Audio Engineering Society

Convention Paper

Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

Subjective Evaluation of Immersive Sound Field Rendition System and Recent Enhancements

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@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. In particular we present detailed subjective result data benchmarking the ISR system in comparison to MPEG Surround and also characterizing the audio quality level at different sub-modes of the system. We also describe enhancements to various algorithmic components in particular the blind 2-to-5 channel upmixing algorithm and describe a novel scheme for providing enhanced stereo downmix at the receiver for improved decoding by conventional matrix decoding systems.

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 & gaming. Plans are currently underway to incorporate such schemes in Satellite Digital Audio Radio (SDAR) and other digital broadcast systems. We have recently introduced a novel technique for the parametric coding of multi-channel audio called the Immersive Sound-field Rendition (ISR) System [1, 2] which is capable of high quality parametric multi-channel coding in the range of 0-12 kbps. In this paper we undertake a comparative evaluation of the ISR system using subjective audio quality tests. For the purpose of comparison, the MP3 Surround [12] and MPEG Surround [13, 16] techniques are used.

The organization of the rest of the paper is as follows. An overview of the ISR system and ISR features are presented in Section 2. Section 3 discusses the various details about the subjective test methodology, and detailed results. Enhancements in ISR system for Matrix decoding are discussed in section 3 followed by conclusion in section 4.

2. ISR SYSTEM OVERVIEW

Immersive Sound-field Rendition (ISR) System is a novel 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 ISR Encoder (which generates an ISR bit stream). The ISR Bitstream consists of localization cues [3, 4] of the original multichannel audio generated on a time-frequency grid of adaptive resolution. Figure 2 shows the corresponding ISR Decoder, it uses decoded stereo carrier and the ISR bitstream information to synthesize 5 channel surround audio output.

2.1. ISR Modes of Operation

The ISR system offers following 4 modes of operation:

- *Mode 1*: Detailed multi-channel reproduction with 12- 14 kbps overhead; including a 12 kbps Constant Bit Rate (CBR) option.
- *Mode 2*: High quality multi-channel reproduction with 8-10 kbps overhead; including an 8 kbps CBR option.
- *Mode 3*: Realistic multi-channel reproduction with 4-6 kbps as overhead; including a 6 kbps CBR option.
- *Mode 4*: Blind/Near Blind upmixing with a 0-2 kbps overhead.

2.2. ISR Architecture

The ISR technique is based upon the following core algorithmic components [1, 2]:

- Analysis and encoding of accurate multi-band temporal envelope. The envelope is computed by analyzing the signal using an over-sampled, high resolution *Utility Filter Bank (UFB)* [5] and computing a suitable time-frequency envelope based upon a signal adaptive resolution. A slew of efficient coding techniques are then employed to jointly encode the multi-channel time-frequency envelopes [5]

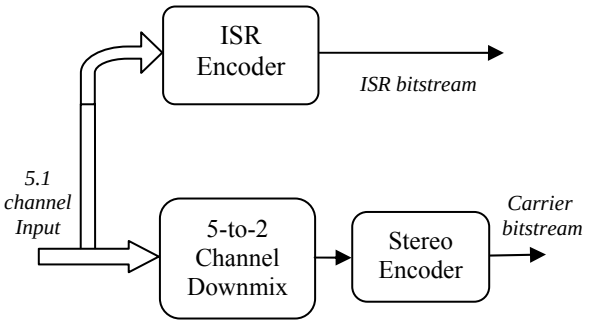

Figure 1 : Architecture of ISR Encoder

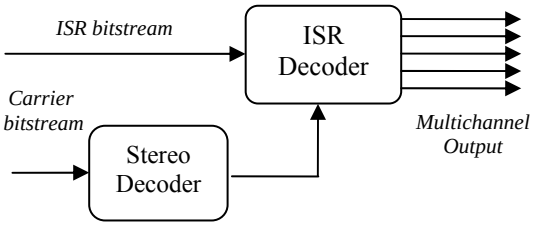

Figure 2 : Architecture of ISR decoder

- 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[6, 7] before the generation of the surround channel.
- Use of a new phase compensated stereo down-mixing scheme. The scheme compensates phases in the ODFT [2] domain using an adaptive estimation algorithm.
- A new blind 2-to-5-channel up-mixing algorithm for *Mode 4* which emphasizes the enhancement of any detected smooth image movements in the stereo downmix to generate a high level of recreated spaciousness.

The structure of ISR Encoder and Decoder in *Modes 1-3* is detailed in Figures 3 and 4 respectively. The blind 2-to-5 channel upmixing algorithm generates the surround channel by adaptively mixing three components – viz. ambience, reverberated dominant PCA signal and an Image Movement Spatialization Component (IMSC). The third component, IMSC, which is a weighted version of original carrier audio, is based on the rate of change of the stereo angle of the dominant signal component; its inclusion increases the sense of spaciousness or immersiveness in the blind ISR system mode. The structure of the blind upmixing technique is shown in figure 5.

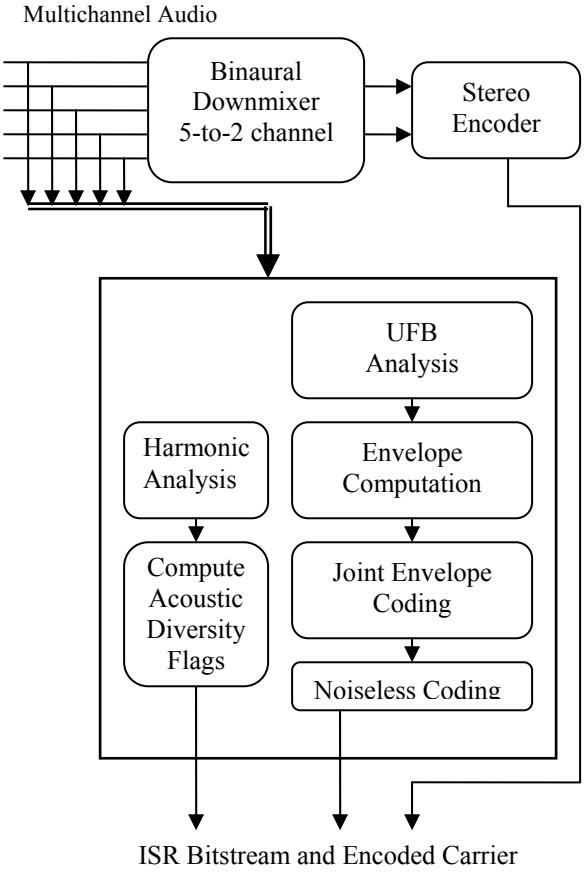

Figure 3: ISR Encoder Architecture

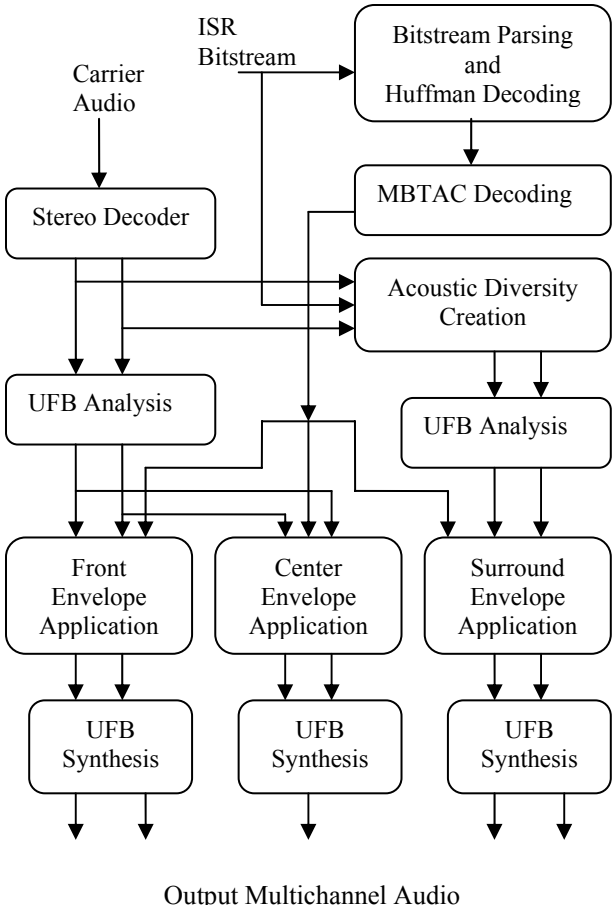

Figure 4: ISR Decoder Architecture

2.3. Current Status

The proposed system was previously compared in informal subjective tests to the MP3 Surround system [12], which is based on the Spatial Audio Coding/Binaural Cue Coding (BCC) [8, 9, 10] techniques. In terms of the overall quality informal listening by trained listeners indicated 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 the combined 5 channel coding rate of 128 kbps). We have now finished formal blind subjective evaluation using a pool of 12 critical listeners. The purpose of these tests was two-fold. Firstly, we wanted to compare and characterize the ISR systems using the MP3 Surround system and MPEG Surround system [11, 16] as a benchmark.

Secondly, the test results were expected to quantify the relative loss is the perceived quality as the bit rate of the ISR system is lowered. Two sets of detailed subjective test data are presented and analyzed in the next section.

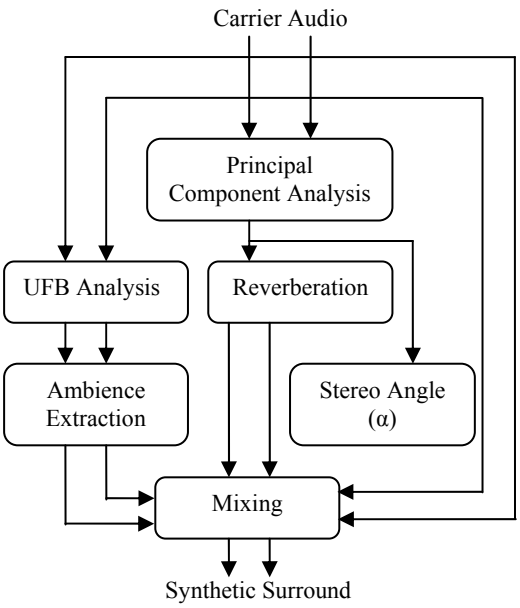

Figure 5: Blind Upmixing Scheme

3. SUBJECTIVE EVALUATION

3.1. Selection of Benchmarks

- For the purpose of evaluating Modes 1-3 of the ISR system, MP3 Surround and MPEG Surround technologies are used as benchmarks. MP3 Surround encoder/decoder is available from [12] and it can operate for 44100 and 48000 kHz sampling frequencies in the bit rate range of 128 Kbps to 192 Kbps. We chose 128 kbps MP3 Surround as one of the benchmarks. As a second benchmark, the MPEG Surround [13] system was used. MPEG Surround employs an evolution of the BCC coding technique used in MP3 surround. Specifically the Reference Model 0 of the MPEG Surround was used in these tests.

- Mode 4, i.e. 2-to-5 Channel Blind Upmixing Scheme was tested in comparison with Dolby ProLogic-II [17] and Creative CMSS [18] technology. Both ProLogic-II decoding and Blind Upmixing are applied on same downmix carrier.

3.2. Test Preparation

The test preparation involved selection of critical multichannel audio samples and the listeners. Initially 20 samples are selected for training the listeners, out of which six were chosen for the final evaluation. The list and the type of samples selected are given in Table 1. Original and ISR coded samples on different bit rates were used to make listeners comfortable with multichannel audio. A number of listener fatigue related factors were considered - the audio sample length was chosen to be between 20 to 30 seconds, the audio levels were equalized for the entire test stimuli corresponding to a particular sample, the range of artifacts were introduced to the listeners during training session (including imaging artifacts). As noted above a pool of 12 critical listeners was used in this test.

Table 1 : Test Samples

Name	Description
River	Female voice in center and birds chirping in other channels
Roxy	Song with rapidly moving and split vocals in the front channels
Glock	Glockenspiel and timpani
Genzmer	Music with strong attacks in front and ambience in surround
Rock	Music in front and center with clapping in surround
Seawash	Moving and breaking ocean waves with rapid image movement
For Blind Upmixing	
Canyon	Image movements and natural sounds
Black Water	Smooth image movement between left and right channel

3.3. Test Methodology

We used a testing methodology based on the MUSHRA intermediate audio quality evaluation standard [14]. In this methodology the listeners are presented with a

reference sample along with versions of the same sample processed by each of the systems under test. The listener is allowed to listen to either the reference of either of the test versions at will as many times as desired. The test versions include a hidden reference as well as a low anchor. The listener is asked to score each of the test samples on a 5 point scale:

Score	Rating
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

3.4. Selection of Tests

3.4.1. Test I: Surround Image Quality Test for the ISR System at 12 kbps and 8 kbps

In this test the ISR system was tested in two modes for image quality. In this test the stereo downmix (carrier) was not quantized (as shown in Figure 6 below). The quality was compared against the MPEG surround system (also with uncompressed downmix). In summary the tests systems consisted for the following (the reference for the test was the original 5 channel recording):

- ISR Mode 1: 12 kbps CBR
- ISR Mode 2: 8 kbps CBR
- MPEG Surround
- Low Anchor: Mono Downmix
- Hidden Reference.

3.4.2. Test 2: Test for Complete Coding Schemes Employing ISR

In this case ISR Mode 1 operating at 12 Kbps CBR was used in conjunction with a core stereo codec. In particular the TeslaPro Codec [15] was utilized as the core coding scheme operating at two different bit rates (56 kbps and 88 kbps). This is illustrated in Figure 7. The MP3 Surround operating at 128 kbps was used as the third system under test. Once again the original multi-channel audio was used as the reference. The test systems therefore consisted of the following:

- TeslaPro 56 kbps + 12 kbps ISR (Total bit rate 68 kbps)
- Tesla Pro 88 kbps + 12 kbps ISR (Total bit rate 100 kbps)
- MP3 Surround at 128 kbps
- Low Anchor: Mono Downmix
- Hidden Reference

3.4.3. Test 3: Comparison of ISR Blind Upmixing with Existing Matrix Schemes

The mode 4 of the ISR System, Blind Upmixing, was tested in comparison with Dolby ProLogic-II and Creative CMSS technology. Both ProLogic-II decoding and Blind Upmixing were applied on same downmixed carrier. In this case no comparative scoring was used but rather the subjects were asked to comment on the relative audio quality.

3.5. Test Results and Analysis

Test 1: ***Surround Image Quality Test***

Detailed scores are presented in Figure 8. It shows the average score of 12 trained listeners with 95% confidence level for 6 subjects. The ISR 12 Kbps (green marks) scored higher for all audio material. In case of *River*, where the dominating audio is female voice in the center channel, both ISR and MPEG Surround scores are comparable. In *Genzmer*, *Seawash*, *Glock* and *Rock*, MPEG Surround scored significantly lower than ISR. The surround image was found to be narrower and not faithful to the original audio. In test samples, such as *Genzmer*, listeners reported a loss of imaging in addition to annoying artifacts in the case of MPEG surround. The ISR 12 kbps mode was found to present superior fidelity and imaging in comparison.

Figure 9 illustrates the mean scores of all the systems under consideration in Test 1 averaged across all audio samples and listeners. The ISR 12 kbps mode performed the best with an average rating of 4.01. The average score of the ISR system in mode 8 kbps was 3.57. In comparison, the MPEG surround scores 3.07.

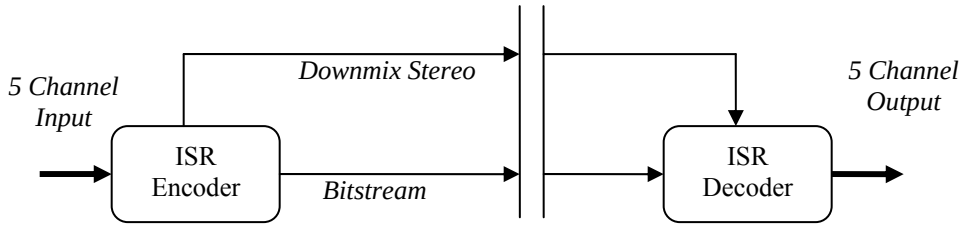

Figure 6: System used in Surround Image Quality Test

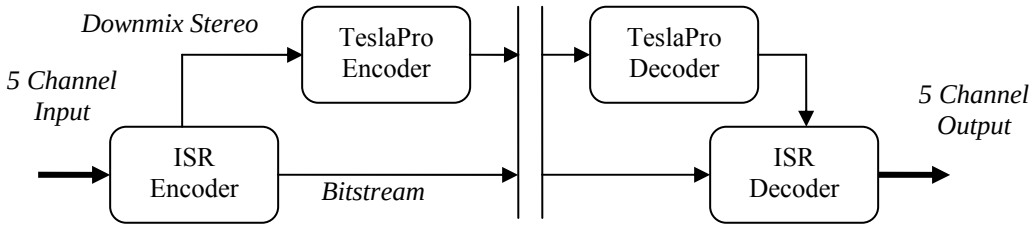

Figure 7: System used in Full Coding Scheme Test

Test 2: ***Audio Quality Test for Complete Coding Schemes***

For the music signals MP3 Surround at 128 Kbps scored lower than ISR at 100 Kbps. Listeners reported a loss of image quality and missing signatures (e.g. clapping in *Rock* and the intensity of attacks in *Genzmer*). It was found that the audio image had shrunk in comparison to the original. This was clearly evident in samples such as *Glock* and *Roxy*. In other samples such as *Seawash* small signatures of falling water are missing in case on MP3 Surround. The ISR system's audio quality remained faithful to the original perhaps primarily due to the superior downmix and coding schemes.

Test 3: ***Descriptive Comparison of the Blind Upmix Schemes***

We have compared ISR Mode 4 with existing schemes such as CMSS and Dolby Pro Logic II. The surround generation methods used in these schemes are different. ProLogic II uses delayed version of ambience signal and ISR uses a combination of ambience, reverberation and IMSC. We applied ProLogic II decoding and ISR Mode 4 on the same audio material. The dominance of the front channels in Pro Logic II is readily noticed where as ISR produces a balanced and spacious output. Better surround image quality is achieved by applying ISR on the *Blackwater*. The image movement from left to right channels is extended towards surround channels in ISR which is not the case in other decoders. In the case of the sample *Canyon*, ISR displays a spacious sounding realistic surround upmix

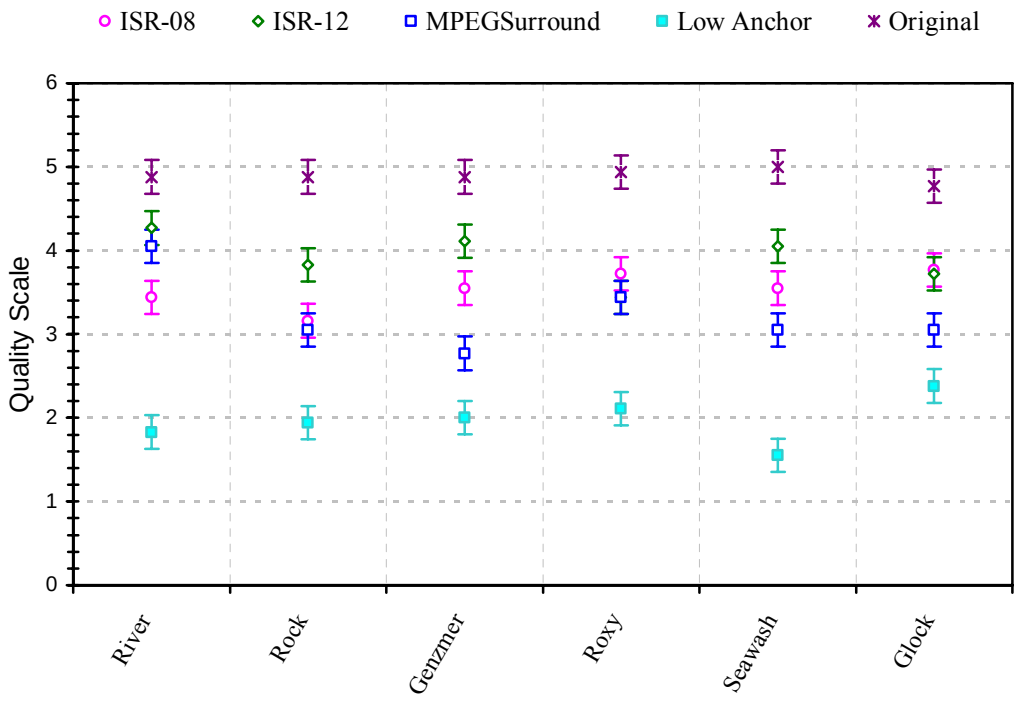

Figure 8: Surround Image Quality Test

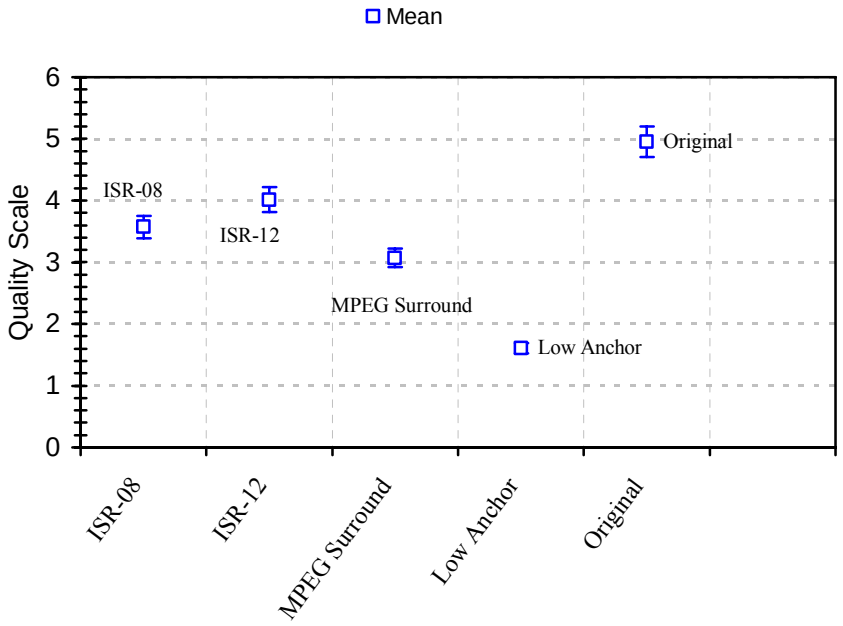

Figure 9: Mean Scores of Surround Image Quality Test

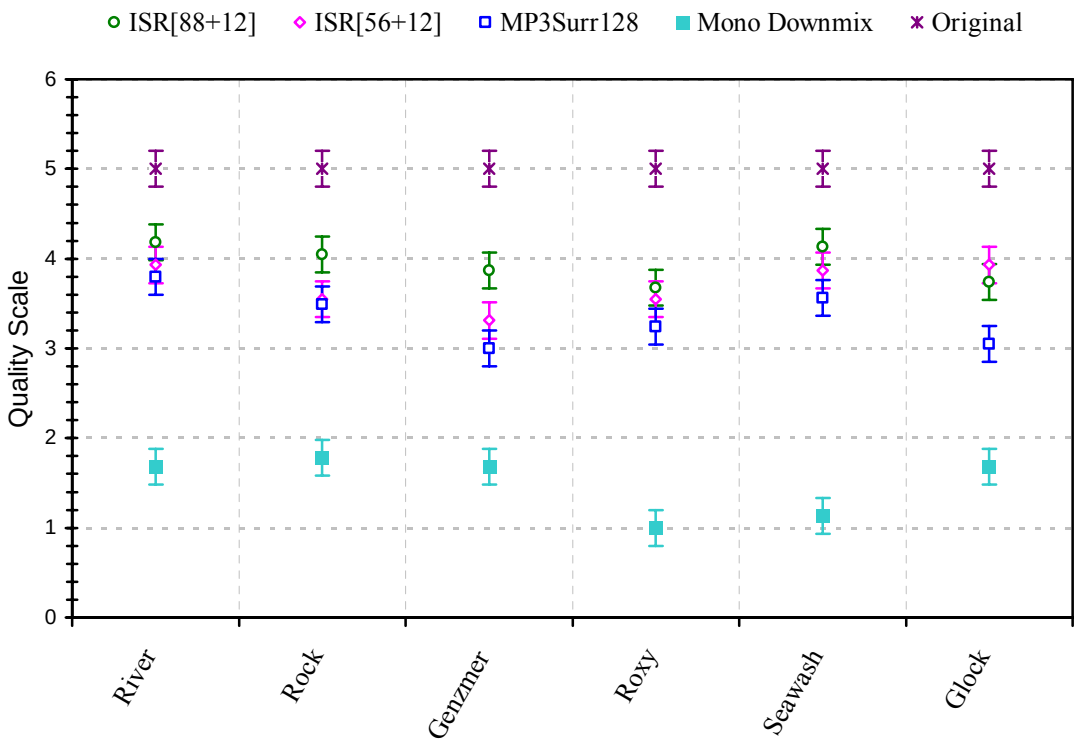

Figure 10: Audio Quality Test for Compete Coding

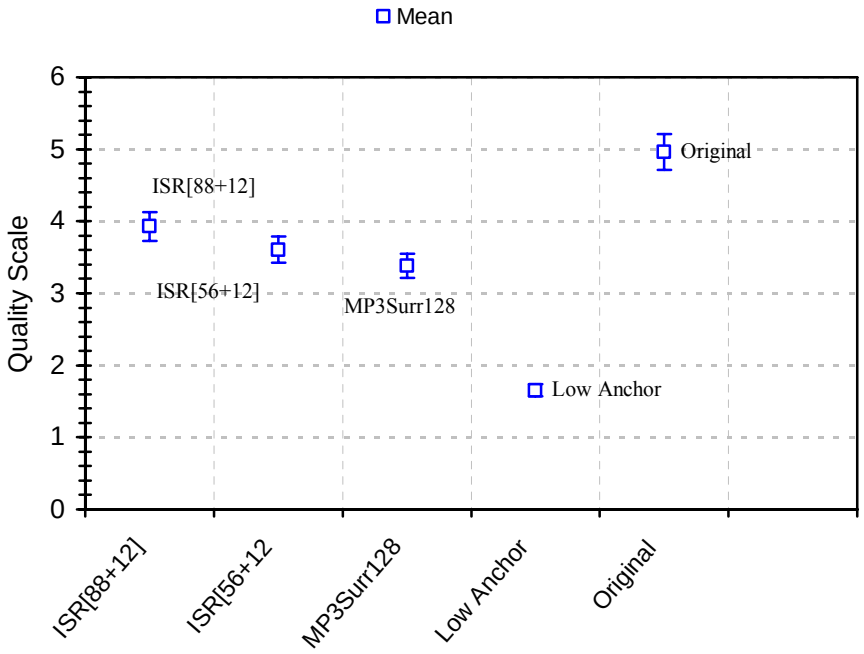

Figure 11: Mean of All Scores in Audio Quality Test

4. OTHER RECENT ADDITIONS TO ISR SYSTEM

In addition to the modes presented above, the functionality of working with Matrix Decoder has now been added to the ISR System. The idea is to enhance the 5-to-2 channel downmix using the ISR bitstream information before sending it to a Matrix Decoder like ProLogic II. As shown in the Figure 7, ISR based matrix encoding block is added just prior to the Matrix Decoder to generate a high quality stereo downmix with the goal that eventual multichannel audio (after matrix decoding) has better image characteristics as compared to the traditional scheme.

The MBTAC parameters are found to be particularly useful in generating a stereo downmix from the synthesize channels. This downmix is used in matrix decoder to generate

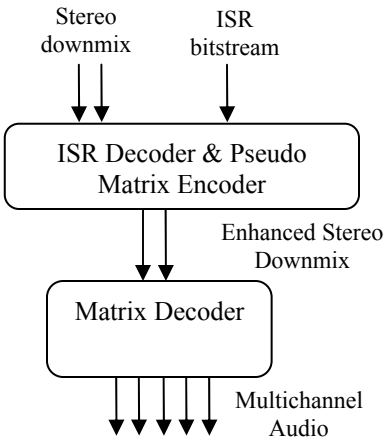

Figure 12: Enhanced Matrix Functionality

Intuitively, this approach is better in the sense that a general matrix decoder uses phase information and ambience signal with few milliseconds delay to recreate the surround channels; while downmix formed using ISR carries more accurate information directly derived from the original signal.

5. CONCLUSION

We have presented subjective test results for the Immersive Soundfield Rendition system. Three tests have been conducted to characterize the quality of the ISR system viz. Surround image quality test, ISR system test with complete coding and a comparison of blind upmix schemes. Listeners reported a higher preference to ISR coded audio in comparison to other popular schemes such as MP3 Surround and MPEG Surround. The surround image quality was found to have a higher fidelity to the original in comparison with the other schemes. The ISR system with the carrier coded at 88 kbps and 56 kbps also performed better than the MP3 surround coded at a net bit rate of 128 kbps in terms of image quality and stability. The ISR blind upmix scheme has also been compared against the Dolby ProLogic II and the Creative CMSS technologies and compares favorably against these technologies.

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*", in the preprints of 121st AES Convention October 5–8, 2006 San Francisco, CA, USA.

[2] Chandresh Dubey, Richa Gupta, Deepen Sinha and Anibal Ferreira, "*New Enhancements to Immersive Sound Field Rendition (ISR) System*", in the preprints of 122nd Convention 2007 May 5–8, Vienna, Austria.

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

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

[5] 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.

[6] 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.

[7] Anibal J. S. Ferreira, "*Perceptual Coding of Harmonic Signals*", in the preprints of 100th Convention of the Audio Engineering Society, 1996.

[8] C. Faller and F. Baumgarte, "*Binaural Cue Coding Applied to Stereo and Multichannel Audio Compression*", in the preprints of the 112th Convention of the Audio Engineering Society, 2002.

[9] 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.

[10] 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.

[11] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, P. Kroon, "*Spatial Audio Coding: Next-Generation Efficient and Compatible Coding of Multi-Channel Audio"*, in the preprints of 117th the Audio Engineering Society Convention, San Francisco 2004

[12] MP3 Surround Executable, http://www.iis.fraunhofer.de/EN/bf/amm/mp3sur/index.jsp

[13] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch1, K. Kjörling, E. Schuijers, J. Hilpert, F. Myburg, "*The Reference Model Architecture for MPEG Spatial Audio Coding*", in the preprints of 118th Convention of the Audio Engineering Society, 2005, May 28–31 Barcelona, Spain.

[14] ITU-T Recommendation BS. 1534-1, *"Method for Subjective Assessment of Intermediate Sound Quality (MUSHRA)"*, International Telecommunications Union, Geneva, Swizerland, 2001.

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

[16] Jurgen Herre, Kristof Kjorling, Jeoren Breebart, Christof Faller, Sascha Disch, Heiko Purnhagen, Jeroen Koppens, Johannes Hilpert, Jonas Roden, Werner Oomen, Karsten Linzmeier and Kok Seng Chong, "*MPEG Surround – The ISO MPEG Standard for Efficient and Compatible Multi-Channel Audio Codin"*, in the preprints of the 122nd Convention of the Audio Engineering Society, May, 2007.

[17] Dolby Prologic II, http://www.dolby.com/consumer/technology/prologic_II.html

[18] Creative Multi-Speaker Surround, http://www.creative.com/products/speakers/tech/?id=62790