Audio Engineering Society

Convention Paper

Presented at the 121st Convention
2006 October 5–8 San Francisco, CA, USA

A Novel Very Low Bit Rate Multi-Channel Audio Coding Scheme Using Accurate Temporal Envelope Coding and Signal Synthesis Tools

Chandresh Dubey[1], Richa Gupta[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@atc-labs.com

ABSTRACT

Multichannel audio is increasingly ubiquitous in consumer audio applications such as satellite radio broadcast systems; surround sound playback systems, multichannel audio streaming and other emerging applications. These applications often present challenging bandwidth constraints making parametric multichannel coding schemes attractive. Several techniques have been proposed recently to address this problem. Here we present a novel low bit rate five channel encoding system that has shown promising results. This technique called the Immersive Soundfield Rendition (ISR) System emphasizes accurate reproduction of multi-band temporal envelope. The ISR system also incorporates a very low over-head (blind upmixing) mode. The proposed multichannel coding system has yielded promising results for multi-channel coding in 0-12 kbps range. More information and audio demos are available at http://ww.atc-labs.com/isr.

1. INTRODUCTION

With the prevalence of multi-channel audio reproduction equipment (4-5 speakers on or more) in home entertainment systems, computer audio setups, and automotive audio; the need for multichannel audio is being increasingly felt in even very low bit rate audio broadcast and playback applications. Several initiatives are currently underway to incorporate multichannel audio in Satellite Digital Audio Radio (SDAR) systems, Terrestrial Digital Audio Broadcasting (DAB) systems, and other low bit rate audio applications. Research in the area of parametric low bit rate multichannel audio coding has therefore proven to be attractive.

It has long been recognized that binaural audio image perception is based on certain primary cues such as Interaural Level Difference (ILD) and Interaural Time Difference (ITD) [1],[2]. The spectral level differences between the audio channels are referred to as ILD. ILD is the primary localization cue at frequencies above approximately 1.5 kHz. At the lower frequencies sound waves travel through the head and are thus not substantially attenuated (between the two ears); hence at lower frequencies ITD cues are the dominant localization cues. Although ITD cues may be perceptually significant for frequencies up to 5 kHz (or even higher), absolute phase differences are significant only at frequencies below 1.5 kHz. At higher frequencies ambiguity in the perception of absolute phase differences makes these relatively unimportant, and, inter-aural group delay differences (*time difference between envelopes*) and inter-channel coherence are the remaining pertinent aspects of ITD. Several parametric techniques for the coding of stereo audio signals using the synthesis and/or efficient coding of these cues have therefore been proposed [3], [4], [5]. It has been further recognized that similar localization model and synthesis may be extended to multichannel audio perception and coding [6].

In the parametric multichannel coding, a 5 channel audio source – consisting of a Left (L), Right (R), Center (C), Left Surround (Ls), and Right Surround (Rs) - is coded as a stereo downmix and a low overhead side stream. The side stream should contain enough information to allow the decoder to recreate a 5 channel upmix from the stereo downmix (which itself has typically been coded by low bit rate audio coding scheme such as the codecs described in [9], [10], [11], [12], [13]). The audio quality of the upmix as well as the bit overhead for the side stream is the most two key performance criteria. We propose and describe a new multichannel coding scheme called the Immersive Soundfield Rendition (ISR) system that emphasizes the accurate reproduction of the temporal envelope of the multichannel signal in multiple frequency bands. Additionally, the ISR system attempts to maintain any existing acoustic diversity between the front and surround audio channels (in other words if certain instruments and/or vocals are present only in the front channels these are removed). The proposed ISR system has 4 modes of operation: (1) a high fidelity mode requiring a side stream overhead of 14-16 kbps, (2) a high quality mode requiring an overhead of 8-12 kbps, (3) a realistic multichannel mode requiring an overhead of 4-6 kbps, and, (4) a blind or "near" blind mode requiring an overhead of 0-2 kbps.

The organization of the rest of the paper is as follows. Section 2 provides an overview of the previous approaches. Section 3 describes the key ingredients of the proposed coding scheme and the synthesis procedure of spatial image reconstruction. Section 4 describes various configurations of the Immersive Soundfield Rendition (ISR) System in detail. Coding results are presented in section 5 followed by conclusions in section 6.

2. OVERVIEW OF PREVIOUS APPROACHES

The main concern in surround sound audio reproduction is to provide a homogeneous and coherent sound field. The stability of the spatial image at the listener's position is referred to as sound field coherence. Sound Field Synthesis is a technique for generation and reproduction of natural audio to overcome the limitation of classical stereophonic techniques. The position of sound in a 3D space can be indicated by viz., time difference of arrival at the ears or the ITD, intensity difference between ears or the ILD, complex distance cues such as reverberation, Head Related Transfer Functions (HRTFs) and sensory augmentation. ILD also captures the spectral differences, depending on both direction and frequency. The prominent localization cue below 1.5 kHz is the ITD while ILD accounts for localization cues at higher frequencies.

Binaural Cue Coding (BCC) [3], [4] synthesizes stereo or multichannel audio signal by approximating inter channel time difference (ICTD), inter channel level difference (ICLD) and inter channel coherence cues (ICC) of the original audio with a certain time frequency resolution. The BCC encoder extracts the binaural spatial cues from the original multichannel audio signal and is transmitted as a *BCC* bitstream along with the downmixed mono audio signal. BCC decoder reconstructs the spatial image by restoring the spatial localization cues applied on the mono audio signal. The aim is to synthesize a multichannel audio signal which is perceptually similar to the original multichannel signal. Good coding results with a BCC bitstream overhead of about 16 kbps have been variously reported [6].

3. PROPOSED CODING SCHEME

We propose a new encoding scheme called the Immersive Soundfield Rendition (ISR) system that emphasizes the accurate reproduction of the temporal envelope of the multichannel signal in multiple frequency bands. Experiments indicate that a higher fidelity reconstruction of the multi-channel soundfield is possible using this approach. The key components of this system are as follows:

- Analysis and coding of the multi-channel envelope using an over-sampled filterbank called a *Utility Filter Bank (UFB)* [15], [16].
- Techniques for the computation and efficient coding of the multichannel time-frequency envelope.
- Mechanism to create acoustic diversity between the front and surround channels; e.g., if an instrument (or vocal) with a detectable harmonic pattern is present only in the front channels then it is removed from the downmix using accurate tone detection and subtraction techniques - developed as part of the Accurate Signal Representation (ASR) technique [7], [8] - before the generation of the surround channel.
- Mechanism for estimating and incorporating suitable time delays between the front and surround channels.
- A multi-band downmixing tool that rotates the phase angle of channel component before downmixing to avoid phase cancellations.
- A technique for multi-channel recreation from stereo downmix using very low (or zero overhead) (blind upmixing). This is similar to [14] whereby an effective sound distribution to the surround channel is achieved by cross correlating the stereo channel's information, and an improved stereo image is obtained using Principal Component Analysis. PCA is used to detect the direction of stereo image, and which is then used to derive center and surround channels.

Here we describe some of the above components in more detail as well as overall structure of the ISR encoder and Decoder. The blind upmixing technique is described in more detail in Section 4.4.

Figure 1 illustrates the architecture of the ISR encoder. The heart of ISR encoder is the formation of Multi Band Temporal Envelope of the multi channel audio. This method is called Multi Band Temporal Amplitude

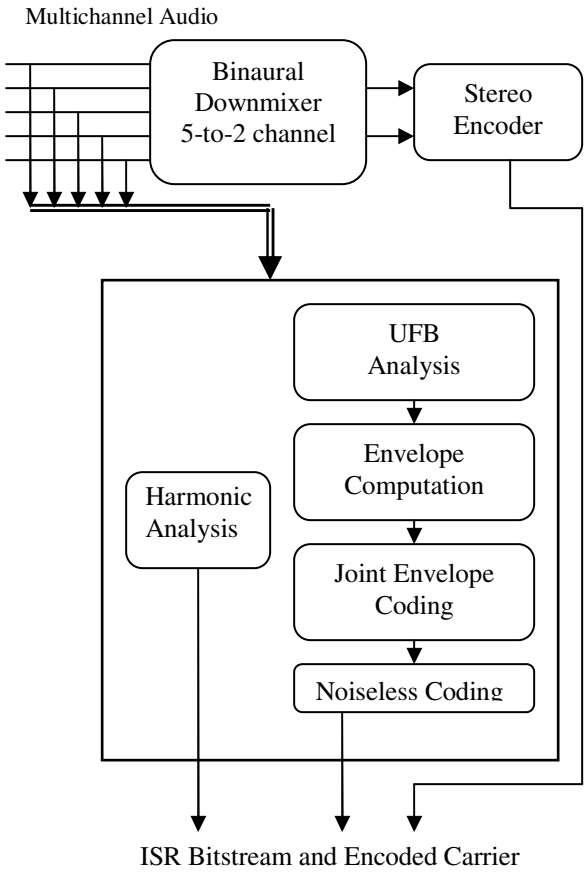

Figure 1: ISR Encoder Architecture

Coding (MBTAC) [15], [16]. A conventional stereo encoder is upgraded to a multichannel encoder utilizing an *ISR bitstream* as overhead. A binaural downmixer is used to generate a stereo audio, which behaves as the carrier audio. It is the downmix signal that is compressed and sent rather then multichannel signal. The ISR encoder in turn generates the localization cues based on the temporal envelope in the time frequency plane. Stereo coding scheme such as [9] are used to encode carrier audio signal.

The temporal envelope coding is designed as a cascade of time-frequency grouping schemes followed by an innovative joint envelope coding scheme. By operating multi band temporal envelope coding scheme on channel pairs, multichannel bit rates can be reduced

significantly. ISR coding is able to prepare the spatial image of input audio channel pairs and recreate an accurate image using synthesis driven by small coded spatial information.

The computation of temporal envelope is divided in four steps as follows:

- Utility filter bank (UFB) analysis: Analysis and coding of the multichannel envelope using an over-sampled filterbank over a non-uniform time-frequency grid with an adaptive (time-varying) resolution. The (starting) envelope resolution is adapted to the non-stationarities in the signal.

- Mapping of UFB subbands on critical bands: The first level of grouping involves the mapping of subbands on the bark scale. Auditory perceptual based grouping of frequency bands is performed to utilize the masking effects.

- Time and frequency grouping: There is a first level of grouping in critical band followed by second level grouping in time and frequency. This is based on the characteristics of the audio signal.

- Redundancy removal and coding: several strategies are used to reduce the bit overhead for coding the temporal envelope. These include joint inter-channel envelope coding, differential coding, a redundancy removal technique described in [18], and adaptive Huffman coding.

We now describe a novel technique used in ISR decoder for creating acoustic diversity to enhance the surround image and prevent potentially annoying leakage of front channel components into the surround channel. Subjective tests have indicated that removal of leaked harmonics between the front and surround channel yields improved results. Harmonic pattern of front and surround channels is determined through a pitch finding and harmonic detection algorithm. The possibility of leaked unwanted harmonics in synthesized surround audio is flagged by cross checking the front pair harmonic structure with the frequency spectrum and harmonic structure of the surround. The detected harmonic pattern, if found to be unwanted, is thereby removed from the surround channel using ASR sinusoidal synthesis and removal techniques. Figure 2 illustrates the architecture of the harmonic removal procedure.

Figure 3 illustrates the corresponding full architecture of the ISR decoder. The main blocks consist of UFB analysis, MBTAC decoding, acoustic diversity creation, MBTAC application for accurate temporal envelope synthesis and UFB synthesis.

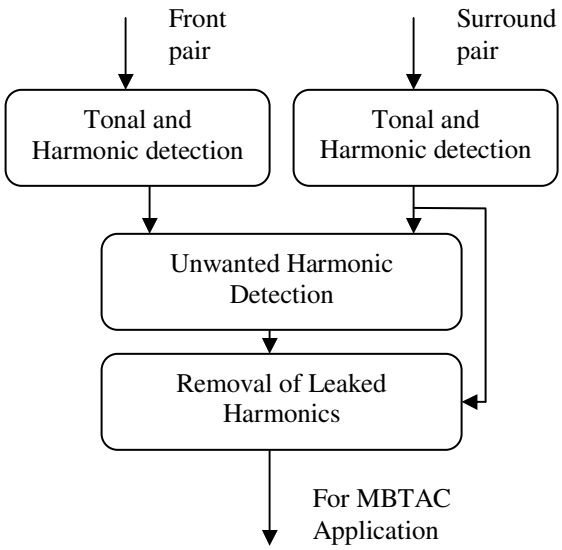

Figure 2: Acoustic Diversity Creation through Harmonic Removal

4. CODEC ARCHITECTURES

The proposed ISR System can operate in 4 different configurations. This classification is based on bit rate overhead required to code the *ISR Bitstream*. ISR system can be operated at 14 - 16 Kbps (detailed multichannel reproduction), 8 - 12 Kbps (high quality multichannel reproduction), 4 - 6 kbps (realistic multichannel), and bit rates 0-2 kbps (blind/near blind upmixing). The architecture of all configurations is explained below.

4.1. Detailed Multichannel Reproduction (14 – 16 Kbps Overhead):

14 – 16 Kbps overhead configuration is used for detailed multichannel reproduction. Temporal envelope for multichannel audio input is computed using the MBTAC algorithm explained above. The encoded data is multiplexed and transmitted in addition with encoded carrier audio file. This mode gives the best possible multichannel reconstruction. MBTAC encoding and envelope application is performed with the goal of creating the spatial image of original audio as closely as possible.

.

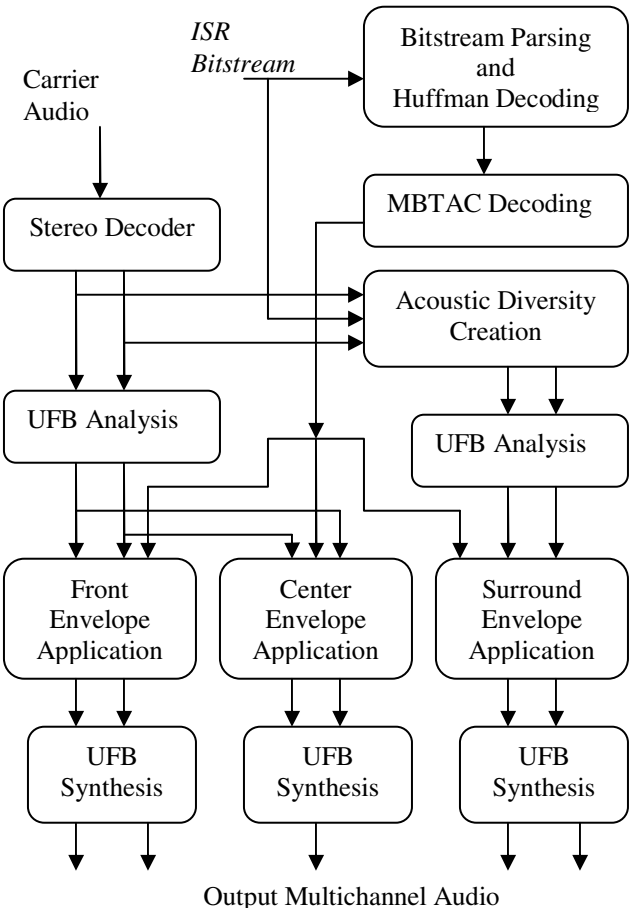

Figure 3: ISR Decoder Architecture

.

4.2. High Quality Multichannel Reproduction (8 – 12 Kbps Overhead):

8 – 12 Kbps bit overhead is obtained by computing and transmitting the MBTAC of front and surround pair only as in Figure 4. It is assumed that the center channel is dominated by front pair and can be synthetically generating from front MBTAC. Hence instead of computing the MBTAC of center the Sum MBTAC of front pair is used on downmix of carrier audio.

4.3. Realistic Multichannel (4 – 6 Kbps Overhead):

This mode has bit overhead of 4 – 6 Kbps and leads to realistic multichannel reproduction by creating delays in surround and enhancements in front channels. The ISR encoder and decoder diagrams for this mode are given in Figures 5 and 6. In this mode the MBTAC of surround pair only is transmitted. The key idea behind this scheme is to use the front pair as a carrier signal and create the center channel at the decoder as a downmix of front channels. The surround MBTAC decoding is time delayed with respect to front so as to enhance the spaciousness. The front pair may also be enhanced by decorrelation methods for better stereo image perception.

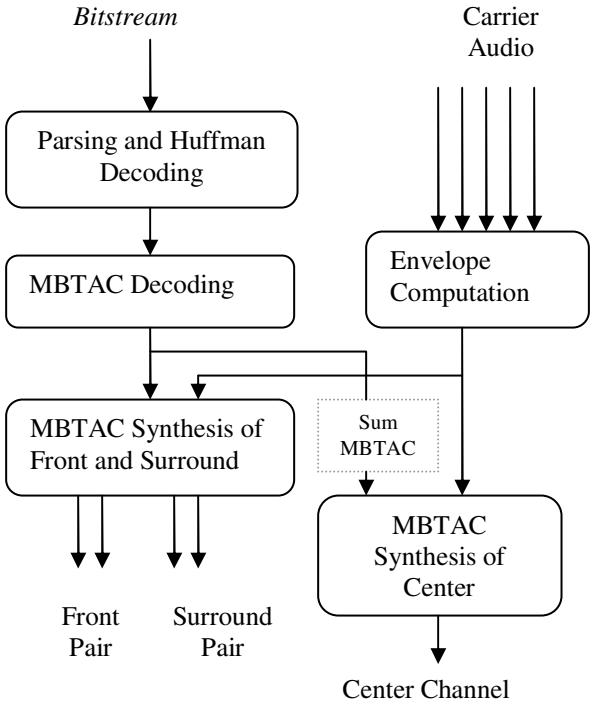

Figure 4: 8 – 12 Kbps Mode Decoder (Partial)

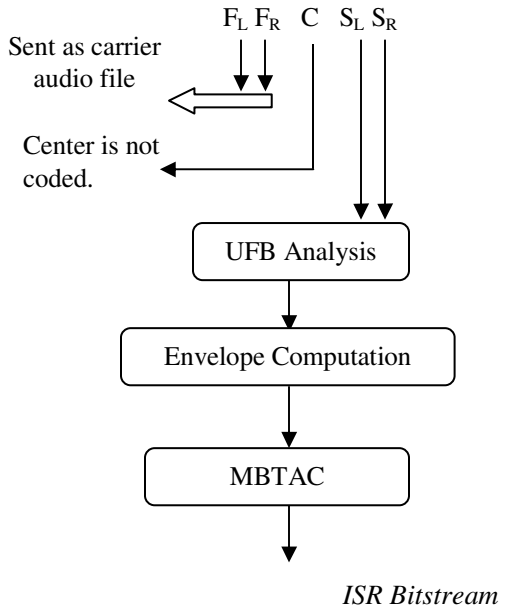

Figure 5: 4 – 6 kbps ISR encoder

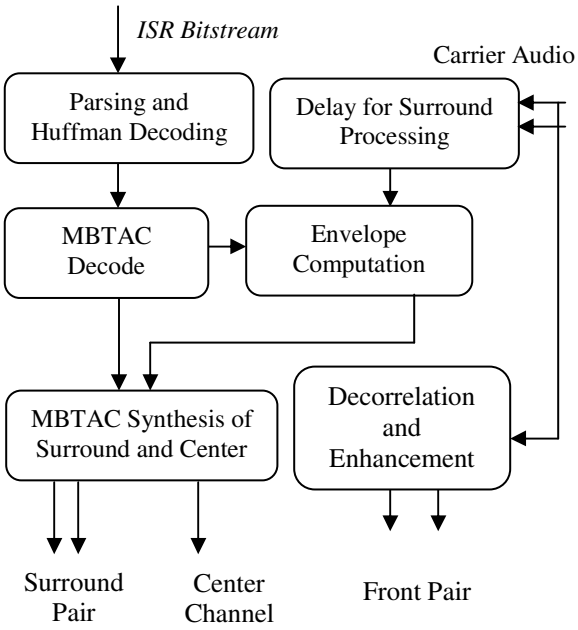

Figure 6: 4 – 6 Kbps Decoder

4.4. Blind/Near Blind Up-mixing (0 – 2 Kbps Overhead):

Blind upmixing is a process of signal format conversion from two-channel audio to five-channel audio without any overhead in addition to the coded stereo. ISR System incorporates a blind upmixing scheme which is adapted from [14]. However, we utilize a novel method for Principal Component Analysis in UFB domain as explained below.

The algorithm is illustrated in Figures 7 and 8, in which the channels are labeled as F_L (left), C (center), F_R (right), S_L (left surround), S_R (right surround), I (intensity) and S (side).

A Utility Filter Bank (UFB) is used to analyze the input audio frame. The output of any frequency band (subband) for a short-time series can be interpreted as the response of the filter bank about the center frequency for that band. To reduce the correlation between the left and right channels in the subband domain, rotational transformation [17] is used as shown in Figure 7. This can be seen as a rotation of the left and right axes to the so-called intensity and side axes. The rotation angle (α) is adapted to every block of samples and it depends upon correlation coefficients between the left and right channels. In effect, this transformation

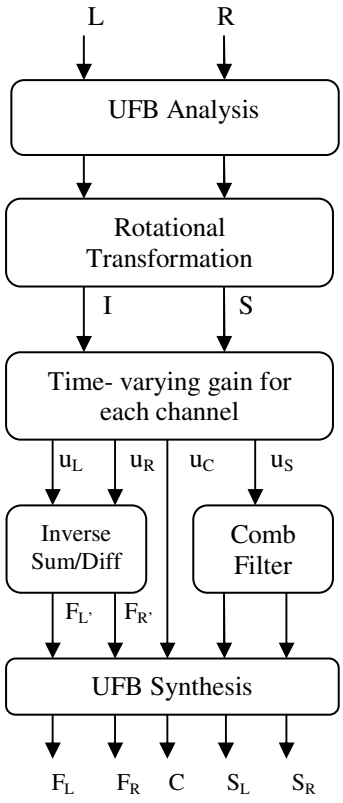

Figure 7: Blind Upmixing

behaves as Principal Component Analysis (PCA). Consider l[k] and r[k] as the left and right audio samples for k = 0, 1… N-1: N being the number of subbands. The correlation coefficients are now computed as:

$$c_{lr} = \frac{1}{N}\sum_{k=0}^{N-1} l[k]r[k]. \quad (1)$$

And

$$\tan(2\alpha) = \frac{2c_{lr}}{c_{ll} - c_{rr}} \quad (2)$$

PCA has proved to be a powerful tool to detect the direction ($\gamma(k)$) of a stereo image, which in turn can be used to derive center channel's gain and a three-dimensional mapping is used to derive a time-varying gain for mono surround channel [14]. In terms of an angle (in radians), the direction of stereo image can be computed as:

$$\gamma(k) = \arctan(\frac{w_L(k)}{w_R(k)}) \quad (3)$$

Where w_L and w_R are the weights corresponding to the left and the right channels and can be computed as:

$$w_L(k) = w_L(k-1) + \mu I(k-1)\times[I(k-1) - w_L(k-1)I(k-1)]$$
$$w_R(k) = w_R(k-1) + \mu I(k-1)\times[S(k-1) - w_R(k-1)I(k-1)] \quad (4)$$

Here μ is the step size, and it must satisfy the following condition to make the algorithm stable:

$$0 < \mu < \frac{2}{(I \times I) + (S \times S)} \quad (5)$$

When the amount of the intensity signal equals or almost equals that of the side signal, an ambiguity appears since there is no way of determining the direction vector uniquely, so extra information is needed when dealing with this sort of ambiguity. The use of both the direction of stereo image and correlation coefficient is necessary to obtain multichannel audio signals. The angle $\beta(k)$ can be defined to represent the actual surround information by means of the adaptive correlation coefficient $\rho_o(k)$ [14], for example, by using the expression

$$\beta(k) = \arcsin[1 - \rho_o(k)] \quad (6)$$

The signals for the left, right, center and mono surround can be obtained by the following equations:

$$u_L(k) = c_L(k)\times I(k) + \cos(\beta(k))\times w(k)\times S(k)$$
$$u_R(k) = c_R(k)\times I(k) + \cos(\beta(k))\times w(k)\times S(k)$$
$$u_C(k) = 2\times w_L(k)\times w_R(k)\times I(k) \quad (7)$$
$$u_S(k) = \sin(\beta(k))\times S(k)$$

And

$$c_L(k) = \begin{cases} -C_{LR}(k), C_{LR}(k) < 0 \\ 0, otherwise \end{cases}$$

$$c_R(k) = \begin{cases} C_{LR}(k), C_{LR}(k) \geq 0 \\ 0, otherwise \end{cases}$$

$$C_{LR} = w_R(k)\times w_R(k) - w_L(k)\times w_L(k)$$

The Lauridsen [19] decorrelator is used to obtain stereo surround because of its simplicity. this decorrelator can be viewed as two FIR comb filters (h_L and h_R) with two taps each for surround left and surround right.

Listening tests have confirmed that audio quality improves if we apply inverse Sum/Difference technique on front stereo instead of inverse rotational transformation, i.e. the left channel audio $F_{L'}$ is obtained by 0.5 * (u_L + u_R) and the right channel audio $F_{R'}$ by 0.5 * (u_L - u_R).

5. RESULTS

Audio tests were performed using different modes of operation on a set of standard audio corpus which are critical for multichannel audio evaluation like, clapping in an auditorium setup. Of all different modes the detailed multichannel reproduction mode is perceptually superior at the expense of bit overhead. Threshold values and transition frequencies have been tuned for all individual modes based on perceptual hearing.

For detailed and high quality multichannel reproduction modes (mode 1 and mode 2) the image is stable and the overall multichannel perception is improved when increasing the time resolution of MBTAC. In high quality multichannel reproduction mode the center channel image is more domination by front because it is generated form down mix of front pair. One of the attributes for improved image is the capability of ISR

system to create acoustic diversity by removing the leaked harmonics from surround channel.

The realistic multichannel reconstruction mode (mode 3) is dominated by front pair. It is found that perception of surround effects is dependent on enhancement of front pair and the time delay applied to surround pair. Applying time delay to surround decreases the front dominance on surround leading to a better overall multichannel perception on low bit overhead.

All multichannel audio samples categorized on their bit rates will be made available on www.atc-labs.com/isr.

6. CONCLUSIONS

Different operating modes of ISR system have been tested. A scheme for blind upmixing and a new structure for harmonic diversity have been included to enhance the multichannel audio image. In blind upmixing, cross correlation technique and principal component analysis make the audio better in terms of sound distribution. The above techniques have yielded better results as mentioned above. More intense subjective based tests are presently being done for fine tuning of multiple parametric values involved in MBTAC algorithm. More professional multichannel listening tests will be conducted and be made available on www.atc-labs.com/isr.

7. REFERENCES

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

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

[3] F. Baumgarte and C. Faller, ``Binaural Cue Coding Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.

[4] C. Faller and F. Baumgarte, ``Binaural Cue Coding - Part II: Schemes and applications, "*IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.

[5] D. Sinha, "Technique for parametric coding of a signal containing information", *U.S. Patent No.* US6539357, Filed 1999 (Published March, 2003).

[6] Christof Faller and Frank Baumgarte, "Binaural Audio cue coding Applied to stereo and Multichannel Audio Compression" in the preprints of 112th Convention of the Audio Engineering Society, München, Germany, 2002.

[7] 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.

[8] A. J. S. Ferreira and D. Sinha, "Accurate Spectral Replacement", in the Preprint of 118th Convention of the Audio Engineering Society, Barcelona, Spain. Convention Paper 6383, May 2005.

[9] Deepen Sinha and Anibal Ferreira "A New Broadcast Quality Low Bit Rate Audio Coding Scheme Utilizing Novel Bandwidth Extension Tools," *119th Convention of the Audio Engineering Society*, October 2005. Paper 6588.

[10] Anibal J. S. Ferreira and Deepen Sinha, "Audio Communication Coder," *in the preprints of* 120*th Convention of the Audio Engineering Society*, May 2006.

[11] J. D. Johnston, D. Sinha, S. Dorward, and S. R Quackenbush, "AT&T Perceptual Audio Coding (PAC)," *in AES Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. 1996, pp. 73-82.

[12] Marina Bosi et al., "ISO/IEC MPEG-2 Advanced Audio Coding," *101st Convention of the Audio Engineering Society*, November 1996, Preprint n. 4382.

[13] Mark Davis, "The AC-3 Multichannel Coder," *95th Convention of the Audio Engineering Society*, October 1993, Preprint n. 3774.

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

[15] Deepen Sinha, Anibal Ferreira, and, Deep Sen "A Fractal Self-Similarity Model for the Spectral Representation of Audio Signals," *118th Convention of the Audio Engineering Society*, May 2005, Paper 6467.

[16] 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.

[17] 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.

[18] Raghuram A., Harinarayanan. E.V, Anibal Ferreira, and Deepen Sinha, "A Novel Very Low Bit Rate Multi-Channel Audio Coding Scheme using Accurate Temporal Envelope Coding and Signal Synthesis Tools", *In the Preprints of AES 121st Convention.*

[19] H. Lauridsen, "Experiments Concerning Different Kinds of Room-Acoustics Recording", Ingenioren, vol. 47, 1954.