A SPATIAL SQUEEZING APPROACH TO AMBISONIC AUDIO COMPRESSION

Bin Cheng, Christian Ritz and Ian Burnett

Whisper Laboratories, University of Wollongong, Wollongong, NSW, Australia
bc362@uow.edu.au, critz@uow.edu.au, ianb@uow.edu.au

ABSTRACT

Spatially Squeezed Surround Audio Coding (S^3AC) has been previously shown to provide efficient coding with perceptually accurate soundfield reconstruction when applied to ITU 5.1 multichannel audio. This paper investigates the application of S^3AC to the coding of Ambisonic audio recordings. Traditional Ambisonics achieve compression and backward compatibility through the use of the UHJ matrixing approach to obtain a stereo signal. In this paper the relationship to Ambisonic B-format signals is described and alternative approaches that derive a stereo or mono-downmix signal based on S^3AC are presented and evaluated. The mono-downmix approach utilizes side information consisting of spatial cues that are quantized based on novel source localization listening experiments. Objective and subjective tests demonstrate significant improvements in the localization of sound sources resulting from decoding the compressed B-format signals to a 5.1 speaker playback.

Index Terms— Audio Coding, Audio Systems

1. INTRODUCTION

There have been many recent techniques proposed for *Spatial Audio Coding* (SAC) [1] that have shown great improvements in coding efficiency and perceptual quality compared to earlier techniques [1, 2]. In these existing techniques, spatial audio is represented by a stereo (or mono) downmix signal plus side information containing cues representing the inter-channel mathematical relationships of the multichannel audio signals e.g. phase/level difference and correlation. Recently, *Spatially Squeezed Surround Audio Coding* (S^3AC) [3, 4] has been proposed as an alternative approach to spatial audio coding. Rather than other approaches that derive relationships between individual channels, S^3AC is based on analysis of the localized soundfield sources and 'squeezing' them into a stereo space. Results for coding of ITU 5.1 multichannel [5] audio signals have shown significant advantages of this approach in preserving the correct sound localization information [3, 4].

In this paper, S^3AC is applied to the compression of Ambisonics recordings of spatial audio. Ambisonics is a widely used format in professional audio studios that allows accurate reproduction of two or three dimensional sound over any speaker layout [6]. Ambisonics signals are traditionally recorded as B-Format signals, which require three full bandwidth channels representing the directional information in the 3D Cartesian coordinates and one channel representing the omnidirectional sound pressure [6]. For compression and backward compatibility, conventional Ambisonic coding uses a matrix approach called UHJ [7] to downmix the B-Format signals to stereo.

In this paper, an alternative compression and downmixing approach based on S^3AC is investigated and the relationship to Ambisonics will be described. Also presented is a technique for representing the 2D components of the Ambisonic signals with a mono-downmix signal representing the sound field and frequency dependent quantized spatial cues representing the location of each sound source. Efficient spatial cue compression is achieved using a novel variable bit rate scheme based on listening tests evaluating the location-dependent perception of spatial sound sources.

Section 2 reviews S^3AC applied to coding ITU 5.1 multichannel audio while Section 3 presents the new application of S^3AC to the compression of Ambisonic signals. Experimental results including localization dependent quantization and S^3AC compression of Ambisonics are presented in Section 4 with conclusions presented in Section 5.

2. S^3AC APPLIED TO AN ITU 5.1 CHANNEL SIGNAL

S^3AC [3, 4], described previously for coding ITU 5.1 multichannel audio signals, achieves compression by exploiting the localization redundancy of the human auditory system [8]. In [4], a compressed (squeezed) stereo sound field was demonstrated as being able to carry the perceptual localization information of a 360° horizontal sound scene without side information. To apply the squeezing approach to 5.1 recordings, the algorithm (illustrated in Fig.1) is based on the assumption that each frequency bin in the soundfield contains just one virtual source (this is similar to other spatial coding approaches). The azimuth θ_k of that frequency bin is estimated by analyzing the energy of the frequency components of a pair of speaker signals A_k^1, A_k^2 using an inverse amplitude panning law, given by:

$$\theta_k = \arctan\left[\frac{A_k^1 - A_k^2}{A_k^1 + A_k^2} \cdot \tan(\psi_{12})\right] \qquad (1)$$

where k is frequency index and ψ_{12} is the azimuth separation between the two speakers. Encoding is achieved by a linear azimuth mapping approach that re-pans each frequency dependent source from the (5 channel) 360° surround soundfield into a (stereo) 60° squeezed soundfield. Decoding is achieved by reversing the soundfield squeezing process. In addition, when coding complex sound environments containing discriminated sound sources with coincident time-frequency components, side information can be added to S^3AC to further improve localization accuracy [3]. This side information is derived from the S^3AC frequency-azimuth analysis and directly represents the source localization information of each frequency bin, θ_k.

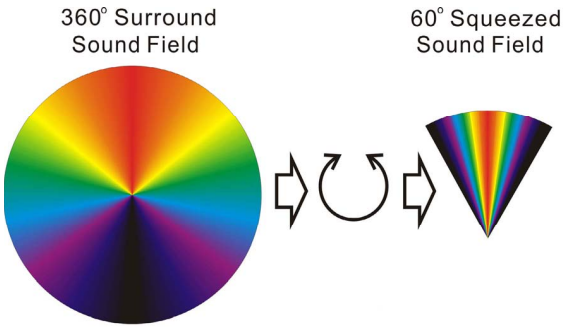

Fig. 1. The Squeezing Approach of S^{3}AC

3. S^{3}AC APPLIED IN AMBISONICS SIGNAL

Ambisonics [6], introduced in the 1970's, is known as one of the best spatial audio recording techniques and provides excellent soundfield and source location recoverability. It is shown in this section that the localization principle of a common Ambisonics playback layout can be derived into pure amplitude panning. Based on this result, the compression of Ambisonic signals using S^{3}AC will be described. Two types of S^{3}AC compression of Ambisonics B-Format signal are introduced: stereo downmixing and mono downmixing with side information.

3.1. Amplitude Panning and Ambisonics Localization

First Order Ambisonics soundfield microphones generate four-channel B-Format and the constituent WXYZ channels are related to source azimuth and elevation according to:

$$W = S/\sqrt{2},\ \ X = \cos(\theta)\cdot\cos(\alpha)\cdot S$$
$$Y = \sin(\theta)\cdot\cos(\alpha)\cdot S,\ \ Z = \sin(\alpha)\cdot S \qquad (2)$$

where S is the source and θ and α are the source azimuth and elevation respectively. When reproducing the B-Format signal set over speakers on a sphere, the speaker feed signal F is calculated according to the speaker azimuth φ and elevation β and a directivity factor d, such that [9]:

$$g_w = \sqrt{2},\ g_x = \cos(\varphi)\cdot\cos(\beta),$$
$$g_y = \sin(\varphi)\cdot\cos(\beta),\ g_z = \sin(\beta), \qquad (3)$$
$$F = 0.5\cdot\left[(2-d)g_w W + d\left(g_x X + g_y Y + g_z Z\right)\right]$$

Considering two channels on the horizontal surface (i.e. β=0°) with azimuth $\pm\varphi$, and substituting Eq. (2) into Eq. (3), the resulting speaker feed signals are found to be:

$$F_1 = 0.5\left[(2-d)S + d\left(\cos(\varphi)\cos(\theta)S + \sin(\varphi)\sin(\theta)S\right)\right]$$
$$F_2 = 0.5\left[(2-d)S + d\left(\cos(\varphi)\cos(\theta)S - \sin(\varphi)\sin(\theta)S\right)\right] \qquad (4)$$

Or in fractional form:

$$\frac{F_1 - F_2}{F_1 + F_2} = \frac{d\sin(\varphi)\sin(\theta)}{(2-d) + d\cos(\varphi)\cos(\theta)} \qquad (5)$$

For $d = 2\cos(\varphi)/\left(\cos(\varphi) + \cos(\theta) - 2\cos(\theta)\cos^2(\varphi)\right)$, Eq. (5) can be expressed in the form of amplitude panning, such that:

$$F_1 - F_2 / F_1 + F_2 = \tan(\theta)/\tan(\varphi) \qquad (6)$$

While the value of d is dependent on both speaker layout φ and source azimuth θ, in the most commonly used loudspeaker layout in Ambisonics, where four speakers are placed symmetrically at ±45° and ±135°, d becomes a constant value of 2 to satisfy Eq. (6). This demonstrates that amplitude panning underpins the localization theory in common Ambisonics playback. The following sections use this common core of amplitude panning to show that S^{3}AC can be used efficiently to compress Ambisonics signals while retaining stereo/mono backward compatibility in a downmixed signal.

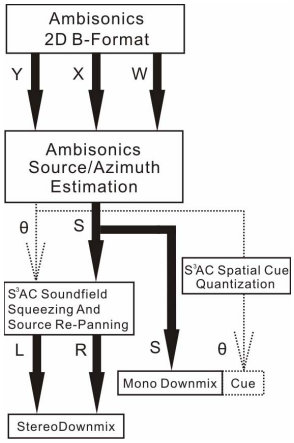

Fig. 2. S^{3}AC Compression of Ambisonics Signals

3.2. UHJ Compression of Ambisonics Signals

Conventional Ambisonics applications use UHJ [7] as a two-channel downmix method to attain backward compatibility with classical stereo systems. Considering a 2D Ambisonics signal, in the frequency domain, the B-Format given in Eq. (2) has only WXY information with α=0, resulting in:

$$W_k = S_k/\sqrt{2},\ \ X_k = \cos(\theta_k)\cdot S_k,\ \ Y_k = \sin(\theta_k)\cdot S_k \qquad (8)$$

where subscript k is the bin frequency index. UHJ encoding on the 2D B-Format signal is then performed according to:

$$L_K = (0.4699 - 0.171j)W_k + (0.0928 + 0.255j)X_k + 0.3277Y_k$$
$$R_k = (0.4699 + 0.171j)W_k + (0.0928 - 0.255j)X_k - 0.3277Y_k \qquad (9)$$

These relationships do not give lossless transmission of B-format and this is confirmed in the tests reported in Section 4. While S^{3}AC using a stereo downmix is also lossy, Section 4 demonstrates that it significantly outperforms the legacy UHJ approach.

3.3. S^{3}AC Compression of Ambisonics Signals

As illustrated in Fig.2, the S^{3}AC approach to compressing 2D Ambisonics signals starts from frequency domain source and azimuth estimation, where source S_k can be derived from W_k in Eq. (8) and its azimuth θ_k in a 360° sound field can be obtained from the following trigonometric relationships:

$$\cos(\theta_k) = X_k \Big/ \sqrt{2}\cdot W_k,\ \ \sin(\theta_k) = Y_k \Big/ \sqrt{2}\cdot W_k \qquad (10)$$

This process can be performed for either every frequency or in (perceptual) frequency bands and various time-frequency transforms can be used. Here, a STFT was utilized. The estimated sources and azimuths are fed to standard S^{3}AC azimuth squeezing process, as illustrated in Fig.1. A new azimuth μ_k in the squeezed soundfield is calculated according to its original azimuth θ_k in the 360° sound field and assigned to the source. Consequently, the source is re-panned into a pair of stereo channels using this azimuth in the squeezed field:

$$L_k = S_k\cdot\left[\tan(\eta) + \tan(\mu_k)\right],\ \ R_k = S_k\cdot\left[\tan(\eta) - \tan(\mu_k)\right] \qquad (11)$$

where η is the azimuth separation between the two stereo speakers, typically 30°. This in turn results in a 60° stereo soundfield

containing the information of a 360° soundfield from the B-Format. As a consequence, by analyzing the stereo soundfield at the decoder, the source spectral information S'_k can be re-estimated and the azimuth μ'_k in the squeezed field can be recovered to 360° domain to form θ'_k. Hence, with the source and its directional information, the B-Format signal can be recovered by applying Eq. (8). This process provides a fully stereo compatible conversion and compression of Ambisonics B-Format which is undistorted and listenable compared with a UHJ downmix. We note that, as with DirAC [12], S^{3}AC localization is directly derivable from B-format.

3.4. Creating a Mono Downmix Using Side Information

Adding side information provides further flexibility and extensibility to S^{3}AC compression of B-Format Ambisonic signals. For a 2D B-Format signal, similarly to Section 3.3, the sound source S_k can be estimated in the frequency domain from the W-channel in Eq. (8) and then transformed to the time domain to produce a mono downmix. The azimuth θ_k can then be estimated from Eq. (10) for each frequency domain source. This azimuth information forms the S^{3}AC side information for the mono downmix. The decoder can then recover the B-Format based on the source S_k and related localization information θ_k using Eq. (8).

The quantization of S^{3}AC side information can be efficiently achieved by exploiting both conventional monophonic perceptual psychoacoustics and spatial localization psychoacoustic. The latter is further investigated in Section 4.1.

4. EXPERIMENTS AND EVALUATIONS

4.1. Localization Dependent Side Information Quantization

Existing spatial audio coders [1, 2] exploit human auditory frequency sensitivity for spectral quantization; but human auditory localization is not directly exploited. S^{3}AC utilizes localization blur to effect compression of the space through squeezing, but this approach can be improved by recognizing that localization blur is, in itself, location dependent. Psychoacoustic research has shown a localization precision of approximately 0.5°~1° in front of a listener, reducing to more than 10° on the sides and to the rear of the listener [8]. This leads to approximately 7 bits and 3 bits effective azimuth precision for the 60° front region and 140° rear region respectively, in the ITU 5.1 channel setup.

To further investigate these theories and exploit them in coding applications, listening tests based on MUSHRA [10] were performed, where listeners were asked to compare the localization accuracy between a reference and coded source. Four types of moving sound sources were used, including a 500Hz tone, 1kHz tone, band-pass noise with two critical-band pass-band with central frequency at approximately 2kHz and a car siren source. For an ITU 5.1 channel setup, each object is panned into four horizontal areas: 30° to -30° in the front, ±30° to ±110° in the left and right, ±110° to ±180° in the rear respectively. The original signals were compared with sources panned to discrete azimuths ranging from 64 linearly discriminated azimuths (6 bits) to 4 azimuths (2 bits). A non-moving anchor source signal was used and six listeners participated in the tests.

The results including mean and 95% confidence intervals are shown in Fig.3. It is shown that, for the front and side sources, the perceived distortion increases with the decreasing azimuth precision while there is strong ambiguity for rear sources. According to the results, by using 5 or 4 bits (32 and 16 discrete azimuths respectively) for the front and side azimuth quantization,

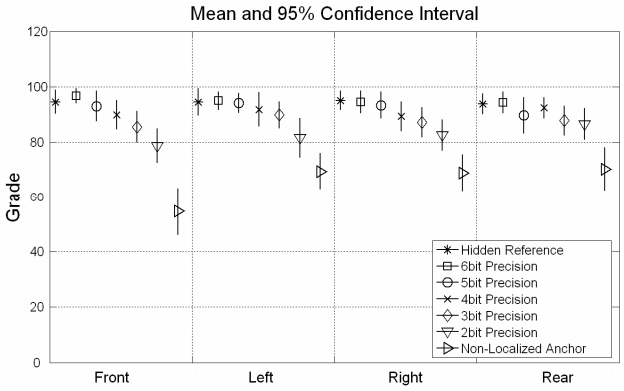

Fig. 3. Listening Tests Results of Localization Dependency of Spatial Cue Quantization

the accuracy of the coded material is within 90% comparing to the original. However, in the ambiguous rear plane, similar, but unreliable, accuracy results precision ranging from 4 to 32 discrete azimuths. These results suggest that, while previous psychoacoustical research indicates higher precision for perceptually undistorted quantization, reduced precision is adequate in coding applications.

Based on these results, 5, 4 and 2 bits of precision were used in the front, side and rear planes respectively, for quantization of the S^{3}AC cues in this work; an extra 2 bits are then needed to indicate the source region. This results in a variable bit rate scheme, with direct quantization resulting in approximately 260kbps for an average of 4 bits per spatial cue and all coefficients. However, quantization of one cue for each of 20 Bark spectral frequency bands (as used in existing spatial audio coders [2]) reduces this bit rate to approximately 10kbps. Further compression utilizing entropy coding could reduce this bit rate further, however this is beyond the scope of this paper.

It is also possible to generate a fixed bit rate azimuth quantization scheme by allocating a non-uniformly spaced codebook to the 360° of azimuth values. Based on the precision requirements for the frontal, side and rear regions, a codebook of 64 values and hence 6 bits per spatial cue was found to be suitable. This results in a fixed rate of 10kbps for 20 Bark spectral bands.

4.2. Objective Evaluation

S^{3}AC compression of Ambisonics signals was evaluated objectively against the UHJ method. Three modes of S^{3}AC were evaluated: a stereo downmix without side information, a mono downmix with un-quantized side information and a mono downmix with quantized side information (abbreviated as S^{3}AC SD, S^{3}AC MD-UQ and S^{3}AC MD-Q respectively). The perceptual results and the scalar bit allocations from Section 4.1 were utilized during the quantization step. The Kullback-Leibler Spectral Distance measurement [11] was used for objective evaluation of the encodings and was calculated between the original signal and all four coding conditions. For each component signal, the average Kullback-Leibler Spectral Distance for each coefficient in each channel of 2D B-Format was calculated according to:

$$D_i = \frac{1}{N \cdot K} \sum_N \sum_K \left(P_i(N,K) - Q_i(N,K)\right) \log \frac{P_i(N,K)}{Q_i(N,K)} \quad (12)$$

where N, K are frame and frequency index respectively and i=*W, X or Y* for the three channels of B-Format. Eight 2D Ambisonic recordings, including immersive soundfield, live concert

Table.1. W, X, Y Channel and Average Kullback-Leibler Spectral Distance of Eight 2D Ambisonics Recordings

	W ($\times 10^{-2}$)	X ($\times 10^{-1}$)	Y ($\times 10^{-1}$)	WXY Average ($\times 10^{-1}$)
UHJ	7.59	1.99	1.68	1.47
S^{3}AC SD	3.06	1.35	1.38	1.01
S^{3}AC MD-UQ	0.00	1.37	1.24	0.87
S^{3}AC MD-Q	0.00	1.77	1.53	1.10

recordings and surround rendered music, were used as test signals and the average results are given in Table 1. While distortion in W channel relates to perceptual quality degrading, distortion in X and Y channel will result in error in localization. Since only sound pressure information is stored in the W channel, the X and Y channels contain the azimuth-localization information. In all three modes, S^{3}AC gives more precise recovery of the spectrum than UHJ for all of the B-format channels. This indicates that, in comparison with UHJ, S^{3}AC more accurate represents both source and its localization. The W channel for S^{3}AC MD-UQ and MD-Q is undistorted, as the mono downmix in these two modes is a perfectly scaled version of the original W channel, as described in Section 3.4. While the quantization of azimuth side information adds distortion, we now show that there is no perceptual impact.

4.3. Subjective Evaluation

Listening tests were performed using the same test materials and coding conditions detailed in Section 4.2. The original and coded B-Format files were converted into 5-channel format according to the ITU 5.1 channel setup and the MUSHRA [10] methodology was employed. An un-localized 3.5kHz low-pass filtered version was used as an anchor signal and six listeners participated in the tests; the results including mean and 95% confidence intervals are shown in Fig.4. Compared with UHJ, all three S^{3}AC approaches show significantly higher scores, with an average 25% improvement in the MUSHRA score. In addition, it should be noted that, while quantization of the S^{3}AC side information objectively increases the spectral distortion, no perceptual distortion is detected subjectively in the listening tests. This further indicates that location dependent spatial cue quantization, as described in Section 4.1, can be efficiently used to further reduce the bit-rates of S^{3}AC side information without introducing perceptual localization distortion. This approach is applicable to any spatial audio coding technique which transmits source location related side information.

5. CONCLUSIONS AND FURTHER WORK

This paper has demonstrated that S^{3}AC is a versatile and efficient representation of multi-channel spatial audio signals. Supplementing S^{3}AC with azimuth based side information provides advantages when compressing complex sound environments as well as introducing further flexibility to S^{3}AC. It has been shown that S^{3}AC, shares with the usual 4 speaker ambisonics setup a common basis of amplitude panning. In addition, S^{3}AC shows significant advantages in producing stereo/mono compatible compression of 2D Ambisonics signals when compared with the conventional UHJ approach. The paper showed that within such a scheme, the localization dependency of spatial cue quantization could be exploited to advantage when creating S^{3}AC side information. Psychoacoustic experiments were performed to test the requirements of perceptual quantization of localization information and the results indicate that in compression, previous psychoacoustic research is unnecessarily pessimistic in quantization requirements. This was verified in subjective tests comparing the original soundfield and a range of encodings of the Ambisonics B-format representation. S^{3}AC also offers the advantage over UHJ that the W (omnidirectional) component of the ambisonics signal is not distorted by the spatial encoding process. Further work will investigate and evaluate the compression of 3D Ambisonics using S^{3}AC, as well as a more comprehensive spatial quantization and masking theory.

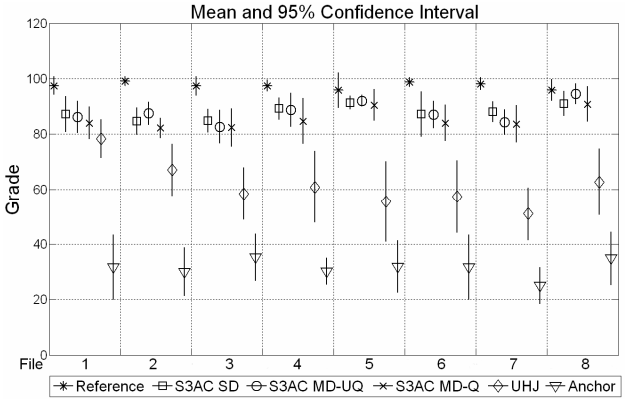

Fig. 4. Listening Tests Results Comparing S^{3}AC and UHJ Compression of 2D Ambisonics

6. REFERENCES

[1] C. Faller, F, Baumgarte, "Binaural Cue Coding – Part II: Schemes and Applications", *IEEE Trans. on Speech and Audio Proc.*, vol.11, No.6, Nov., 2003.

[2] J. Breebaart, et al., "MPEG Spatial Audio Coding/MPEG Surround: Overview and Current Status", in *Proc. 119th AES Convention*, New York, USA, Oct., 2005.

[3] B. Cheng, C. Ritz, I. Burnett, "Encoding Independent Sources in Spatially Squeezed Surround Audio Coding", in *Proc. PCM2007*, HongKong, China, Dec., 2007.

[4] B. Cheng, C. Ritz, I. Burnett, "Principles and Analysis of the Squeezing Approach to Low Bit Rate Spatial Audio Coding", in *Proc. IEEE ICASSP 2007*, Honolulu, USA, Apr., 2007.

[5] ITU-R BS.775-1, "Multichannel Stereophonic Sound System with and without Accompanying Picture", 1994.

[6] M. A. Gerzon, "Ambisonics, Part Two: Studio Techniques," *Studio Sound*, vol. 17, pp. 24-30, Aug., 1975.

[7] M. A, Gerzon, "Ambisonics in Multichannel Broadcasting and Video", *J. Audio Eng. Soc.,* vol.33, No.11, Nov., 1985.

[8] J. Blauert, "Spatial Hearing: the Psychophysics of Human Sound Localization", *MIT Press*, Cambridge, MA, USA, 1996.

[9] A. Farina, et al., "Ambiophonic Principles for the Recording and Reproduction of Surround Sound for Music", in *Proc. 19th AES Inter. Conf. of Surround Sound*, p26-46, Germany, 2001.

[10] ITU-R BS. 1534, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems (MUSHRA)", 2001.

[11] R. Veldhuis, E. Klabbers, "On the Computation of the Kullback-Leibler Measure for Spectral Distances", *IEEE Trans. on Speech and Audio Processing*, vol. 11, No. 1, Jan. 2003.

[12] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding", *J. Audio Eng. Soc.*, vol. 55, No. 6, June 2007.