2007

Principles and analysis of the squeezing approach to low bit rate spatial audio coding

B. Cheng
University of Wollongong

C. H. Ritz
University of Wollongong, critz@uow.edu.au

I. Burnett
University of Wollongong, ianb@uow.edu.au

Publication Details
This paper was originally published as: Cheng, B., Ritz, C. & Burnett, I., Principles and analysis of the squeezing approach to low bit rate spatial audio coding, IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007), Honolulu, Hawaii, USA, 15-20 April 2007, 1, I-13-I-16. Copyright 2007 IEEE.
Principles and analysis of the squeezing approach to low bit rate spatial audio coding

Abstract
This paper presents a novel solution to multichannel spatial audio coding: spatial squeezing surround audio coding (S\(^3\)AC). The S\(^3\)AC 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' 360\(^\circ\) of a horizontal soundfield to a 60\(^\circ\) stereo signal pair. In comparison with other spatial audio coding techniques, S\(^3\)AC significantly advances in-band encoding of the localization information in the original sound scene and achieves accurate recoverability of dynamic localized sources.

Disciplines
Physical Sciences and Mathematics

Publication Details
This paper was originally published as: Cheng, B., Ritz, C. & Burnett, I., Principles and analysis of the squeezing approach to low bit rate spatial audio coding, IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007), Honolulu, Hawaii, USA, 15-20 April 2007, 1, I-13-I-16. Copyright 2007 IEEE.

This conference paper is available at Research Online: http://ro.uow.edu.au/infopapers/518
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 (SSAC). The SSAC 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’ 360° of a horizontal soundfield to a 60° stereo signal pair. In comparison with other spatial audio coding techniques, SSAC 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 SSAC 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 relevance and difference), SSAC 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 SSAC makes its usage attractive for precisely mixed audio or conferencing applications. This paper will present new results comparing the localization performance of SSAC with current spatial audio coders.

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

2. SQUEEZING THE AUDITORY SPACE

The fundamental concept of SSAC 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 0.5°-1°, i.e. displacement of an auditory object by around 1° 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 360° horizontal sound field (rendered by e.g. the standard ITU 5-channel speaker setup (see Fig. 1) [8]) to a 60° 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 SSAC 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 SSAC evaluates the ‘rendered’ surround sound scene and virtual sources. The 360° sound field can be divided into
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'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 $A_{i}(k)$ and $A_{j}(k)$ of the virtual source of frequency $k$ rendered by two channels $i$ and $j$ are computed as:

$$\{A_{i}(k), A_{j}(k)\} = \max_{p} \{[A_{p}(k), A_{p}(k)]\}$$

(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 $\theta_{ab}$ is computed as:

$$\theta_{ab} = \arctan\left[\frac{A_{j}(k)-A_{i}(k)}{A_{i}(k)+A_{j}(k)}\right] \tan(\phi_{ab})$$

(2)

where $\phi_{ab}$ 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: $\{A_{i}(k), A_{j}(k), \theta_{ab}\}$. The virtual source can then be represented as a single mono signal as:

$$S(k) = \sqrt{A_{i}^{2}(k) + A_{j}^{2}(k)} e^{j \phi_{ab}}$$

(3)

where the phase $\phi_{ab}$ 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 360° to 60° linear mapping criteria such that $\theta_{60} = f(\theta_{360})$. 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 sound field as shown in Fig. 3.

$$Y_{60}(k) = \frac{Y_{360}(k)}{L} \frac{\tan(\theta_{360}) + \tan(\theta_{60})}{\tan(\theta_{60}) - \tan(\theta_{360})}$$

(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 centralize frontal audio in the downmix for stereo compatibility. For decoding, the S'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 360° soundfield. A more detailed description of S'AC is given in [6].

3. TIME-FREQUENCY ANALYSIS

Fig. 4 illustrates the filterbank system of S'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 $\tilde{Y}_{1}$ to $\tilde{Y}_{5}$. Following S'AC encoder frequency azimuth analysis, downmixed frequency domain signals $\tilde{Y}_{1}$ and $\tilde{Y}_{2}$ are converted back to the time domain resulting in downmixed signals $\tilde{x}_{1}$ and $\tilde{x}_{2}$, prior to possible coding and transmission. In the decoder, the received stereo signals are transformed back to the frequency domain resulting in signals $\tilde{Y}_{1}$ and $\tilde{Y}_{2}$. Following virtual source estimation and reconstruction, the resulting five frequency domain output signals are converted back to the time domain to produce signals $\tilde{x}_{1}$ to $\tilde{x}_{5}$. As this process relies on both accurate frequency and time domain reconstruction, the S'AC filterbank should achieve both time-domain and frequency-domain aliasing cancellation.

3.1. Analysis of Aliasing in S'AC

To avoid blocking effects, the initial time-frequency transform
Table 1. Average Localization Error in Squeezed Domain in Degrees

<table>
<thead>
<tr>
<th></th>
<th>PQMF</th>
<th>SEFT</th>
<th>MDCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airplane</td>
<td>0.485</td>
<td>0.354</td>
<td>0.423</td>
</tr>
<tr>
<td>Car</td>
<td>0.255</td>
<td>0.446</td>
<td>0.398</td>
</tr>
<tr>
<td>Female Speech</td>
<td>0.164</td>
<td>0.184</td>
<td>0.152</td>
</tr>
<tr>
<td>Male Speech</td>
<td>0.207</td>
<td>0.179</td>
<td>0.167</td>
</tr>
<tr>
<td>Mosquito</td>
<td>0.395</td>
<td>0.312</td>
<td>0.362</td>
</tr>
</tbody>
</table>

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

$$A(k) = |X_i(k)| |\mathcal{W}_i(k)| \quad i = 1 \text{ to } 5$$

where $X(k)$ and $W(k)$ 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\(^2\)AC relies on multiplication of the amplitudes of frequency domain virtual source components during re-panning (see Eqs. (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\(^2\)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 0.5\(^\circ\) in the squeezed space at worst, a maximum of 3.5\(^\circ\) error will be experienced in the recovered 360\(^\circ\) sound field. Furthermore, S\(^2\)AC squeezes the space on the basis of perceptual localization ability and hence for recovered frontal sources the maximum error is 1.5\(^\circ\).

4. LOCALIZATION EVALUATION

In this section, the localization accuracy of the decoded multichannel audio of S\(^2\)AC and MPEG Surround is evaluated both objectively and subjectively.

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 360\(^\circ\) 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\(^2\)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\(^2\)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 1\(^\circ\) requirement for perceptually indistinguishable localization. In addition, in comparison with MPEG Surround NonGuided mode, S\(^2\)AC offers a 30 fold improvement in average localization accuracy.

Fig. 6 compares S\(^2\)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 360\(^\circ\) horizontal plane. The performance improvements offered at no bit rate increase are again apparent.

4.2. Subjective Tests
Subjective listening tests were also performed on S\textsuperscript{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\textsuperscript{3}AC cannot be perceptually detected. Further the results show that listeners have difficulty in distinguishing the hidden reference, S\textsuperscript{3}AC and MPEG Surround 525. However, the MPEG Surround NonGuided coder is easily distinguished as it results in significant distortion of source position.

5. CONCLUSIONS AND FURTHER WORK

An efficient approach to spatial audio coding, called S\textsuperscript{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 localization of virtual sources. When compared with a state-of-the-art spatial audio coder, S\textsuperscript{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


