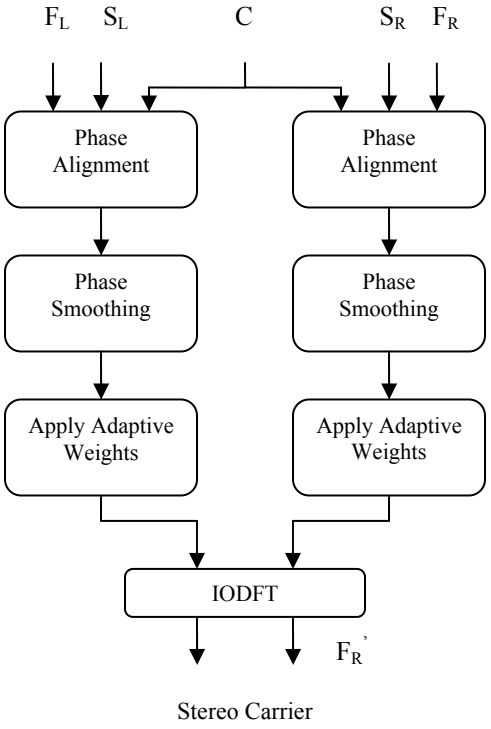

Figure 2: Downmixing in ODFT Domain

3.2. Improved Acoustic Diversity Creation

As previously discussed, the ISR system uses a carrier downmix audio to synthesize the front, surround and center channels. Strong harmonics in the front channel has an impact on the synthesized surround due to the nature of the downmix. This is not a desirable situation if a particular harmonic patter (from a musical instrument or vocal sound) is present only in the front channels and not the surround channel.

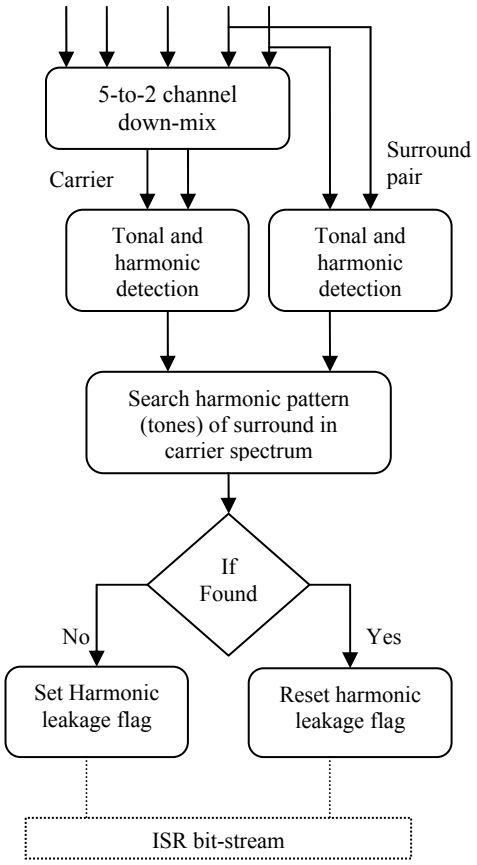

Figure 3: Harmonic Removal at ISR Encoder

In order to mitigate this problem an Acoustic Diversity creation technique for preparing a new carrier (at the decoder) for use in the generation of surround channel was found to be desirable [1]. Here we describe an enhanced scheme for Acoustic Diversity Creation. The new scheme as shown in Figure 3 has now been incorporated in the ISR system. Accurate harmonic analysis of the carrier downmix and surround channels at the encoder yields information of the harmonic structure in the respective channels. A transmitted flag is used to indicate harmonic leakage in the surround channels at the decoder. The decoder removes the harmonics from the carrier when the transmitted flag indicates absence of a particular harmonic pattern in the surround channel. This new updated carrier used for synthesizing the surround channel.

The rationale for the above described scheme is as follows. For example, if Fc_0 and Fc_1 are the harmonic

complexes present in downmixed carrier, the presence of these complexes may be either from the front channel or the surround channel or both. If the harmonic complexes are from only the front channel (i.e. surround is devoid of these harmonics), the effect of these harmonics on the synthesized surround is evident. A harmonic analysis of the carrier and surround will suffice to counter this problem and a harmonic leakage flag is set when harmonic patterns in surround and carrier don't match.

3.3. New Blind/Near Blind Up-mixing:

A new blind upmixing approach has been incorporated in the ISR system. This involves a novel method for generation of the surround channel. In this scheme, the front and center channels are generated using a scheme similar to [1]. The surround channels are generated using a formulation summarized by Figure 4. Principal Component Analysis (PCA) of the input stereo produces two vectors indicating the direction of both a dominant signal ***y*** and the remaining signal ***q***.

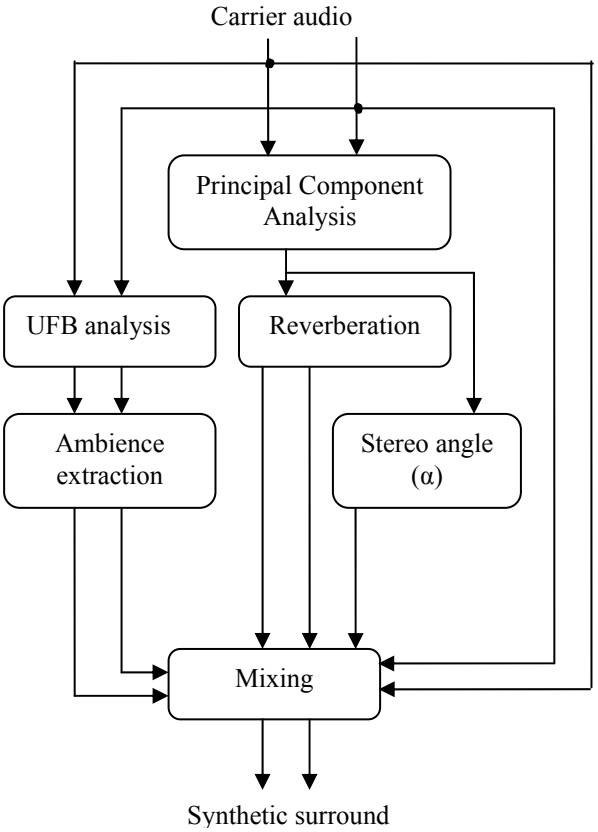

Figure 4: Surround Synthesis

The surround channel is synthesized by adaptively mixing three components viz. ambience extracted signal of the carrier audio, reverberation of the dominant principal component vector ***y*** and a third Image Movement Spatialization Component (*IMSC*) based on the original carrier audio. The third, *IMSC*, component which is weighted version of the original carrier audio is based on the rate of change of the stereo angle of the dominant signal component. The sections that follow describe in brief the different steps involved in the process.

3.3.1. Ambience Extraction

A method for frequency domain ambience extraction [18, 12] has been adapted for this scheme. An oversampled utility filter bank analysis on the carrier audio is used to generate a time-frequency grid.

Assuming the left and right channels are weakly correlated, a coherence function of the stereo is computed using cross correlation of left and right channels [13]. If $C_L(t,f)$ and $C_R(t,f)$ are frequency domain representation of left and right channels respectively, where t index is the time instant and f is sub-band index respectively, then the coherence function is given by

$$\varphi_{lr}(t,f) = \frac{E(C_L(t,f)C_R^*(t,f))}{[E(C_L(t,f)C_L^*(t,f)).E(C_R(t,f)C_R^*(t,f))]^{\frac{1}{2}}} \quad (1)$$

Where E is expectation operator and φ is coherence measure function. A non linear function is derived through this coherence measure which is used as panning function to extract the ambience signal between left and right channels and is given by,

$$Amb_i(t,f) = C_i(t,f)\theta(t,f) \quad (2)$$

3.3.2. Reverberation Generation

Reverberation [14, 15, and 18] is the persistence of sound in a particular space after the original sound is removed. When sound is produced in a space, a large number of echoes build up and then slowly decay as the sound is absorbed by the walls and air, creating reverberation, or reverb.

Since reverberation is essentially caused by a very large number of echoes, feedback delay networks (FDN) [16] have been used to create large and decaying echoes.

Feedback delay networks can simulate the time and frequency domain responses of real rooms based upon room dimensions, absorption and other properties.

In our scheme, two de-correlated reverberated audio signals are created from the dominant component obtained from the carrier PCA analysis. Figure 5 shows one possible delay network to obtain reverberated signals. The transfer function for the block diagram in Figure 5 is given by Equation 3.

$$H(z) = \mathbf{c}^T.[1 - \mathbf{D}(z).\mathbf{A}]^{-1}.\mathbf{D}(z)\mathbf{b} + d \qquad (3)$$

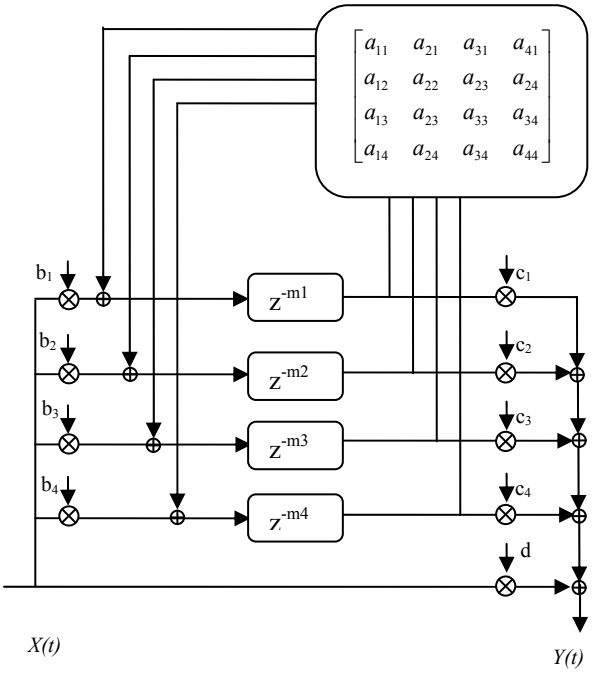

Figure 5: FDN in ISR System

Matrix **A** is an unitary feedback matrix. Matrices **D**(z), **b** and **c** are given by

$$\mathbf{D}(z) = \begin{bmatrix} z^{-m1} & & 0 \\ & \ddots & \\ 0 & & z^{-mN} \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} b1 \\ b2 \\ \cdot \\ \cdot \\ \cdot \\ bn \end{bmatrix} \qquad \mathbf{c} = \begin{bmatrix} c1 \\ c2 \\ \cdot \\ \cdot \\ \cdot \\ cn \end{bmatrix}$$

3.3.3. Image Movement Spatialization Component (IMSC) Generation

Synthesizing the IMSC component utilizes a novel scheme exploiting the stereo image angle (α) of the dominant signal component that is available as a byproduct the Principal component analysis [9]. The rate of change of stereo image angle gives a measure of the movement of the stereo image over the front plane in the stereo downmix. This information is utilized in the final mixing process for the creation of the synthetic surround allowing for a tangible shift in the surround audio image from front to back (or back to front), thereby increasing the perception of spaciousness in the synthetic multi-channel signal.

Figures 6 and 7 illustrate the multichannel blind upmixing and the audio image as perceived before and after the incorporation of the above mentioned IMSC component. For example, for a image moving from left to right, the harmonics from the left front are leaked to left surround based of the rate of change of the stereo angle, giving the listener a feel of the image moving in from the behind. This space filling effect creates a pleasant perception.

Figure 8 shows an example plot of stereo image movement for the audio sample "Blackwater.pcm". The stereo angle oscillates around $\pi/4$ indicating continuous stereo image movement from left to right and back. Stereo down-mixes with similar characteristics benefit significantly from the addition of the IMSC component in the sense that any regular and smooth movement of stereo image from one side to other results in a corresponding diagonal movement of the 3-D image in the surround space for the synthetic multi-channel signal hence increasing the sense of spaciousness or immersiveness in the ISR system.

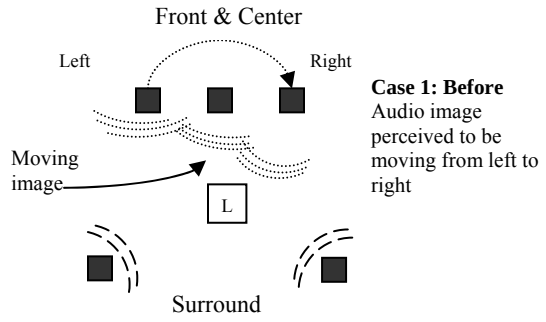

Figure 6: Example illustration of perceived audio image movement without IMSC component

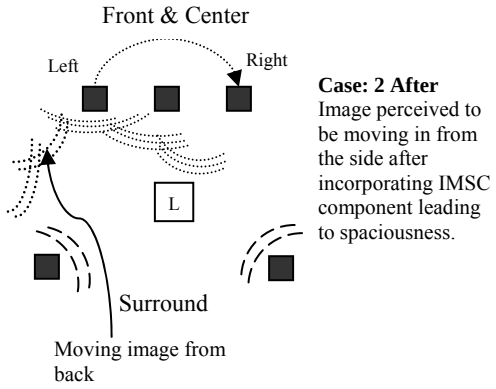

Figure 7: Perceived Image in ISR System with the incorporation of IMSC component.

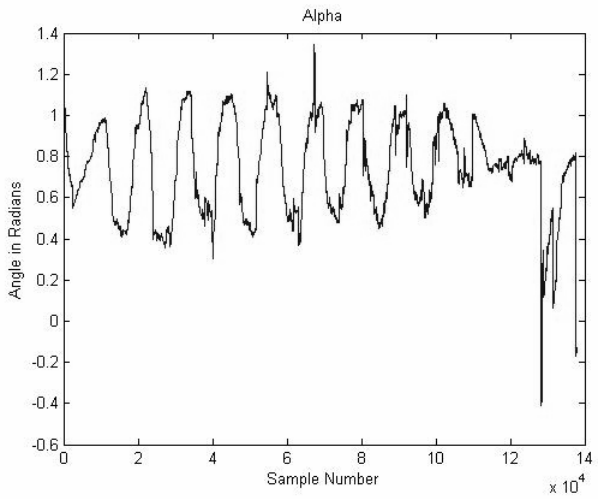

Figure 8: Fluctuation of the stereo image angle in the "Blackwater" signal.

The final surround audio is synthesized by mixing Ambience (complex audio content), Reverberation of the dominant component (naturalness) and the IMSC component.

3.4. Constant Bit Rate Implementation

Figure 9 illustrates an implementation of a rate control mechanism in the ISR system in order to maintain a constant bit rate. The various grouping thresholds in the joint inter-channel envelope coding are re-adjusted based on the available bits in the bit buffer and the harmonic measure of the input audio.

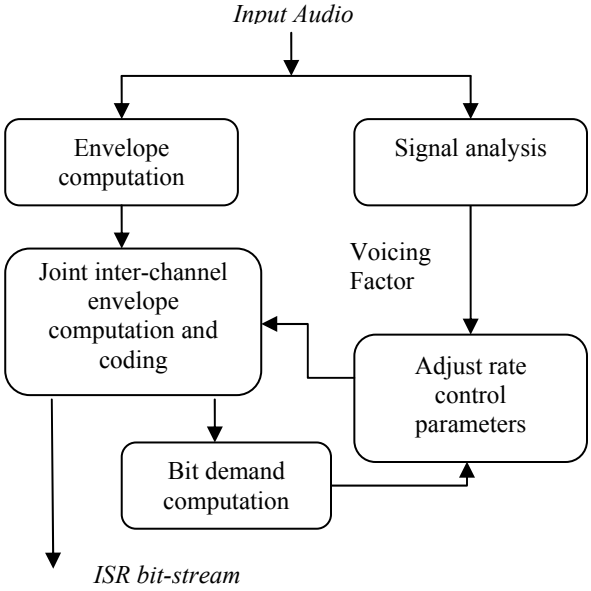

Figure 9: Rate Loop in ISR System

4. RESULTS

Preliminary listening tests to analyze the audio quality in each of the ISR system modes have been conducted. The audio quality has also been compared against MPEG Surround system [19] which is based on the Spatial Audio Coding/Binaural Cue Coding (BCC) techniques [20]. In terms of the overall quality informal listening by trained listeners indicates that the image quality and accuracy for the constant bit rate 12 kbps ISR system is noticeably better than the MP3-Surround system operating at 128 kbps.

In another set of listening tests the performance of the proposed downmixing scheme has been compared

against the standard ITU-T downmix technique. Informal listening tests have revealed that there is a significant loss of surround signal component using the standard downmix. This has been substantially mitigated using the phase-compensated downmixing scheme presented here.

5. CONCLUSIONS

We have presented some of the algorithmic enhancements in the recently introduced ISR system. The ISR system is enhanced with the help of (i) an improved Acoustic Diversity Creation module to reduce the leakage of unwanted signal components to the surround channels, (ii) incorporation of an enhanced blind upmixing scheme, and, (ii) incorporation of an improved downmixing scheme involving phase compensation. The new and improved blind upmixing scheme adaptively mixes three components – viz. ambience, reverberated dominant PCA signal and the Image Movement Spatialization Component (IMSC). A rate control implementation to facilitate constant bit rate operation has also been presented. Listening tests indicate that audio quality of the proposed system is superior to the MP3-Surround system. Furthermore, listening tests also indicate that there is a loss of signal components due to phase cancellation in the conventional downmixing which is mitigated by the proposed algorithm. Audio demonstrations and more details are available at http://www.atc-labs.com/isr.

6. REFERENCES

[1] Chandresh Dubey, Richa Gupta, Deepen Sinha and Anibal Ferreira, "Novel Very Low Bit Rate Multi-Channel Audio Coding Scheme Using Accurate Temporal Envelope Coding and Signal Synthesis Tools", Presented at the 121st AES Convention October 5–8, 2006 San Francisco, CA, USA.

[2] Raghuram Annadana, Harinarayanan. E.V, Anibal Ferreira, and Deepen Sinha, "New Results in Low Bit Rate Speech Coding and Bandwidth Extension", Presented at the 121st AES Convention October 5–8, 2006 San Francisco, CA, USA.

[3] D. Sinha, A. J. S Ferreira and Harinarayanan E. V., "A Novel Integrated Audio Bandwidth Extension Toolkit (ABET)", in the preprints of 120th Convention of the Audio Engineering Society, May 2006.

[4] Anibal J. S. Ferreira and Deepen Sinha, "Accurate and Robust Frequency Estimation in ODFT Domain," in the proceedings of the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2005.

[5] Anibal J. S. Ferreira and Deepen Sinha, "Accurate Spectral Replacement", in the Preprint of 118th Convention of the Audio Engineering Society, Barcelona, Spain, Convention Paper, May 2005.

[6] Lord Rayleigh, J.W. Strutt, "Our perception of sound direction," Philosophical Magazine 13:214-232, 1907.

[7] Jens Blauert, "Spatial Hearing", Revised Ed. MIT Press (1996). ISBN 0-262-02413-6.

[8] J. Stautner, M. Puckette, "Designing multi-channel reverberators", Computer Music Journal, 6(1), 1982.

[9] T. W. Lee, "Independent Component Analysis: Theory and Applications" Kluwer, Boston, MA, 1998.

[10] Van Der Waal, R.G. and Veldhuis, R.N.J. "Subband coding of stereophonic digital audio signals", Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), IEEE Computer Society Press, Los Alamitos, California, pp. 3601-3604, 1991.

[11] Anibal J. S. Ferreira, "Perceptual Coding of Harmonic Signals", 100th Convention of the Audio Engineering Society, May 2005, 1996.

[12] Carlos Avendano, Jean-Mark Jot, "A Frequency-Domain Techniques For Stereo to Multichannel Upmix", 22nd AES international Conference on Virtual, Synthetic and Entertainment Audio, Espoo, Finland, 2002.

[13] J. Allen, D.A. Berkeley and J. Blauert, "Multi-microphone Signal-Processing Technique to Remove Room Reverberation from Speech Signals" Journal of Acoustical Society of America, Vol. 62, No.4, pp 912-915, Oct 1977.

[14] M.R. Schroeder, "Natural Sounding Artificial Reverberation", J. Audio Eng. Soc., 10(3), 1962.

[15] I.A. Moorer, "About This Reverberation Business", Computer Music Journal, 3(2), 1979.

[16] Jot, J.M., "Digital Delay Networks for Designing Artificial Reverberators", AES 90th Convention Preprints, 1991.

[17] R. Irwan and Ronald M. Aarts, "Two-to-Five Channel Sound processing", J. Audio Eng. Soc., 50(11):914-926, November 2002.

[18] T. Holman, "Mixing the Sound", Surround Magazine", June 2001.

[19] Breebaart, J., Disch, S., Faller, C., Herre, J., Hotho, G., Kjörling, K., Myburg, F., Neusinger, M., Oomen, W., Purnhagen, H., and, Rödén, J., "MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status", AES 119th Convention Preprints, 2005.

[20] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger and P. Kroon, "Spatial Audio Coding: Next-generation efficient and compatible coding of multi-channel audio", Preprint AES 117th Convention, 2005.