1-1-2013

A general compression approach to multi-channel three-dimensional audio

Bin Cheng
University of Wollongong, bc362@uow.edu.au

Christian Ritz
University of Wollongong, critz@uow.edu.au

Ian Burnett
RMIT University, ianb@uow.edu.au

Xiguang Zheng
University of Wollongong, xz725@uowmail.edu.au

A general compression approach to multi-channel three-dimensional audio

Abstract

This paper presents a technique for low bit rate compression of three-dimensional (3D) audio produced by multiple loudspeaker channels. The approach is based on the time-frequency analysis of the localization of spatial sound sources within the 3D space as rendered by a multi-channel audio signal (in this case 16 channels). This analysis results in the derivation of a stereo downmix signal representing the original 16 channels. Alternatively, a mono-downmix signal with side information representing the location of sound sources within the 3D spatial scene can also be derived. The resulting downmix signals are then compressed with a traditional audio coder, resulting in a representation of the 3D soundfield at bit rates comparable with existing stereo audio coders while maintaining the perceptual quality produced from separate encoding of each channel.

Keywords

audio, approach, dimensional, compression, general, three, channel, multi

Disciplines

Engineering | Science and Technology Studies

Publication Details

B. Cheng, C. Ritz, I. Burnett & X. Zheng, "A general compression approach to multi-channel three-dimensional audio," IEEE Transactions on Audio, Speech and Language Processing, vol. 21, (8) pp. 1676-1688, 2013.

A General Compression Approach to Multi-Channel Three-Dimensional Audio

Bin Cheng, Christian Ritz, *Senior Member*, *IEEE*, Ian Burnett, *Senior Member*, *IEEE* and Xiguang Zheng

***Abstract* — This paper presents a technique for low bit rate compression of three-dimensional (3D) audio produced by multiple loudspeaker channels. The approach is based on the time-frequency analysis of the localization of spatial sound sources within the 3D space as rendered by a multi-channel audio signal (in this case 16 channels). This analysis results in the derivation of a stereo downmix signal representing the original 16 channels. Alternatively, a mono-downmix signal with side information representing the location of sound sources within the 3D spatial scene can also be derived. The resulting downmix signals are then compressed with a traditional audio coder, resulting in a representation of the 3D soundfield at bit rates comparable with existing stereo audio coders whilst maintaining the perceptual quality produced from separate encoding of each channel.**

Index Terms — Audio Coding, 3D Audio

I. INTRODUCTION

Over the last two decades, research into audio coding technologies has led to significant achievements in compressing mono/stereo signal formats. Successful techniques, such as MP3 and AAC [1]–[4], have been widely used over the globe on all kinds of electronic devices, e.g. computers, mobile phones, portable audio players, etc. It has changed how people

Manuscript received November 3, 2012. This work has been partly supported by the Australian Research Council (ARC) through the grant DP1094053.

Bin Cheng was with the ICT Research Institute and School of Electrical Computer and Telecommunications Engineering, University of Wollongong, NSW2500, Australia. He is now working as a Senior Research Engineer at Dolby Laboratories (Beijing). (email: bchen@dolby.com)

Christian Ritz and Xiguang Zheng are with the ICT Research Institute and School of Electrical Computer and Telecommunications Engineering, University of Wollongong, NSW2500, Australia (email: critz@uow.edu.au and xz725@uow.edu.au,).

Ian Burnett is a Professor in School of Electrical and Computer Engineering, RMIT University, Australia. (email: ian.burnett@rmit.edu.au)

store, distribute and listen to music or other audio content by providing perceptually lossless (or nearly lossless) auditory quality at a low bandwidth cost. Multi-channel audio formats, such as ITU-5.1 [5], which includes 5 channels of standard loudspeaker audio plus a low frequency effects channel, were introduced about 15 years ago to provide the ability to reproduce sound source directions over the entire horizontal (2D) plane and therefore provide listeners with an enhanced audio experience for applications such as cinema and home theatre. Approaches to compress such 2D multichannel audio signals were introduced, e.g. Dolby AC-3 [6], DTS [7] to reduce the storage and transmission requirements for these multiple audio channels. This was achieved through downmixing approaches, however the bit rate of these approaches increase significantly as the number of channels increase.

An overview of recent approaches to further reducing this bit rate is provided in [8]. Techniques such as Parametric Stereo (PS) [9], Binaural Cue Coding [10], [11] and the MPEG Surround standard [12], [13] are based on encoding parametric models of spatial sound perception as well as parameters representing mathematical relationships between multichannel loudspeaker signals. Spatial Audio Scene Coding (SASC) [14], [15] and Directional Audio Coding (DirAC [16], [17]), also based on models of spatial sound perception, encode parameters representing spatial location information of virtual sound sources that are identified within a scene. The virtual sound sources, which are time-frequency components perceived by a listener as a single spatial sound source, are identified through processing the loudspeaker signals [14], [15] or derived from microphone array recordings [16], [17]. Spatially Squeezed Surround Audio Coding (S^3AC) [18], which is also based on estimating the position of virtual sources, achieves compression by mapping the positions of these sources from their original position in the 360° soundfield to a position in a 60° soundfield as represented by a stereo downmix signal. Compared to the approaches of [10]–[17], this approach does not require the transmission of side information representing the spatial attributes of the soundfield [8], [18].

While most of this existing research has focused on coding for multichannel 2D spatial audio, particularly 5.1 channel surround sound audio, there is less research into efficient compression of multichannel 3D spatial audio. In comparison, 3D video technologies are being more widely deployed in the market and attract increasing interest from consumers, e.g.

3D movies, 3D TV. Existing techniques for reproducing 3D audio using loudspeakers include Ambisonics [19] and Wave Field Synthesis [20]. There are now 3D audio solutions in the cinema, such as Dolby ATMOS [21], which brings significant increased interest in using true 3D multichannel audio systems in the market. All of these techniques require significantly higher numbers of channels than those required for 2D audio, e.g. 16-channel in 3D using Ambisonics [22]. Hence, there is an increasing need for solutions to compressing multichannel audio to complement these 3D video applications.

The multichannel 3D audio coding approach presented here is conceptually based on existing spatial audio coding techniques that encode sound sources and information about their location [14], [16], [18]. While an alternative might be to adopt the approach of [12], [13], extending this to many more channels than 5 may lead to distortion and degradation of audio quality [8] as it is optimized for decoding on the same loudspeaker setup as used in the encoding stage [23]. Existing research has investigated the benefits of applying S^3AC (conceptually similar to [14], [16]) for coding of 2D multichannel audio signal [18], [24]–[26]. In [18] it was shown that S^3AC is more accurate at maintaining the location of sound sources within a 5.1 channel audio scene when compared to two operating modes of the MPEG Surround standard, with the most significant improvements shown when compared to the MPEG Surround non-guided mode (similar to S^3AC, this mode does not transmit side information additional to the downmix signal). While these objective results could be explained by the differences in spatial audio models used in each coder (S^3AC is based on panning coefficients derived for virtual time-frequency sources rather than binaural models of spatial sound perception as used in MPEG Surround), subjective listening tests were also conducted to compare the localization quality. Results from these tests of compressed multichannel audio containing mostly localized sound sources confirmed that S^3AC provides comparable quality to MPEG Surround in stereo downmix mode and superior quality to MPEG Surround non-guided mode. For operating in mono-downmix mode, perceptual-based quantization techniques were proposed and subjectively validated in [24]–[26]. Hence, this paper further extends these techniques to 3D audio coding.

The key contributions of this paper include: an investigation of generalized orthogonal analysis techniques for

deriving the location of virtual sound sources through joint processing of loudspeaker signals representing 3D audio; the introduction of the Spatial Localization Quantization Point (SLQP) method for encoding virtual 3D sound source locations; the extension of spatial squeezing to encoding 3D multichannel audio as a stereo or mono downmix signal; and objective and subjective listening test results evaluating the performance of the technique for compressing 16-channel 3D audio.

Section II of this paper describes the extension of S^3AC to 3D for the virtual source localization estimation. Section III presents a new 3D source localization quantization, downmix generation and 3D soundfield reproduction method. Section IV presents both objective and subjective evaluation results of the proposed technique, applied to a database of 16-channel 3D spatial audio signals. Section V draws conclusions and discusses possible further work.

II. Spatial Squeezing for 3D Multi-channel Audio

The compression framework proposed in this section extends the S^3AC spatial squeezing approach [24] and psychoacoustic-based cue quantization [25] to 3D soundfields.

A. System Overview

The proposed compression scheme is aimed at low bit rate and backward compatible transmission of 3D soundfields rendered by an arbitrary number of loudspeakers, where minimum sound source localization distortion is desired. Backward compatibility is defined here as providing a downmix signal that can be compressed with an existing standard audio coder and stored in the corresponding file format. In the proposed encoding system, as shown in Fig.1, an orthogonal localization analysis is applied to estimate the location of 'virtual sources', which are defined here as individual (or groupings) of time-frequency components of the input multichannel audio signals. This is followed by quantization of derived 3D spatial localization, which exploits the perceptual localization redundancy for bit-rate efficiency as well as a further S^3AC spatial squeezing analysis.

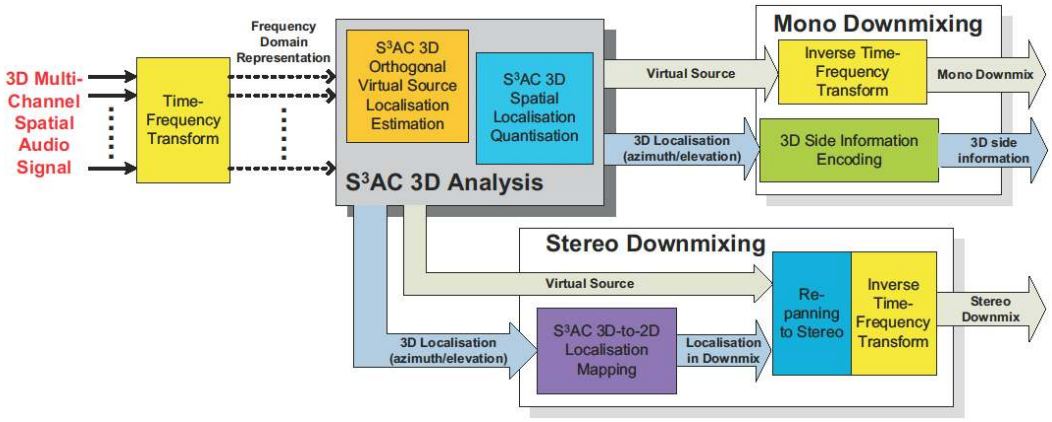

Fig.1. : S3AC 3D Encoding System

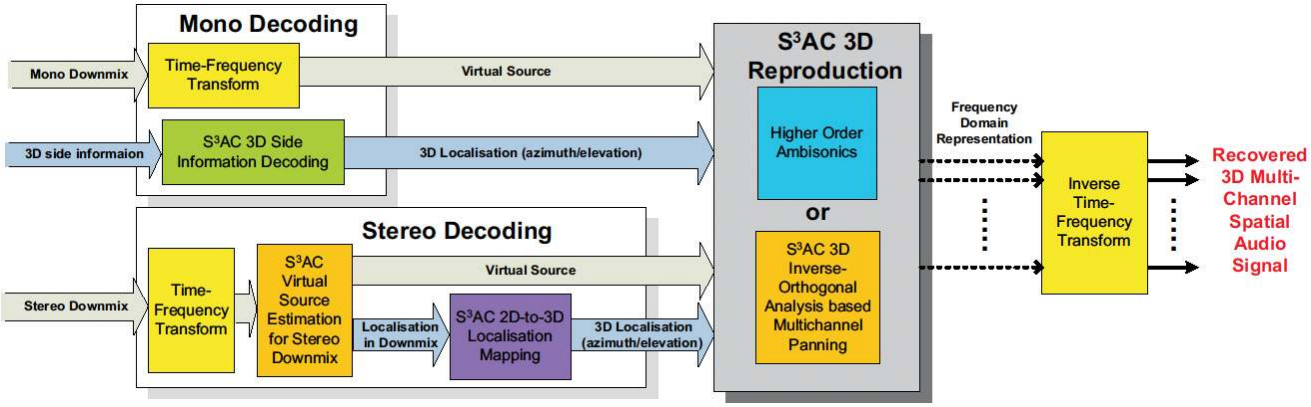

Fig.2.S^{3}AC 3D Decoding System

The resulting quantized 'virtual' 3D sound sources can be saved in an S^{3}AC stereo downmix, with the stereo downmix soundfield intelligently designed to have a unique mapping between Spatial Localization Quantization Points (SLQPs) in the original 3D soundfield and the downmixed localization points in the downmix stereo soundfield. The quantized virtual 3D sound sources can also be saved as a mono downmix while the SQLP information is saved as side information, which can be further quantized to reduce the bit rate. Based on this encoding approach, a 3D S^{3}AC decoder, as illustrated in Fig.2, can derive a sound source with its direct localization information in a 3D soundfield from either a 3D S^{3}AC stereo downmix or a mono downmix with accompanying 3D localization side information. This 'source + 3D localization' format provides flexibility in that any major 3D reproduction method can be applied. Hence, a user can choose the same method as used in producing the original multichannel signals ensuring minimum distortion, or the most appropriate rendering solution that meets the application's requirement.

B. Orthogonal Analysis for Deriving Time-Frequency Virtual Sources and their Azimuth-Elevation

S^3AC has been proposed in [24] for 2D soundfield compression where two dominant time-frequencies among five channels are considered to generate the virtual source. Here, the 2D S^3AC has been extended for compressing 3D soundfields where a significant number of channels are considered (in the evaluation 16 channels are used). The extended orthogonal analysis described in this section aims to jointly consider all of the channels to form the virtual source. The proposed method starts with a time-frequency decomposition applied separately to each of the N channels of the input spatial audio signal. Any modern time-frequency decomposition can be used such as a short-time Fourier transform (STFT) [1], [18] or a Pseudo Quadrature-Mirror Filterbank (PQMF) with further perceptual bank decomposition (as used in MP3/AAC [2], [3]). Consider a loudspeaker positioned at azimuth μ_i, elevation η_i, as shown in Fig.3. The resulting time-frequency representation of the i^{th} loudspeaker signal $p_i(k,n)$ can be decomposed into x, y and z components as:

$$p_i(k,n) = g_i(k,n) \cdot \begin{bmatrix} \vartheta_i \\ \varepsilon_i \\ \varsigma_i \end{bmatrix} \tag{1}$$

where $g_i(k,n)$ represents the gain of this loudspeaker, $[\vartheta_i \;\; \varepsilon_i \;\; \zeta_i]^T$ is the unit vector representing the loudspeaker location, $\vartheta_i = \cos\mu_i \cdot \cos\eta_i$, $\varepsilon_i = \sin\mu_i \cdot \cos\eta_i$, $\zeta_i = \sin\eta_i$ and k and n are frequency and temporal frame indices, respectively. Assuming a given virtual time-frequency source is reproduced by N loudspeakers, the overall source level $g_s(k, n)$ is given by:

$$g_s^{\,2}(k,n) = \left[\sum_{i=1}^{N} g_i(k,n) \cdot |\vartheta_i|\right]^2 + \left[\sum_{i=1}^{N} g_i(k,n) \cdot |\varepsilon_i|\right]^2 + \left[\sum_{i=1}^{N} g_i(k,n) \cdot |\varsigma_i|\right]^2 \tag{2}$$

where only the magnitude of location vectors are considered in (2) in order to avoid front/back, left/right energy cancelation, which does not occur in natural 3D sound.

The source signal $S(k,n)$ can be generated by applying the phase information $e^{\phi M}$ chosen from the channel with the highest amplitude (defined as the M^{th} channel) in order to maintain phase consistency:

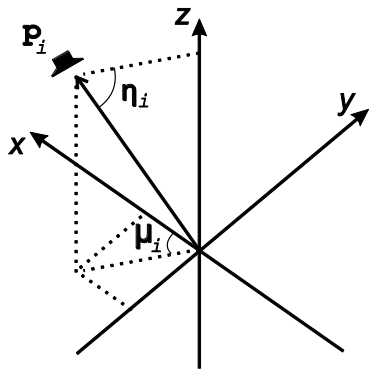

Fig.3. 3D Loudspeaker Signal Positioned at azimuth μ, elevation η

$$S(k,n) = \sqrt{g_s^2(k,n)} \cdot e^{\phi M} \tag{3}$$

Meanwhile, the azimuth $\mu_s(k,n)$ and elevation $\eta_s(k,n)$ of the source can be derived by analyzing its orthogonally decomposed components on the x, y and z axis as:

$$\tan \mu_s(k,n) = \frac{\sum_{i=1}^{N} g_i(k,n) \cdot \vartheta_i}{\sum_{i=1}^{N} g_i(k,n) \cdot \varepsilon_i} \tag{4}$$

$$\tan \eta_s(k,n) = \frac{\sqrt{\left[\sum_{i=1}^{N} g_i(k,n) \cdot \vartheta_i\right]^2 + \left[\sum_{i=1}^{N} g_i(k,n) \cdot \varepsilon_i\right]^2}}{\sum_{i=1}^{N} g_i(k,n) \cdot \varsigma_i} \tag{5}$$

C. Comparison with Existing Approaches

The proposed approach can also be viewed as a generalization of amplitude panning methods such as classical tangent panning and Vector Base Amplitude Panning (VBAP) [27]. Specifically, 2D amplitude panning is equivalent to the proposed method with $\eta_i = 0$ and $N = 2$ in equations (2)-(4), whilst 3D VBAP is equivalent to the proposed method when $N = 3$ in equations (2)-(4). Note that the derivation of equations (1) to (3) is also conceptually similar to the orthogonal analysis approach used in DirAC [16], [17], [23]. The derivation of the spatial parameters ($\mu_s(k,n)$ and $\eta_s(k,n)$) using (4) and (5) is similar to the approach used to identify the directions of time-frequency components as used in [28]. DirAC has recently been extended to the creation of virtual B-Format recordings from surround sound loudspeaker signals [29], [30], using an approach similar to (2) but using N=5, an elevation of zero and factors to create a compatible

first order B-format recording.

One difference with the approach described here is that it does not attempt to decompose the virtual time-frequency sources into directional and diffuse components. While such decomposition could allow for the application of different quantization techniques for the direct and diffuse components, this was not evaluated here. Further, the accurate estimation and rendering of diffuse components can be difficult [30] and as suggested in [29], it is expected that diffuseness will be represented by the variability of the direction estimates for these time-frequency components. A further difference to the evaluations provided in [29], [30] is that here an investigation into the performance for compressing 16 loudspeaker channels designed to reproduce 3D sound is presented. While the approaches of [29] [30] are not theoretically restricted to 2D, results were restricted to evaluating 5 channel audio material. Another key difference in the proposed work is the application of the spatial squeezing technique [18] for embedding the spatial location information within a stereo downmix signal, which is further described in the next section. Finally, the approach in this paper could also be modified to be compatible with SASC [14], [15] in a similar way to that described for DirAC [23]. A key difference is that SASC is based on an estimating an optimized primary and ambient decomposition using Principal Component Analysis (PCA) applied to STFT representation of the loudspeaker channels [14], [15]. By not aiming to decompose the audio scene into ambient and primary components, the approach described in this paper could be regarded as less complex than DirAC or SASC. Similar to the proposed approach, directional analysis of the sound scene without separating into primary and ambient components has previously been proposed for binaural synthesis[31]. By avoiding the primary-ambient decomposition, the authors' claim that this simplifies the approach whilst assuming that diffuseness is implicitly represented by the variability found in the time-frequency directional estimates [31]. While not directly investigated, such an assumption could also be made for the approach in this paper.

III. QUANTIZATION AND SQUEEZING 3D MULTI-CHANNEL AUDIO

This section describes how the squeezed recordings, including the spatial location information, can be efficiently

encoded and decoded.

A. Spatial Localization Quantization Points

The 3D orthogonal analysis algorithm presented in the Section III-B derives the azimuth and elevation localization information of the 3D sound source rendered by a number of loudspeakers in a 3D soundfield. The precision of the derived azimuth/elevation is based on the input signal bit precision, e.g. 16 bits. This can be reduced for higher bit rate efficiency, similar to the cue quantization approach used in 2D S^3AC coding [25]. Psychoacoustic research shows that in the frontal plane, which is the most sensitive listening area, the human auditory system has approximately a 1° azimuth and 5° to 10° elevation resolution, respectively, for localizing tonal sources with frequency components most sensitive to the human ear [32]. This phenomenon, also referred as localization blur [32], is exploited in 2D S^3AC coding to achieve bit rate efficiency. During the S^3AC compression of 3D soundfields, both perceptual azimuth resolution and perceptual elevation resolution are exploited, so as to effectively describe a continuous 3D sphere by a discrete number of localization points whilst ensuring minimum loss in perceptual localization. These points are defined as the Spatial Localization Quantization Points (SLQP).

Each SLQP consists of localization information described as source azimuth/elevation. Two example SLQP designs are described here, which have comparatively higher and lower quantization precision, as shown in

Fig.4 and

Fig.5. Based on the available experimental facilities [22], both examples are designed for an upper-hemisphere 3D sound scene, and used for evaluation in Section IV. The two designs are described as follows:

- Based on perceptual experiments on quantization of 2D S^3AC side information in [24], an azimuth precision of 2° (3° for low precision) is used for the 0° elevation plane.
- Compared with the location dependent azimuth quantization precision of S^3AC side information described in previous 2D S^3AC approaches [24], [25], a uniform azimuth quantization precision is used here, since in 3D surround sound scenarios, listeners may turn their head or body to fully exploit the impression of a 3D surround audio scene.

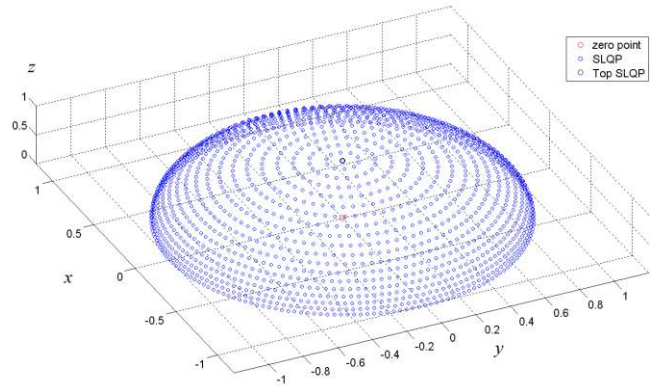

Fig.4. S[3]AC SLQP with High Precision

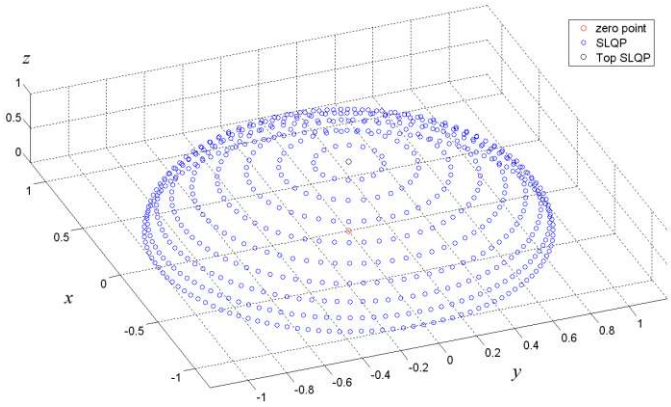

Fig.5. S[3]AC SLQP with Low Precision

- However, as in some scenarios listener head rotation may be limited, the proposed quantization technique represents the upper limit on bit rate that would be required.
- Based on the elevation resolution suggested by psychoacoustics [32], a 5° (10° for low precision) elevation resolution is utilized. This effectively results in SLQPs with parallel layers of different elevations.
- As the perceptual localization precision degrades when the elevation of a source increases away from the 0° elevation horizontal plane, the azimuth precision used for quantization is decreased with increasing elevation. Specifically, in the higher precision example, the azimuth quantization precision used for the adjacent higher layer is reduced by 10 quantization points, e.g. while the 0° elevation layer has 180 SQLPs, the 5° elevation layer has 170 SLQPs. In the lower precision example, the quantization precision reduction between layers is 12 points.
- A 90° elevation SLQP is reserved for both designs and it's the only SLQP on the 90° elevation layer.

- The resulting numbers of SLQPs are 1729 and 658, for the high and low precision design, respectively. Hence, it requires 10.7 bits and 9.4 bits to encode each derived SLQP, for the high precision and low precision design respectively.

Note that, these two SLQP examples are designed based on a linearly decreasing resolution with increasing elevation, for evaluating the proposed 3D audio coding approach. In practical applications, SLQP design with a 'power of 2' number of SLQPs (e.g. 1024 SLQPs) can be adopted for better bitrate efficiency. By exploiting available psychophysical theory and experimental results, the goal of the two SLQP designs described above is to ensure minimum perceptual localization distortion while efficient bandwidth reduction can be achieved. The performance of these designs will be evaluated and justified, both objectively and subjectively, in Section IV.

B. Stereo Downmixing based on Spatial Squeezing

A key novelty of 2D S^{3}AC is the ability to implicitly encode the spatial information within a stereo downmix signal that is compressed with a standard audio coder such that the separate encoding and storage of side information is not needed. This could be attractive for transmitting 3D audio content using existing file formats, which could then be flexibly reproduced in 2D or 3D using software on a client. This is further investigated here. Section IV will compare results for the stereo downmix approach with the mono-downmix approach described in Section III.C.

Similar to the approach to re-pan the virtual source to a unique position in the stereo downmix described in 2D S^{3}AC compression, each SLQP derived in Section III-A is given a unique mapping into a 60° stereo downmix soundfield. The approach adopted here is to uniformly divide the 60° downmix soundfield into discrete azimuths according to the number of SLQPs in the 3D soundfield. For instance, to save the high SLQP design described in Section III-A and

Fig.4, which has 1729 points, two adjacent downmix localization points then have approximately a 0.035° azimuth discrimination. This can be expressed mathematically by:

$$\varphi_{dm}\left(k,n\right) = f\left(SLQP\left(k,n\right)\right) \tag{6}$$

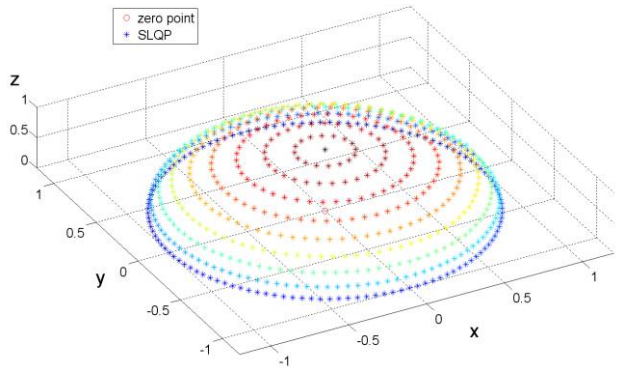

Fig.6. S^{3}AC SLQP with Low Precision while each SLQP is distinguished by Color

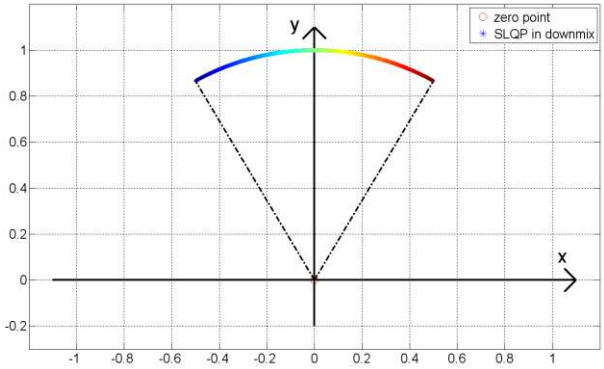

Fig.7. S^{3}AC SLQP in a Stereo

where $\varphi_{dm}(k,n)$ is the assigned azimuth in the 60° downmix soundfield and $f(\cdot)$ represents the 3D SLQP to 2D soundfield mapping methodology, which can be defined by the user. An illustrative example of this 3D-to-2D localization mapping is given in Fig.6 and Fig.7. Fig.6 shows the low precision SLQP design as described in Section III-A, while each SLQP layer is distinguished by color. Fig.7 illustrates an example localization mapping to generate a stereo downmix where each SLQP in the 3D soundfield is given a unique azimuth in the 2D 60° soundfield. This is indicated by different colors in Fig. 7 for the different layers of (azimuth, elevation) points in Fig. 6. For example, given that the 60° region is divided into 658 points for the low precision SLQP, the lowest layer of Fig. 6 containing 120 points is then mapped to the region from -30° to approximately 10.9°. Similarly, as in 2D S^{3}AC coding, the stereo downmix is generated by amplitude panning the derived virtual source from (3) to the location in the stereo downmix derived from (6):

$$L_{dm}(k,n) = \frac{S(k,n)\cdot\left[\tan 30° + \tan\left(\varphi_{dm}(k,n)\right)\right]}{\sqrt{2\cdot\tan^2 30° + 2\cdot\tan^2\left(\varphi_{dm}(k,n)\right)}}$$
$$R_{dm}(k,n) = \frac{S(k,n)\cdot\left[\tan 30° - \tan\left(\varphi_{dm}(k,n)\right)\right]}{\sqrt{2\cdot\tan^2 30° + 2\cdot\tan^2\left(\varphi_{dm}(k,n)\right)}} \quad (7)$$

This is followed by an inverse time-frequency transform for a backward compatible time domain stereo signal representation.

Previous research in [26] investigated the required azimuth resolution in the S^{3}AC stereo downmix soundfield and its relationship to the virtual source amplitude. Equivalently, the number of derivable azimuths in the S^{3}AC downmix is:

$$\left\| g_s\left(k,n\right) \right\| + 1 \tag{8}$$

where $\| \cdot \|$ stands for rounding to the nearest integer and $g_s(k, n)$ is the amplitude of the derived virtual source. Although a derived virtual source amplitude higher than 1729 (approximately -25.5dB for a signal with 16-bit quantization level and defining the level of 2^{16} as 0 dB) is required for maintaining adequate downmix azimuth resolution, by building the localization mapping from 3D to stereo downmix such that adjacent SLQPs in 3D are given adjacent downmix azimuths, it is ensured that, for sources having amplitudes lower than the required amplitude, the recovered localization in 3D has minimum deviation from the original position. The impact of this approach on the localization accuracy is further evaluated in Section IV.

Based on this, a stereo downmix containing squeezed localization information of a 3D surround sound field can be further compressed by conventional perceptual audio coders, such as AAC, which results in bit rates equivalent to conventional stereo audio for transmitting a 3D spatial soundfield, e.g. 128kbps. Since no side information is required, a further advantage is the ability to store the compressed downmix as an AAC compatible audio file. While not evaluated here, the approach is not restricted to a specific mapping approach and hence alternative mapping approaches could also be investigated. In the case of a stereo downmix, the mapping functions of S^{3}AC applied to 5 channel audio could be modified such that the virtual sources are mapped based on standard downmix equations [5]. Similarly, for 3D, the mapping table used for the SLQPs of Fig. 6 covering the original 360° could be modified so that sources in the left and right halves of the hemisphere are panned to the left and right halves of the stereo soundfield, respectively. These downmix equations could also be based on artistic preferences, such as discussed in [13]. It should be noted that the

primary purpose of the squeezing approach adopted here is to enable efficient compression of the 3D spatial information and previous research has shown that fixed downmix equations such as used in Dolby Prologic does not provide the most efficient compression scheme for spatial audio [13]. While outside the scope of this paper, a further investigation into alternative downmix approaches for 3D sound that result in subjectively acceptable playback on stereo systems and the impact of these approaches on compression is recommended.

C. Mono Downmixing and Differential Quantisation of the SLQPs

The SLQP derived in Section III-C can also be saved as accompanying 3D localization side information for a mono downmix generated by the virtual sound source given by (2). According to the evaluation presented [25], a frame-wise localization differential coding based on a code-book representation of the derived source localization can efficiently reduce the bitrate of spatial side information without introducing any distortion. In this work, similarly to the algorithm described in [25], the SLQP set representing a 3D soundfield is transformed into a codebook representation, where each SLQP has a unique index in the codebook as:

$$I(k,n) \tag{9}$$

The codebook distance between two adjacent time frames of a frequency component is calculated as:

$$E(k,n) = I(k,n) - I(k,n-1) \tag{10}$$

To ensure tolerance to transmission errors, the code book index in (9) is recorded every number of frames, similar to the approach described in [25], while the differential sequence of (10) is entropy coded, e.g. using Rice Coding [33]. While this differential coding approach applied in 2D S^{3}AC has been investigated in [25], it is not the objective of this paper to further evaluate this approach.

D. Decoding

For a received S^{3}AC 3D stereo downmix, following a time-frequency transform, the decoder performs a virtual source localization in the 60° stereo downmix soundfield by inverting (7), where a virtual source $\hat{S}(k,n)$ and its azimuth

$\hat{\varphi}_{dm}(k,n)$ in the 60° downmix soundfield is recovered as:

$$\hat{S}(k,n) = \sqrt{\hat{L}^2_{dm}(k,n) + \hat{R}^2_{dm}(k,n)} \cdot e^{\phi_{dm}} \tag{11}$$

$$\hat{\varphi}_{dm}(k,n) = \tan^{-1}\left[\frac{\hat{L}_{dm}(k,n) - \hat{R}_{dm}(k,n)}{\hat{L}_{dm}(k,n) + \hat{R}_{dm}(k,n)} \cdot \tan(30^\circ)\right] \tag{12}$$

where $\hat{L}_{dm}$ and $\hat{R}_{dm}$ are the decoded downmix channels and the phase parameter $e^{\phi_{dm}}$ is chosen as the phase information from the channel with the highest energy of the two stereo channels. This is followed by inverting (6) for re-mapping the localization to an SLQP in the 3D soundfield, such that:

$$\hat{SLQP}(k,n) = f^{-1}\left(\hat{\varphi}_{dm}(k,n)\right) \tag{13}$$

Note that the compression of the stereo downmix by the perceptual audio means that the decoded SLQP of (13) may not perfectly match the original SLQP of (6). Further, the quantization noise introduced by the coder will be higher for masked compared to non-masked frequency components. However, (12) is based on the amplitude relationships between the two channels. It was shown in [26] for 2D audio that the minimum signal amplitude required to accurately encode the spatial locations is well above the masking levels. One issue that could arise is when a frequency component masked by an adjacent frequency component in one channel is severely distorted due to excessive quantization noise compared to the corresponding frequency component in the other channel. This could arise if, for example, adjacent frequencies belong to different virtual sources with differing directions and hence different masking curves are derived for each channel. In practice, this problem could be minimized by operating the perceptual audio coder at a higher bitrate (in this paper AAC at 64 kbps/channel is adopted).

Alternatively, for maximum accuracy in decoded spatial locations, the S^{3}AC 3D mono downmix mode can be used. This mode results in a sound source and accompanying decoded SLQP containing localization azimuth/elevation information. The compressed 3D side information is decoded to recover the SLQP in the 3D soundfield and the mono downmix is decomposed into a frequency domain representation. A comparison of the localization accuracy of both downmix approaches is provided in Section IV.

E. Reproduction

Based on the available reproduction facility (see Section IV) with a symmetrical loudspeaker array, higher order Ambisonics reproduction is used for 3D soundfield decoding and reproduction. To reproduce a source signal using m^{th} order Ambisonics in an array with m^2 loudspeakers, the source signal is firstly encoded as [22]:

$$B(k,n) = s(k,n) \cdot y\left(\mu(k,n), \eta(k,n)\right) \tag{14}$$

where $s(k,n)$, $\mu(k,n)$ and $\eta(k,n)$ are the time-frequency representation of the virtual source, and its azimuth/elevation in the 3D soundfield, respectively, while y($\mu(k,n)$, $\eta(k,n)$) is defined as a vector containing spherical harmonics functions as:

$$y\left(\mu(k,n), \eta(k,n)\right) = \left[Y_1\left(\mu(k,n), \eta(k,n)\right), Y_2\left(\mu(k,n), \eta(k,n)\right), \ ..., Y_{(m+1)^2}\left(\mu(k,n), \eta(k,n)\right)\right] \tag{15}$$

where $Y_n(\mu(k,n), \eta(k,n))$ are the higher order spherical harmonics functions [22]. The encoded source signal is transformed into a loudspeaker signal matrix *LS(k,n)*, defined as:

$$LS(k,n) = D \cdot B(k,n) \tag{16}$$

where D is the pseudo-inverse of the re-encoding matrix C such that:

$$D = pinv(C) = C^T \cdot \left(C \cdot C^T\right)^{-1} \tag{17}$$

The re-encoding matrix C is defined based on the configuration of the reproduction loudspeakers system as:

$$C = \left[c_1, c_2, ..., c_i, ..., c_N\right] \tag{18}$$

where i is the speaker index and N the total number of loudspeakers in the reproduction system, vector c_i defined as the series of spherical harmonics functions similar as in (15) but based on the azimuth μ_i and elevation η_i of the i^{th} loudspeaker, such that:

$$c_i = \left[Y_1\left(\mu_i, \eta_i\right), Y_2\left(\mu_i, \eta_i\right), ... Y_{(m+1)^2}\left(\mu_i, \eta_i\right)\right] \tag{19}$$

Note that compared with the sound source azimuth $\mu(k,n)$ and elevation $\eta(k,n)$ derived on a time-frequency basis, the azimuth μ_i and elevation η_i for the i^{th} loudspeaker is fixed as long as the loudspeaker configuration remains unchanged.

IV. EVALUATIONS

The proposed S^{3}AC 3D spatial audio compression technique, including both the stereo and mono downmix modes, is evaluated both objectively and subjectively in this section. For this purpose, the algorithms and methodologies presented in this paper are implemented based on a 16-channel hemisphere loudspeaker array designed by G. Potard etc. (see Section 5.4.1 of [22]), called the Configurable Hemisphere Environment for Surround Sound (CHESS). This system is pictured in Fig.8 and the loud speaker positioning configuration is described in Table I. A detailed description of the loudspeaker location setting in CHESS can be found in [22].

A. 16-Channel 3D Audio Files for Evaluation

For evaluation of the proposed S^{3}AC compression of 3D audio, based on the 16-channel CHESS loudspeaker system, eight 16-channel 3D audio signals with differing audio content were created. Reproduction of the content in an anechoic room was achieved using the method described in Section III.D and [22] to derive the loudspeaker signals for 4th order Ambisonics audio reproduction. A set of files were created to provide a range of 3D auditory experiences. These files are described in the following:

- **File1**. A clear male speech signal presenting the loudspeaker channel number (from 1 to 16) panned to only that channel. Each channel is spoken sequentially in order and the total duration is approximately 22 seconds. The locations of the channels and hence intended source directions are provided in Table I.
- **File2**. An airplane moving over-head from the rear right (at -144° azimuth) to the front left (at 36° azimuth). This was synthesized using amplitude panning and online tuning of the reproduction parameters to ensure the designed source movement is perceptually achieved. The duration is approximately 17 seconds.
- **File3**. Male speech in the presence of ambient noise. The male speech is rendered by a single loudspeaker using a recorded speech sentence and ambient noise is simulated using noise recorded at a busy restaurant that is equally panned to all 16 channels. The speech file is sequentially played from different locations (rendered by a single

loudspeaker) throughout the duration of the file, which is approximately 25 seconds. (Note that the noise recordings used in Files 3, 4 and 5 are not identical.)

- **File4**. Ambient noise simulated using the same approach as for File 3 but without any localized content and a different recording noise at a restaurant. The duration is approximately 30 seconds. This file was designed to evaluate the coding quality for ambient noise only.
- **File5**. A moving source in the presence of ambient noise simulated using the same approach as for File 3 but with a third recording of restaurant noise. The resulting audio is perceived as a directional source moving around the listener. The duration is approximately 32 seconds. This file was designed to evaluate mono sources whose location smoothly varies in the presence of ambient noise.
- **File6**. A 4^{th} order 16-channel reproduction of an Ambisonics B-format recording, featuring music and localized sound sources (original available from [34]). The duration is approximately 20 seconds. To create the loudspeaker signals, 4^{th} order Higher Order Ambisonics (HOA) encoding equations specified in [22] (based on [20]) were first applied to the source signals and spatial information derived from the B-format signals using the (14) and (15) of Section III.E. The resulting HOA Ambisonic signals were then decoded to obtain the 16 loudspeaker signals using (16)-(19). This approach was also used for Files 7 and 8.
- **File7**. A 4^{th} order 16-channel reproduction of an Ambisonics B-format recording, featuring the sound of percussion instruments whose location varies over time (original file obtained from [35]. The duration is approximately 18 seconds.
- **File8**. A 4^{th} order 16-channel reproduction of an Ambisonics B-format recording, featuring music and localized sound sources that also vary over time (original available from [36]). The duration is approximately 21 seconds.

These files are used for both objective and subjective evaluation presented in the following sections.

B. Objective Evaluation

File 2 described in Section IV-A is firstly used for evaluation of the localization accuracy after S^3AC 3D coding.

Table I: CHESS Loudspeaker Configuration

Channel	Azimuth	Elevation	Channel	Azimuth	Elevation
1	0	0	9	252	30
2	72	0	10	324	30
3	144	0	11	0	60
4	216	0	12	72	60
5	288	0	13	144	60
6	36	30	14	216	60
7	108	30	15	288	60
8	180	30	16	0	90

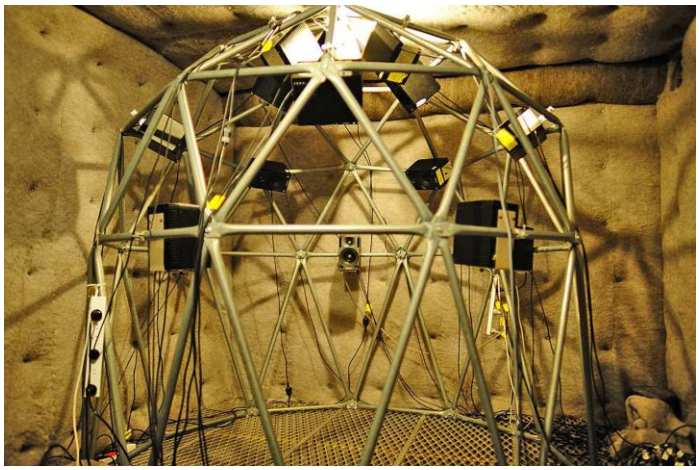

Fig.8. 16-Channel loudspeaker array CHESS in an anechoic environment

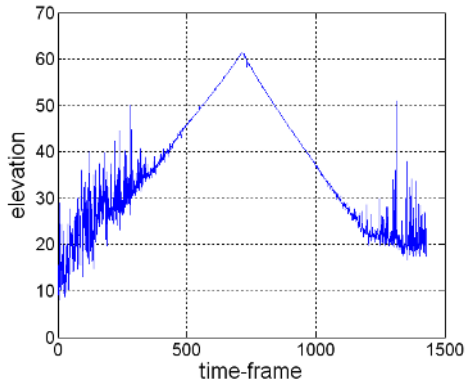

Fig.9. Elevation Feature of the 80th Frequency

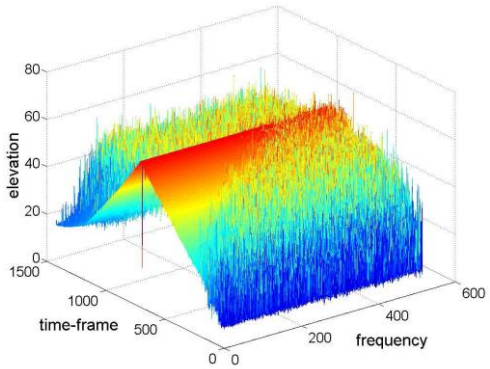

Fig.10. Time-frequency-elevation Mesh of the Original Signal

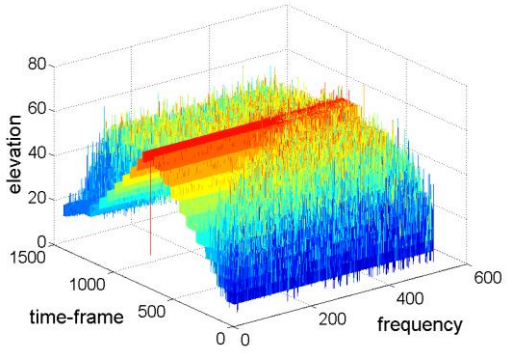

Fig.11. Time-frequency-elevation mesh of the signal encoded by S^{3}AC 3D mono downmix with High-precision SLQP quantization

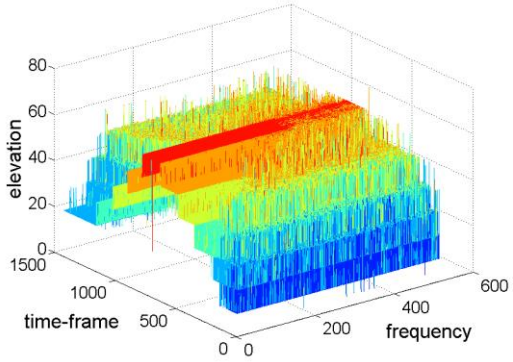

Fig.12 Time-frequency-elevation mesh of the signal encoded by S^{3}AC 3D mono downmix with Low-precision SLQP quantization

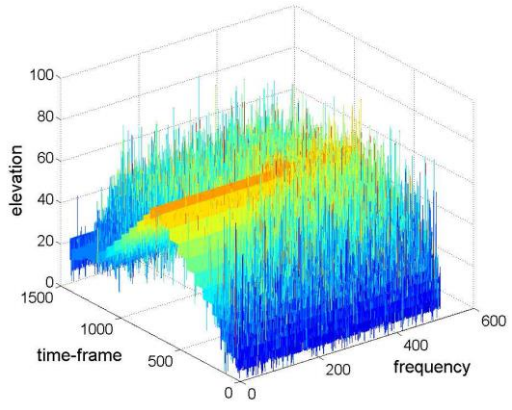

Fig.13 Time-frequency-elevation mesh of the signal encoded by S^{3}AC 3D stereo downmix, SLQP recovered from stereo downmix soundfield

This is a recording of an airplane with synthesized azimuth/elevation localization parameters. It is reproduced into a 16-channel signal by amplitude panning to ensure correct 3D rendering of an airplane flying over-head and moving from the rear right (-144° azimuth) to the front left (36° azimuth). This signal is processed using the 3D source localization orthogonal analysis presented in Section III based on a 1024-point 50% overlapping STFT transform, to derive the virtual time-frequency source locations in the 3D soundfield. While this primarily examines the objective error resulting from

the quantization of the spatial locations, this audio file was carefully prepared using manual adjustment of the amplitude panning such that informal listening tests confirmed the intended source directions using the adopted spatial rendering technique were perceptually achieved.

Fig.10 illustrates the derived source elevation feature on a time-frequency basis. Furthermore, Fig.9 shows the derived elevation of a single frequency, the 80^{th} frequency bin, which contains the highest amplitude information of the whole signal. The elevation noise at the beginning and end of the plot of Fig. 10 is due to a lack of signal energy for the 80^{th} frequency component in this region. Similar noise is evident in Fig. 12 as well as additional noise at other frequencies that is caused by a lack of energy in certain time-frequency regions. The derived source azimuth/elevation localization information is spatially quantized using the SLQP approach described in Section III, followed by both the stereo downmixing approach presented in Section III-D and the mono downmix approach presented in Section III-E.

The resulting quantized elevation using high-precision SLQP design is illustrated in Fig.11, while the result of using low-precision SLQP design is illustrated in Fig.12. In addition, based on the methodology presented in Section III-F, Fig.13 gives the elevation feature derived from the stereo downmix, which is synthesized by mapping the high-precision SLQP to the 60° stereo soundfield.

By comparing Fig.11, Fig.12 and Fig.13 with Fig.10, the localization and source movements in the original signal are estimated and recovered in the S^3AC 3D encoding/decoding process. Quantization distortion is introduced due to the SLQP quantization process, while the high-precision SLQP designed introduces less quantization distortion compared with the other two coding approaches.

Further analysis is performed by mathematically evaluating the localization error caused by the S^3AC 3D coding. Three different coding modes presented above, including stereo downmixing, mono downmixing with high precision SLQP design and mono downmixing with low precision SLQP, are evaluated for each of the original 16-channel 3D audio signals described in Section IV-A,. The error is calculated as:

$$\mu_{error} = \frac{1}{KM} \cdot \sum_{k=1}^{K} \sum_{n=1}^{M} \left\lfloor \mu_{original}(k,n) - \mu_{decoded}(k,n) \right\rfloor$$
$$\eta_{error} = \frac{1}{KM} \cdot \sum_{k=1}^{K} \sum_{n=1}^{M} \left\lfloor \eta_{original}(k,n) - \eta_{decoded}(k,n) \right\rfloor \quad (20)$$

where $\mu_{original}(k,n)$, $\eta_{original}(k,n)$ are the original time-frequency azimuths and elevations, respectively, derived by using the proposed 3D orthogonal localization analysis, $\mu_{decoded}(k,n)$, $\eta_{decoded}(k,n)$ are the time-frequency azimuths and elevations, respectively, derived at the decoder and K, M are the total number of frequency bins and frames respectively. The three proposed coding modes, including stereo downmix, mono downmix with high-precision SLQP and mono downmix with low-precision SLQP, are evaluated. The resulting average azimuth/elevation errors for each test file are given in Tables II to IV. While the resulting errors are file-dependent, it is shown that the stereo coding mode and high-precision SLQP mono coding mode have similar error performance. The low-precision mono coding mode introduces higher error. An average error of 5 degrees was observed, with the maximum error being approximately 15 degrees for File 3, which contains a speech source whose location varies from one loudspeaker to another during the file. The relatively higher azimuth errors for Files 1, 3 and 6 result from the quantization of the spatial location. As described in Section II.A, the SLQP utilizes lower azimuth resolution for sources with high elevation. Hence, it is important to note that the azimuth error would be much lower for these files for sources close to the horizontal plane. The resulting perceptual impact will be evaluated in the next Section.

It should be noted that the approach relies on the assumption that orthogonal analysis derives virtual source directions that correspond to the perceived directions of the virtual sources. To provide an indication of the objective localization error resulting from the orthogonal analysis, a subset of the test files were chosen that contained predominantly localized content (these were Files 1 and 2 and Files 6, 7 and 8). The error between the intended directions and those derived from orthogonal analysis are shown in Table V for time-frequency components with non-zero energy. With average errors less than 2 degrees, these results show that the orthogonal localization analysis is highly accurate for localized sources, with the resulting errors for these files close to the minimum perceivable errors [32]. For Files 3, 4 and

Table II: Average azimuth/elevation error (in degrees) from S^{3}AC 3D mono coding with high-precision SLQP design

File	1	2	3	4	5	6	7	8
Azimuth Error	11.84	0.08	15.10	0.16	2.70	7.35	1.19	4.60
Elevation Error	1.19	0.73	1.30	1.90	3.55	1.26	0.42	1.22

Table III: Average azimuth/elevation error (in degrees) from S^{3}AC 3D mono coding with low-precision SLQP design

File	1	2	3	4	5	6	7	8
Azimuth Error	12.17	0.70	15.37	0.22	5.36	7.64	1.29	4.89
Elevation Error	2.27	1.45	2.87	3.88	7.01	2.53	0.82	2.36

Table IV: Average azimuth/elevation error (in degrees) from S^{3}AC 3D stereo coding

File	1	2	3	4	5	6	7	8
Azimuth Error	13.03	0.06	15.04	0.08	1.75	7.56	3.81	3.85
Elevation Error	1.17	1.23	1.31	1.90	2.02	1.26	1.34	1.27

Table V: Average azimuth/elevation error (in degrees) from S^{3}AC 3D stereo coding

File	1	2	6	7	8
Azimuth Error	0.67	1.04	0.79	1.31	1.85
Elevation Error	0.09	0.06	0.16	1.08	1.75

5, which contain significant levels of ambient noise, direction estimates were more variable. This could be explained by the significant levels of ambient noise in these files and while not investigated here, accurate estimation of directional sources would require additional processing, such as separation into directional and diffuse components [23]. The overall perceived localization accuracy, resulting from both the orthogonal analysis and subsequent direction quantization, is evaluated in the next section.

C. Subjective Evaluation

The proposed S^{3}AC multi-channel 3D spatial audio compression system is further evaluated by subjective listening experiments. The eight 16-channel 3D audio files described in Section IV-A are used. Three proposed types of S^{3}AC 3D multichannel audio compression approaches were evaluated, including stereo downmix, mono downmix with high-precision SLQP, mono downmix with low-precision. The coding process is based on a 1024-point 50% overlapping STFT.

In the stereo downmix approach, azimuth/elevation localization information is derived for every frequency for

mapping into the 60° downmix soundfield. In the mono downmix approach, SLQP is also derived for every frequency and further quantized using either the high precision or low precision design. The high precision design results in a bit-rate of 474 kbps for the 3D side information, while the low precision design results in a bit-rate of 413 kbps. These bitrates are used for maximizing the coding performance in the mono downmix mode when compared to the stereo downmix mode. However, by further exploiting the fact that the human auditory system perceives a single location for sources within the same Equivalent Rectangular Band (ERB) [32]), the bandwidth required for transmitting 3D side information could be significantly reduced. This results in bit rates of 18 kbps and 16 kbps, for the high precision and low precision SLQP design, respectively, if one quantized SLQP is used for each double-ERB band, however the subjective evaluation of this additional quantization was not evaluated here. The downmix signals in all three modes are further coded by AAC with 64kbps/channel. The 4^{th} order Ambisonics reproduction method presented in Section III-F is used for reproducing the signal to 16-channels. Based on the 16-channel CHESS loudspeaker configuration, the 4^{th} order Ambisonics can ensure the highest localization reproduction precision by fully utilizing the available 16 loudspeakers.

A perceptual evaluation methodology based on MUSHRA [37] was utilized for the listening test. Besides the three proposed S^3AC 3D coding conditions, an AAC coding condition is incorporated for comparison purposes, where each channel in the original signal is coded individually with 128 kbps AAC, resulting in a total bit rate of 2048 kbps. All coded conditions are randomly mixed with a hidden reference and an anchor signal. Listeners were asked to compare the test files with the reference to judge both the auditory quality and spatial localization accuracy. All listeners were instructed to sit facing loudspeaker 1 and their chair height was adjusted to ensure their head was at the same level as the first ring of loudspeakers. Similar to [12], the anchor signal was created with a 3.5 kHz low-pass filtered version of the original signal, however here this is followed by mono-mixing to all channels (achieved by summing all channels equally) to remove localization. A total of 21 listeners took part in the experiment, including 6 experienced listeners (thus the majority of listeners were non-experts). Following the recommendations of [37], post-screening of listeners was applied to remove outlier results, resulting in a total of 17 listeners. The average scores for each file over all listeners' results are

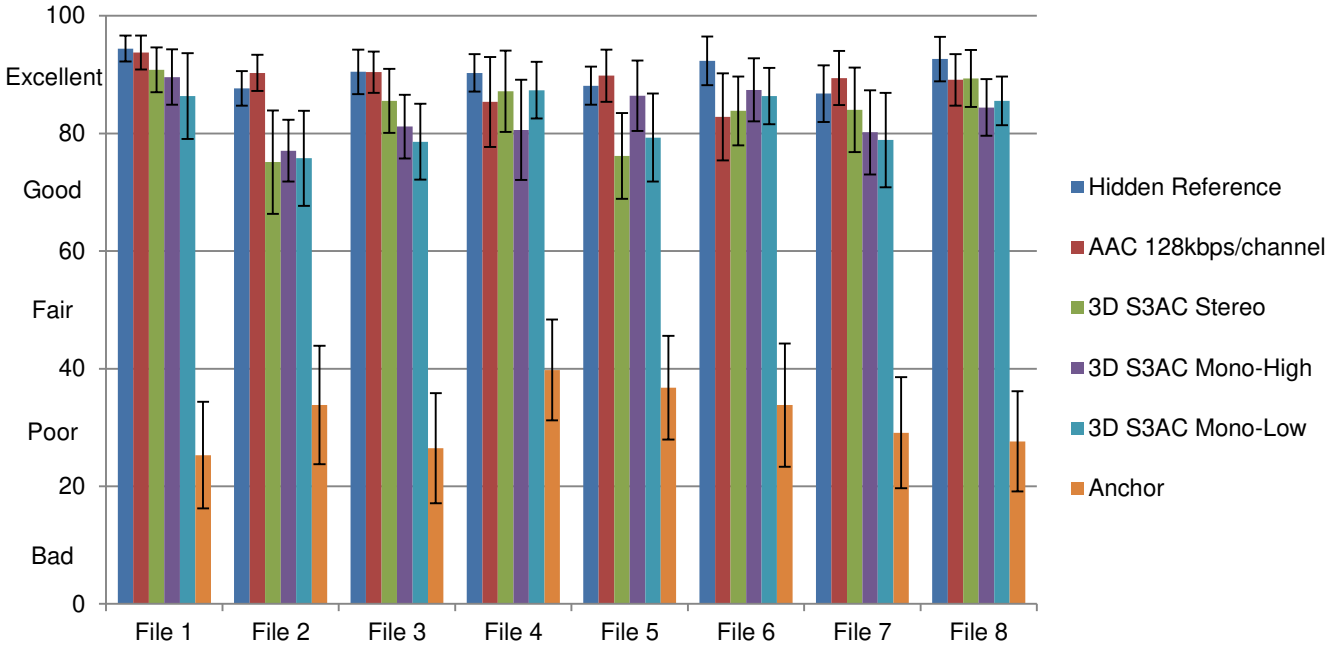

Fig.14. S^{3}AC 3D Listening Test Results. Error bars indicate 95% confidence intervals.

Table VI. ANOVA Test Results. $F_{crit}(4,80) = 2.49$, $\alpha = 0.05$.

File	DF	F	p-value	File	DF	F	p-value
1	4	2.00	0.10	5	4	3.84	0.01
2	4	5.31	0.00	6	4	1.70	0.16
3	4	4.33	0.00	7	4	1.76	0.15
4	4	1.16	0.34	8	4	2.14	0.08

shown in Fig.14, with error bars indicated by 95% confidence intervals.

For all scenarios, the proposed S^{3}AC 3D compression algorithm achieves grades above the MUSHRA 'Good' grade, while most of the results of the proposed conditions lie on the boundary between the MUSHRA 'Excellent' grade and 'Good' grade. Considering the bandwidth reduction from 16-channel to 2-channel (stereo downmix) or less (mono downmix + 3D side information), the advantage of the proposed S^{3}AC 3D multi-channel audio compression technique can clearly be seen. It can be observed that in all cases, the results for the AAC condition appear statistically similar to the hidden reference as judged by the overlapping confidence intervals. This result is expected, since for this condition, each of the 16 channels is compressed at 128 kbps using AAC. For the anchor signals, while Files 1, 3, 7 and 8 achieve scores between 20 and 30, the anchor signals for files 2, 4 and 5 achieve average scores close to 40. This is higher than typically reported for MUSHRA tests of compressed multichannel spatial audio [13]. This could be explained by the

nature of the 3D audio contained in these files, which includes ambient noise and moving sources rather than localized point sources. Hence, the anchor signal for these files, which contains low pass and non-localized content, is more similar to the reference signals when compared to other test files. Further research is required to investigate this in more detail.

An ANOVA test was conducted to analyze the results for each file in more detail. Excluding the anchor from this analysis resulted in the results presented in Table VI for an $F_{crit}(4,80)=2.49$ and at the significance level of $\alpha=0.05$. The results in Fig. 14 and Table VI show that, for Files 1, 4, 6, 7 and 8, the three proposed 3D compression modes resulted in little or no perceptual difference compared to the original reference and AAC conditions, with p-values exceeding the 0.05 significance level. It can also be observed that despite higher average azimuth errors objectively measured for File 1 (see Tables II to IV), this did not result in a significant reduction in subjective quality and hence provides some validation for the SLQP quantization approach for localized speech sources.

In contrast, results for Files 2, 3 and 8 indicate some statistical differences between the reference, AAC coded and the three proposed 3D compression modes, with p-values less than the 0.05 significance level. File 2 resulted in the lowest average MUSHRA scores of all files with the three proposed 3D compression modes all resulting in statistically lower results than the reference and AAC conditions. This is despite achieving the lowest azimuth and elevation errors as shown in Tables II to IV. It is proposed that this is due to the nature of the sound scene as further discussed below. File 3 achieves the lowest average MUSHRA result for the low precision 3D mono coding mode, which can be explained by the high azimuth error as indicated in Table III. File 5 achieves the lowest average MUSHRA for the 3D stereo coding mode although overlapping confidence intervals indicate that this result is statistically similar to the other two 3D compression modes. As File 5 also contains moving sources, the lower perceptual quality for this file could be explained by similar reasons to those used for File 2.

Informal listening and feedback from subjects indicated that the most noticeable distortions were due to the loss in localization accuracy or distortion of the localized sources. For example, for File 2, which contains a moving source flying overhead, there are perceptible ‘jumps’ in the location of the source. This can be observed in the objective results of Figs. 11

to 13, which show the discontinuities in the sound source location, which is particularly noticeable in Fig. 12 showing results for the low precision SLQP quantization. One possible approach to minimizing this distortion would be to employ a technique to smoothly change the source locations (although this was not explored further in this work). Distortion of the sources was most noticeable when there were mixtures of localized sources and diffuse noise. For example, for File 3, containing both localized speech and diffuse noise, distortions in the form of loss of spectral content (perceived as musical distortion) was apparent. This can occur when two or more sources are overlapping in a given time frequency bin, where the orthogonal analysis of Section II.B will combine these sources into a single virtual source. If one source is speech (as in File 3), any portion of this spectrum that overlaps with other sources may be distorted during reproduction, since the exact energy and location of these frequency components is not recoverable. A solution to these problems could be to transmit additional side information about the relationship between sources and their locations (e.g. as used in [38]), although this was not explored further in this work. It should also be noted that a detailed evaluation of the effect of the AAC compression of the stereo downmix on the localization error of the virtual sources was not conducted. However, the bit rate of 64 kbps/channel was chosen to minimize the influence of the AAC coder on such errors and results show that the stereo-downmix mode performed similar in perceptual tests compared to the mono-dowmix mode, where localization information was coded separately rather than derived from the compressed downmix.

V. CONCLUSION

A generalized compression approach to the efficient compression of multi-channel 3D spatial audio is presented. A 3D orthogonal analysis algorithm is proposed for efficient estimation of 3D sound source azimuth/elevation localization information from the loudspeaker signals of an arbitrary 3D reproduction setup. A complete encoding/decoding system is designed based on this algorithm, where the derived 3D source azimuth/elevation localization information is spatially quantized by the proposed SLQP approach. The S^3AC spatial squeezing approach is extended so as to perform a unique mapping between the 3D SLQP and the 60° azimuthal region represented by a stereo downmix. A ‘mono downmix + 3D

side information' approach is also proposed, where the 3D SLQP can be further quantized with different precisions defined by the user. Different 3D reproduction methods for decoding are discussed, where higher order Ambisonics reproduction can be used in symmetrical loudspeaker arrays and a 3D amplitude panning method based on inverse orthogonal analysis can be used in asymmetrical loudspeaker arrays for better source localization.

The objective results of the proposed compression algorithms, including stereo downmixing, mono downmixing with high SLQP precision and mono downmixing with low SLQP precision, are also presented for a series of multichannel 3D audio files containing localized sources. Objective evaluations show that the 3D azimuth/elevation localization information derived using the proposed orthogonal source estimation algorithm can be efficiently quantized without introducing significant quantization distortion. This 3D multi-channel audio compression technique, including the three different modes, is further evaluated by subjective experiments focusing on the perceived localization quality. The results indicate that, while the bandwidth requirement is significantly reduced from 16-channel to 2-channel (or less), the degradation in perceptual quality remains minimal after decoding.

REFERENCES

[1] B. Bosi and R. E. Goldberg,, *Introduction to digital audio coding and standards*. Springer, 2003.

[2] International Organization for Standardization, "Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s — Part 3: Audio," Mar-1999. [Online]. Available: http://dret.net/biblio/reference/iso11172-3. [Accessed: 30-Dec-2011].

[3] ISO/IEC 13818-7:2006, "Information technology -- Generic coding of moving pictures and associated audio information -- Part 7: Advanced Audio Coding (AAC)," 1997.

[4] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, A. Kenzo,, H. Fuchs,, and M. Dietz,, "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc*, vol. 45, no. 10, pp. 789–814, 1997.

[5] ITU-R BS.775-2, "Multichannel stereophonic sound system with and without accompanying picture," Jan. 2006.

[6] "Surround Sound Past, Present and Future, Technical Report," *Dolby Laboratories, Online:http://www.dolby.com/tech*, 1999.

[7] M. Bosi, "High Quality Multichannel Audio Coding: Trends and Challenges," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, 1999.

[8] I. Elfitri, B. Günel, and A. M. Kondoz,, "Multichannel Audio Coding Based on Analysis by Synthesis," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 657–670, Apr. 2011.

[9] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric Coding of Stereo Audio," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1305–1322, 2005.

[10] F. Baumgarte and C. Faller, "Binaural cue coding-Part I: psychoacoustic fundamentals and design principles," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 509– 519, Nov. 2003.

[11] C. Faller and F. Baumgarte, "Binaural cue coding-Part II: Schemes and applications," *IEEE Transactions on Speech and*

Audio Processing, vol. 11, no. 6, pp. 520– 531, Nov. 2003.

[12] L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen, and K. Kjörling, "MPEG Surround: The Forthcoming ISO Standard for Spatial Audio Coding," in *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*, 2006.

[13] J. Breebaart, S. Disch, C. Faller, J. Herre, G. Hotho, K. Kjörling, F. Myburg, M. Neusinger, W. Oomen, H. Purnhagen, and J. Rödén, "MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status," in *Audio Engineering Society Convention 119*, 2005.

[14] M. Goodwin and J. Jot, "Spatial Audio Scene Coding," in *Audio Engineering Society Convention 125*, 2008.

[15] J. Jot, A. Krishnaswami, J. Laroche, J. Merimaa, and M. Goodwin, "Spatial Audio Scene Coding in a Universal Two-Channel 3-D Stereo Format," in *Audio Engineering Society Convention 123*, 2007.

[16] V. Pulkki, "Directional Audio Coding in Spatial Sound Reproduction and Stereo Upmixing," in *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*, 2006.

[17] C. Faller and V. Pulkki, "Directional Audio Coding: Filterbank and STFT-based Design," in *Audio Engineering Society Convention 120*, 2006.

[18] B. Cheng, C. Ritz, and I. Burnett, "Principles and Analysis of the Squeezing Approach to Low Bit Rate Spatial Audio Coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, 2007, vol. 1, pp. I–13–I–16.

[19] M. Gerzon, "Ambisonics part two: Studio techniques," *Studio Sound*, no. 17, pp. 24–30, 1975.

[20] D. Jerome, M. Sebastien, and N. Rozenn, "Further Investigation of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging," in *114 AES Convention*, 2003.

[21] Dolby Laboratories, "Dolby ATMOS Cinema Technical Guidelines," *available at http://www.dolby.com/uploadedFiles/Assets/US/Doc/Professional/Dolby-Atmos-Cinema-Technical-Guidelines.pdf*, 2012.

[22] G. Potard, "3D-audio object oriented coding," *University of Wollongong Thesis Collection*, Jan. 2006.

[23] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[24] B. Cheng, C. Ritz, and I. Burnett, "A Spatial Squeezing approach to Ambisonic audio compression," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, 2008, pp. 369–372.

[25] B. Cheng, C. H. Ritz, and I. S. Burnett, "Psychoacoustic-based quantisation of spatial audio cues," *Electronics Letters*, vol. 44, no. 18, pp. 1098–1099, Aug. 2008.

[26] B. Cheng, C. Ritz, and I. Burnett, "Spatial audio coding by squeezing: analysis and application to compressing multiple soundfields," presented at the The 17th European Signal Processing Conference, 2009, pp. 909–913.

[27] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[28] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc*, vol. 53, no. 12, pp. 1115–1127, 2005.

[29] M. Laitinen and V. Pulkki, "Converting 5.1 audio recordings to B-format for directional audio coding reproduction," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 61–64.

[30] M. Laitinen, "Reproducing Applause-Type Signals with Directional Audio Coding," *Journal of the Audio Engineering Society*, vol. 59, no. 1, p. 29, 2011.

[31] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP J. Adv. Signal Process*, vol. 2010, pp. 2:1–2:13, Feb. 2010.

[32] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT Press, 1997.

[33] R. Rice, *Some practical universal noiseless coding techniques*. National Aeronautics and Space Administration, Jet Propulsion Laboratory, California Institute of Technology, 1991.

[34] H. Walmsey, "Spanish Flea — Ambisonic Surround Sound. Ambisonics, 5.1, audio recordings." [Online]. Available: http://www.ambisonia.com/Members/Henry/ambisonicfile.2006-07-20.0004393664/. [Accessed: 20-Jul-2012].

[35] "Ambisonic Surround Sound. Ambisonics, 5.1, audio recordings — Ambisonic Surround Sound. Ambisonics, 5.1, audio recordings." [Online]. Available: http://www.ambisonia.com/. [Accessed: 16-Aug-2012].

[36] H. Walmsey, "Tijuana Taxi — Ambisonic Surround Sound. Ambisonics, 5.1, audio recordings." [Online]. Available: http://www.ambisonia.com/Members/Henry/ambisonicfile.2006-07-26.7264303593/. [Accessed: 20-Jul-2012].

[37] B. 1534 International Telecommunication Union, "Methods for the subjective assessment of intermediate quality levels of coding systems." 1997.

[38] B. Cheng, C. Ritz, and I. Burnett, "Encoding Independent Sources in Spatially Squeezed Surround Audio Coding," in *Advances in Multimedia Information Processing – PCM 2007*, vol. 4810, H. Ip, O. Au, H. Leung, M.-T. Sun, W.-Y. Ma, and S.-M. Hu, Eds. Springer Berlin / Heidelberg, 2007, pp. 804–813.

Bin Cheng received his B.E. degree in Electrical Engineering from Beijing University of Aeronautics and Astronautics, China, in 2003. He received his MEng and PhD both from the University of Wollongong, Australia, in 2004 and 2011 respectively.

His research interests include spatial audio processing, binaural signal processing, and multimedia content analysis. He is currently working as a Senior Research Engineer at Dolby Laboratories (Beijing).

Christian Ritz (M'97, SM'08) received his B.E. degree in Electrical Engineering and B. Math degree both from the University of Wollongong, Wollongong, Australia, in 1998. He received his PhD degree in 2003 from the University of Wollongong, Wollongong, Australia.

He joined the University of Wollongong in 2003 and is currently a Senior Lecturer there. His current research interests include spatial audio signal processing, multichannel speech signal processing and multimedia signal processing.

Ian S Burnett (M'87–SM'02) is Professor and Head of School at RMIT University, Australia. His current research interests are in multimedia processing and delivery, speech and audio signal processing, 3D spatial audio, and 3D models from images. He received his BSc, MEng, and PhD in electrical and electronic engineering from the University of Bath, UK.

He is currently a participant in the Smart Services CRC working on projects in the multimedia space with various industry partners He was an active participant in MPEG and MPEG-21 until 2008, notably as Australian Head of Delegation and the Chair of the Multimedia Description Schemes sub-group. He was also the CTO of enikospty ltd, a company specializing in MPEG-21 based applications.

Xiguang Zheng received the B.E. degree in telecommunications engineering in 2009 from both Beijing University of Posts and Telecommunications, China and Queen Mary University of London, United Kingdom. He is currently working towards the Ph.D. degree in spatial speech and audio signal processing at University of Wollongong, Wollongong, Australia.

From September 2008 to June 2009, he was a Research Assistant at the Beijing University of Posts and Telecommunications. His research interests include multichannel speech and audio signal processing and joint source-channel coding.