PRINCIPLES AND ANALYSIS OF THE SQUEEZING APPROACH TO LOW BIT RATE SPATIAL AUDIO CODING

Bin Cheng, Christian Ritz and Ian Burnett

Whisper Laboratories, University of Wollongong, Wollongong, NSW, Australia
bc362@uow.edu.au, chritz@elec.uow.edu.au, i.burnett@elec.uow.edu.au

ABSTRACT

This paper presents a novel solution to multichannel spatial audio coding: Spatial Squeezing Surround Audio Coding (S^3AC). The S^3AC scheme analyses a multichannel audio signal and downmixes it into a stereo signal pair containing both the monophonic properties of audio sources and their localization information; this avoids the need for side information. The approach uses time-frequency analysis of a spatial audio scene and exploits virtual sources and amplitude panning techniques to 'squeeze' 360º of a horizontal soundfield to a 60º stereo signal pair. In comparison with other spatial audio coding techniques, S^3AC significantly advances in-band encoding of the localization information in the original sound scene and achieves accurate recoverability of dynamic localized sources.

Key Words: Audio Coding, Audio Systems.

1. INTRODUCTION

Efficient representation of multichannel spatial audio signals has been an area of great interest in recent years. Conventional perceptual codecs such as MP3 and AAC [1] are inefficient for multichannel audio since the bit-rate linearly increases with the number of channels. Thus, coding techniques based on exploiting the inter-channel mathematical relationships such as Binaural Cue Coding (BCC) [2], Parametric Stereo [3] and the recently released MPEG Surround [4] have been proposed. Based on a "downmix + cues" framework, these schemes aim at extracting cross-channel level differences, phase differences and correlation to represent localization information. This data can then be transmitted such that the multichannel signal can be recovered from a stereo or mono downmix. While some recent research investigates transmitting source direction information rather than the cross-channel cues [5], the successful recovery of sound localization is still critically dependent on the extra side information.

The S^3AC solution presented in this paper addresses spatial audio coding by exploiting the localization redundancy of the surround sound and avoids sending side information. Rather, source location information is transmitted within the encoded downmix signal. In comparison with other multichannel audio coding approaches (exploiting cross-channel relevancy and difference), S^3AC has been shown to require much lower computational cost whilst comparing favourably with MPEG Surround in terms of perceptual quality [6]. The precise recoverability of localized sound sources offered by S^3AC makes its usage attractive for precisely mixed audio or conferencing applications. This paper will present new results comparing the localization performance of S^3AC with current spatial audio coders.

Section 2 of this paper describes the principles behind S^3AC while Section 3 analyses the time-frequency characteristics of S^3AC. Section 4 compares the localization performance of S^3AC with an existing spatial audio coder and Section 5 presents conclusions.

2. SQUEEZING THE AUDITORY SPACE

The fundamental concept of S^3AC is that a full surrounding auditory panorama for perceptual listening purposes can be represented by a smaller (squeezed) sound field for the purposes of transmission. Thus, the data rate of the compressed signal is independent of the number of original channels but is instead dependent on the size of the squeezed sound field. Psychoacoustic research has shown that the human ear has limited precision in localizing static sound sources (known as localization blur [7]). On the horizontal plane, the localization blur is approximately 0.5º~1º, i.e. displacement of an auditory object by around 1º is perceptually unnoticeable [7]. Hence the aim of 'squeezing' the space is to ensure that the sound field can be squeezed and expanded in a 'perceptual localization lossless manner' by minimizing the displacement of the rendered sound sources. Typically, the requirement is to squeeze a 360º horizontal soundfield (rendered by e.g. the standard ITU 5-channel speaker setup (see Fig. 1) [8]) to a stereo signal pair. Further compression of the stereo signal by a perceptual audio coder, (e.g. MP3 or AAC [1]), allows encoding at a bit-rate comparable with stereo audio coders but facilitates recovery a full surround auditory scene. The system diagram for S^3AC is illustrated in Fig. 2 and is described below.

2.1. Time-Frequency Transform

If the soundfield is artificial or unrendered, objects may already be 'positioned' and sound estimation is not required. However, where the original soundfield is generated by e.g. a 5-channel setup, each channel of the original signal is transformed into the frequency domain (e.g. using a Short-Time Fourier Transform (STFT) or a Pseudo Quadrature Mirror Filterbank (PQMF)) prior to estimation of virtual sources and generation of a stereo downmix. In the decoder, upmixing of the received stereo signal (by reversing the process described in Section 2.2) and inverse transformations to multichannel output signals are performed. Further details of the time-frequency transform stages are provided in Section 3.

2.2. Virtual Source Estimation and Azimuth Analysis

The second stage of S^3AC evaluates the 'rendered' surround sound scene and virtual sources. The 360º sound field can be divided into

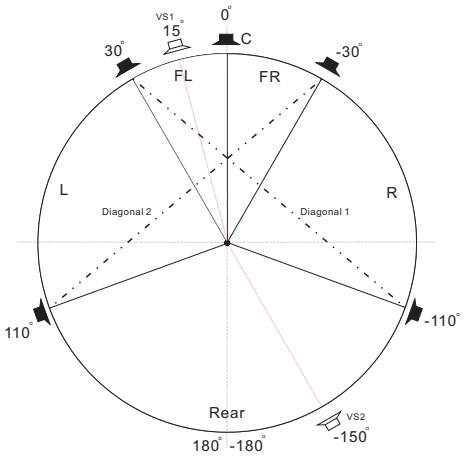

Fig. 1. 5-Channel Sound System and Auditory Sub-Spaces

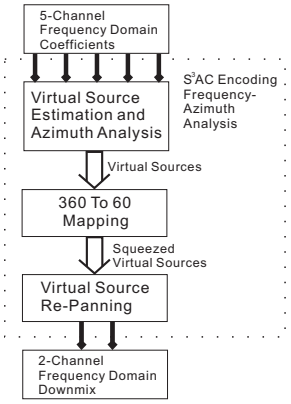

Fig. 2. S^{3}AC Encoder Diagram

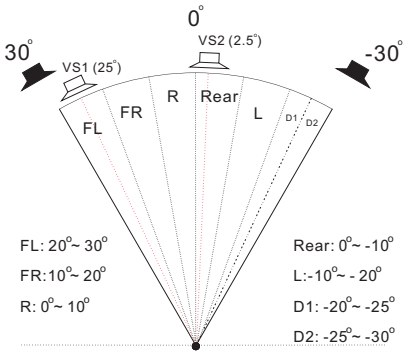

Fig.3. A 'Squeezed' Space Represented by Two Channels

a number of sub-regions based on the speaker setup (see Fig. 1.) such that each sub-region is represented by amplitude panning between a pair of channels. Adjacent channels are used to render virtual sources located at the perimeter while diagonal channel pairs are used to render virtual sources within the central soundfield area. When S^{3}AC is applied to existing recordings, pairs of frequency coefficients (or small sets of coefficients, similar to [4]) from adjacent or diagonal channel pairs are assumed to represent a virtual source.

The amplitudes $A_a(k)$ and $A_b(k)$ of the virtual source of frequency k rendered by two channels *a* and *b* are computed as:

$$\{A_a(k), A_b(k)\} = \max_{ij}\{[A_i(k), A_j(k)]\} \quad (1)$$

where $A_i(k)$ and $A_j(k)$ are the magnitudes of channel pair {i, j} as a function of frequency, k. This method identifies a virtual source as that rendered by the channel pair with the dominant energy for a given frequency component. The corresponding azimuth of the virtual source θ_{ab} is computed as:

$$\theta_{ab} = \arctan\left[\frac{A_a(k) - A_b(k)}{A_a(k) + A_b(k)} \cdot \tan(\varphi_{ab})\right] \quad (2)$$

where φ_{ab} is the angular separation of the chosen channel pair a and b of Fig. 1. A virtual source is thus represented by the frequency coefficient amplitudes of the chosen channel pair and its azimuth: $\{A_a(k), A_b(k), \theta_{ab}\}$. The virtual source can then be represented as a single mono signal as:

$$S(k) = \sqrt{A_a^2(k) + A_b^2(k)}.e^{\phi_{ab}} \quad (3)$$

where the phase ϕ_{ab} is arbitrarily chosen to be the phase of either channel *a* or *b*. This process assumes sources are rendered in a similar way to the amplitude panning approach of [9, 10], however here panning is applied to individual (or groupings) of frequency coefficients. The identified virtual source is then squeezed into the new auditory space as described in Section 2.3.

2.3. 360-To-60 Mapping and Virtual Source Re-panning

Once frequency-domain virtual sources are derived for each channel pair, a new azimuth in the stereo sound field is assigned to this source according to a 360º to 60º linear mapping criteria such that $\theta_{DM} = f(\theta_{ab})$. This source is then re-panned to this new direction in the downmix field using a similar process to [6] but based on the Tangent amplitude panning law [9, 10] as given in Eq. (4) where *L* is a factor to maintain the overall energy. This process results in the 'squeezed' version of the original surround sound field as shown in Fig. 3.

$$\mathbf{Y}_{DM}(k) = \begin{bmatrix} Y_L(k) \\ Y_R(k) \end{bmatrix} = \frac{S(k)}{L}\begin{bmatrix} \tan(\varphi_{DM}) + \tan(\theta_{DM}) \\ \tan(\varphi_{DM}) - \tan(\theta_{DM}) \end{bmatrix} = S(k)\cdot\mathbf{M}_{DM}(\theta_{DM}) \quad (4)$$

The stereo signal pair rendering this squeezed space contains all the auditory and localization information of the original signal; this contrasts with [2-4] where additional spatial cues are required to describe the inter-channel relationships. Additionally, while the azimuth mapping criteria used here maximizes the discrimination of the source localization, other mapping criteria can also be chosen to meet specific requirements, e.g. to centralizes frontal audio in the downmix for stereo compatibility. For decoding, the S^{3}AC decoder applies 'inverse' amplitude panning on the downmixed stereo signal pair, allowing each frequency virtual source to be estimated and re-panned to the full 360º soundfield. A more detailed description of S^{3}AC is given in [6].

3. TIME-FREQUENCY ANALYSIS

Fig. 4 illustrates the filterbank system of S^{3}AC. In the encoding stage of Fig. 4, the five time domain channel signals x_1 to x_5 are decomposed into the frequency domain to produce signals Y_1 to Y_5. Following S^{3}AC encoder frequency azimuth analysis, downmixed frequency domain signals Y_L and Y_R are converted back to the time domain resulting in downmixed signals x_L and x_R, prior to possible coding and transmission. In the decoder, the received stereo signals are transformed back to the frequency domain resulting in signals $\hat{Y}_L$ and $\hat{Y}_R$. Following virtual source estimation and re-synthesis, the resulting five frequency domain output signals are converted back to the time domain to produce signals $\hat{x}_1$ to $\hat{x}_5$. As this process relies on both accurate frequency and time domain reconstruction, the S^{3}AC filterbank should achieve both time-domain and frequency-domain aliasing cancellation.

3.1. Analysis of Aliasing in S^{3}AC

To avoid blocking effects, the initial time-frequency transform

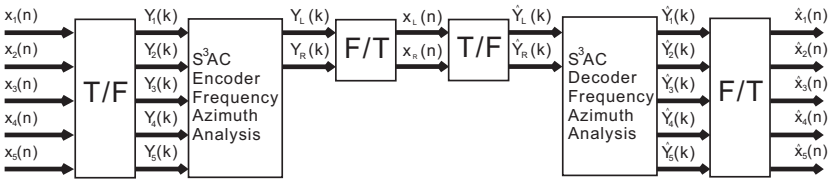

Fig. 4. S^{3}AC and the Filterbank System

Table. 1. Average Localization Error in Squeezed Domain in Degrees

	PQMF	STFT	MDCT
Airplane	0.482	0.354	0.423
Car	0.255	0.446	0.398
Female Speech	0.164	0.184	0.152
Male Speech	0.207	0.179	0.167
Mosquito	0.395	0.312	0.362

applies an analysis window when calculating the frequency domain multichannel signals $Y_i(k)$ of Fig. 4., described as:

$$A_i(k) = \left|Y_i(k)\right| = \left|X_i(k)W_i(k)\right| \quad i = 1 \text{ to } 5 \qquad (5)$$

where $X_i(k)$ and $W_i(k)$ are the frequency (or sub-band) domain coefficients representing the input signal and window functions, respectively. Traditional audio coders apply a synthesis window prior to overlap add reconstruction to remove the effects of the analysis window. However, the spatial squeezing approach of S^{3}AC relies on multiplication of the amplitudes of frequency domain virtual source components during re-panning (see Eq.s (1) to (4)) to create a stereo downmix. This in turn leads to multiplicative modification of the frequency domain window coefficients $W_i(k)$ of Eq. (5). This modification of the window function is frequency and amplitude dependent and is unpredictable. Hence, applying a traditional matched analysis/synthesis window approach will not result in complete aliasing cancellation in S^{3}AC. This leads to errors in both reconstruction of the frequency domain stereo downmix and the reconstruction of the time domain multichannel output signals. The frequency domain errors will be examined in the next sub-section while time domain errors will be examined in Section 4.

3.2 Analysis of the Frequency Domain Errors

The frequency domain errors can be minimized if coding or transmission of the downmix signal is performed in the frequency domain (e.g. by incorporating into an existing audio coder). However, if a time domain stereo downmix is required, there will be some distortion introduced. While the filterbank induced error is important, it is more critical to consider the change in location of the virtual sources due to the changes in the frequency coefficients.

To investigate these errors, a number of typical multichannel audio signals (as used in Section 4) were tested using three time-frequency transform methods (see Table 1). Since the average displacements are within 0.5º in the squeezed space at worst, a maximum of 3.5º error will be experienced in the recovered 360º sound field. Furthermore, S^{3}AC squeezes the space on the basis of perceptual localization ability and hence for recovered frontal sources the maximum error is 1.5º.

4. LOCALIZATION EVALUATION

In this section, the localization accuracy of the decoded multichannel audio of S^{3}AC and MPEG Surround is evaluated both objectively and subjectively.

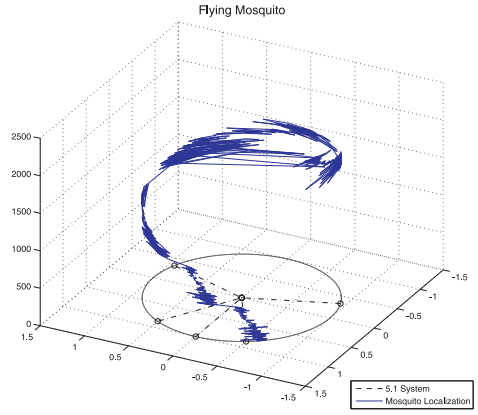

Fig. 5 Flying Mosquito

4.1. Objective Analysis

A 5-channel example signal rendering a mosquito flying slowly around the horizontal plane was generated for objective evaluation purposes. Fig. 5 illustrates this example with the horizontal plane representing the 360º sound scene surface and the z-axis indicates the time frame.

The azimuth of the major source in the soundfield was objectively identified by applying the inverse panning law of Eq. (2) to the two strongest channels as determined by Eq. (1). S^{3}AC and MPEG Surround 525 and MPEG Surround Non-Guided were then compared against the original soundfield using this measure. During the evaluation, the virtual object and the respective azimuth were generated on a per-frequency-coefficient basis, using a 50% overlapped 1024-point STFT. The azimuth difference of corresponding frequency coefficients between the original signal and the three coding methods were calculated as the decoding azimuth error.

Table 2 shows the average results of this evaluation for the mosquito signal. It can be seen that S^{3}AC, even though it requires no side information, offers a 7 fold improvement over MPEG Surround 525 and provides localization recoverability only marginally beyond the 1º requirement for perceptually indistinguishable localization. In addition, in comparison with MPEG Surround NonGuided mode, S^{3}AC offers a 30 fold improvement in average localization accuracy.

Fig. 6 compares S^{3}AC with MPEG Surround NonGuided in a more intuitive way where the x and y axis represent the frame number and FFT coefficients respectively, while the z axis indicates azimuth on the 360º horizontal plane. The performance improvements offered at no bit rate increase are again apparent.

4.2. Subjective Tests

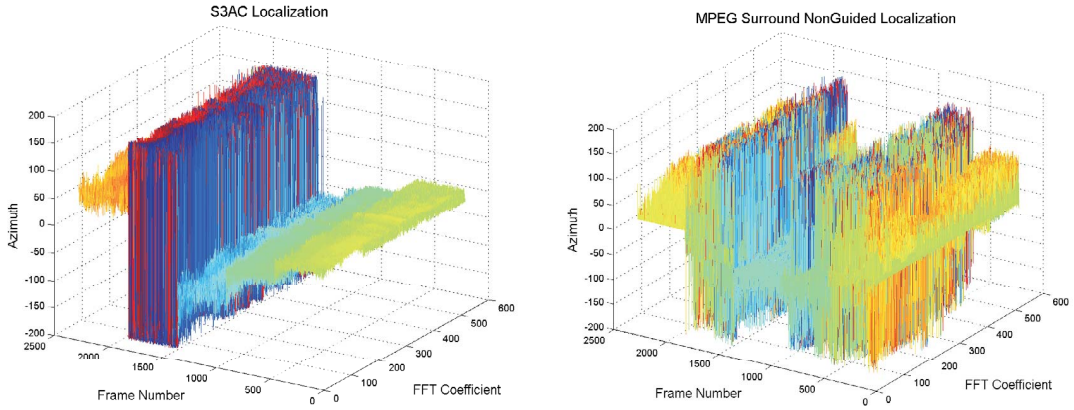

Fig. 6. Localization Recovery of S^{3}AC and MPEG Surround NonGuided

Table. 2. Average Azimuth Error of Three Coders

S^{3}AC	MPEG Surround 525	MPEG Surround NG
1.39º	10.28º	42.75º

Subjective listening tests were also performed on S^{3}AC and MPEG Surround. Besides the mosquito example used in the previous section, several other localized immersive soundfield examples were employed; these included moving speech, aircraft and car sirens. The test methodology employed was based on the MUSHRA approach [11] but the listeners were asked to focus on the localization accuracy of the coded material. A non-localized anchor signal equally mixing a mono signal to each channel was provided during the test. Nine listeners participated in the tests and the results including mean and 95% confidence intervals are shown in Fig. 7. The results indicate that, without any side information,the localization error of S^{3}AC cannot be perceptually detected. Further the results show that listeners have difficulty in distinguishing the hidden reference, S^{3}AC and MPEG Surround 525. However, the MPEG Surround NonGuided coder is easily distinguished as it results in significant distortion of source position.

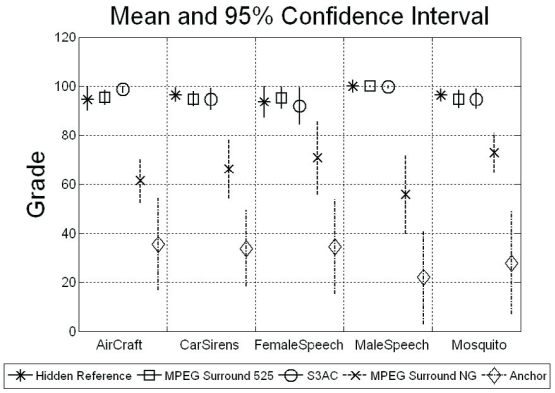

Fig. 7. Listening Test Results

5. CONCLUSIONS AND FURTHER WORK

An efficient approach to spatial audio coding, called S^{3}AC, has been presented. Its fundamental theory is based on squeezing the auditory space by identifying frequency domain virtual sources and exploiting perceptual localization redundancy. An investigation into the impact of the time-frequency transform has been examined and shown to produce minimal errors in the location of squeezed virtual sources. When compared with a state-of-the-art spatial audio coder, S^{3}AC shows a significant advantage in maintaining the localization of audio sources within the soundfield. Further work will investigate the use of alternative perceptually based algorithms for identifying and squeezing virtual sources as well as new algorithms for minimizing the aliasing effects resulting from time-frequency analysis.

REFERENCES

[1] M. Bosi, R.E. Goldberg, "Introduction to Digital Audio Coding and Standards", *Springer Science+Business Media*, New York, USA, 2002.

[2] 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.

[3] J. Breebaart, et al., "Parametric Coding of Stereo Audio", *EURASIP Jour. Applied Signal Proc.*, 1305-1322, Sep. 2005.

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

[5] H. Moon, et al., "A Multi-channel Audio Compression Method with Virtual Source Localization Information", in *Advances in MultiMedia Info. Proc. PCM 2005*, Springer-Verlag, 2005.

[6] B. Cheng, C. Ritz, I. Burnett, "Squeezing the Auditory Space: A New Approach to Multi-Channel Audio Coding", in *Proc. PCM2006*, Hangzhou, China, Nov., 2006.

[7] J. Blauert, "Spatial Hearing: the Psychophysics of Human Sound Localization", *MIT Press,* 1997.

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

[9] I.M. Neoran, "Surround Sound Mixing Using Rotation, Stereo Width, and Distance Pan-Pots", in *Proc. 109th AES Convention*, Los Angeles, USA, Sep., 2000.

[10] V. Pulkki, "Localization of Amplitude-Panned Virtual Sources II: Two- and Three-Dimensional Panning", J. Audio Eng. Soc., Vol. 49, No. 9, Sep., 2001.

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