SPATIAL AUDIO CODING BY SQUEEZING: ANALYSIS AND APPLICATION TO COMPRESSING MULTIPLE SOUNDFIELDS

Bin Cheng[1], Christian Ritz[1], and Ian Burnett[2]

[1]Whisper Laboratories
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong, NSW, Australia, 2500
{bc362, critz}@uow.edu.au

[2]School of Electrical and Computer Engineering
Royal Melbourne Institute of Technology, Melbourne, VIC, Australia, 3000
ian.burnett@rmit.edu.au

ABSTRACT

Spatially Squeezed Surround Audio Coding (S^3AC) proposed by the authors provides efficient compression of multi-channel surround audio. Compression is achieved by exploiting human sound localisation blur to save the surround soundfield information in a squeezed stereo soundfield. In this paper, the localisation loss during the S^3AC analysis/synthesis is evaluated and the minimum size of the S^3AC squeezed soundfield is derived in a frequency dependent form. Results from perceptual listening tests show that, compared with standard squeezing from a 360° surround soundfield to a 60° stereo soundfield, a more intensive squeezing method, such as from 360°-to-5°, does not introduce audible localisation distortion. This leads to a further application of S^3AC for compressing more than one surround soundfield into a single stereo downmix for applications such as spatialised teleconferencing. This application is also described and perceptually evaluate.

1. INTRODUCTION

Efficient representation of surround sound signals has been an area with significant research interests for decades. Recent development in this area has focused on the parametric approach motivated by binaural sound perception, which derives side information to replicate the arithmetical relationships between multiple audio channels, including inter-channel time/level difference and coherence [1]. In these approaches, a backward compatible stereo/mono downmix is created by summing multiple channels representing a surround soundfield. This downmix is accompanied by the derived parametric side information. Alternative parametric approaches derive side information representing source direction and diffuseness [2]. While these approaches have the advantages of bit-rate efficiency and backward compatibility, the extra side information increases data transmission and storage costs.

Spatially Squeezed Surround Audio Coding (S^3AC) was introduced by the authors as an alternative approach to compressing multi-channel spatial audio. [3] This approach is based on deriving sound sources and their location within the soundfield. Rather than requiring side information, S^3AC utilizes its stereo downmix to save the localisation information of the surround soundfield. In S^3AC, a spatial squeezing algorithm compresses a 360° surround soundfield into a smaller stereo soundfield, as illustrated in Fig. 1. By exploiting the psychoacoustic theory of localisation blur [4], which states that the perceptual localisation resolution of human auditory system is limited [4], this squeezing process ensures that all the perceptual localisation information in the original soundfield is recorded and fully recoverable. Furthermore, the stereo downmix signal can be further compressed with conventional stereo audio coders, similar to existing spatial coding coders.

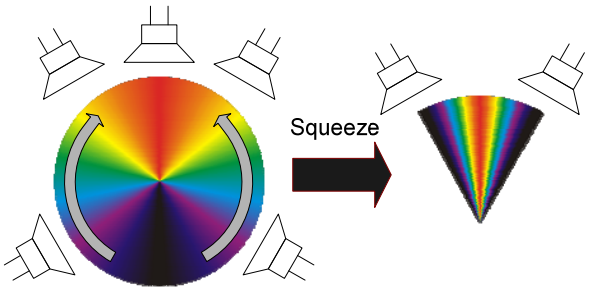

Fig. 1 S^3AC spatial soundfield squeezing, colour regions in the original soundfield are squeezed to the same but smaller colour regions in the downmix soundfield

The coder has been successfully applied to typical multichannel loudspeaker signals [3] (such as ITU 5.1 [5]) as well as B-format soundfield recordings [6]. In addition, an application of S^3AC to spatial teleconferencing was proposed recently [7]. In this application, microphone recordings of one or more participants at one or more sites are squeezed into a single stereo channel. Spatialised rendering of the conference (e.g. over a set of loudspeakers) can be achieved by analysing the received downmix and re-panning to ensure each speaker is spatially separated.

This paper analyses S^3AC to identify the impact of squeezing on the localisation error of decoded sound sources and further investigates the application to squeezing multiple soundfields into a single stereo downmix. In Section 2, an overview of S^3AC is presented, followed by the theoretical analysis of the localisation resolution degradation caused by S^3AC compression and the minimum size of the S^3AC squeezed soundfield for maintaining no perceptual localisation loss. Section 3 presents the subjective results for evaluating different types S^3AC squeezing processes and the resulting perceptual localisation impact. Section 4 presents a new approach to compressing multiple spatial sound scenes with speech content while Section 6 draws the conclusions.

2. THEORETICAL ANALYSIS

2.1 Overview of the S^{3}AC coder

A standard S^{3}AC encoding/decoding procedure is illustrated in Fig. 2. For a multi-channel audio signal containing the source localisation information of a surround soundfield, S^{3}AC starts the encoding by a time-frequency transformation, which can be implemented by either a short-time Fourier transform (STFT) or perceptual filterbank decomposition. This is followed by a frequency domain source estimation process, where both the auditory and localisation information is extracted as a virtual source for each frequency or perceptual frequency band. Typically, for an ITU 5.1-channel [5] audio signal, a channel pair with the highest signal energy is analyzed to obtain a localised dominating virtual sound source for a particular frequency *k*. This can be efficiently achieved by applying inverse tangent pan-pot law, defined by:

$$\theta(k) = \arctan\left(\frac{L(k)-R(k)}{L(k)+R(k)} \cdot \tan\varphi(k)\right) \quad (1)$$

where $L(k)$ and $R(k)$ are the amplitude of the two selected channels with angular separation of $2\times\varphi(k)$, $\theta(k)$ is the derived source azimuth. The total energy of this virtual source is thus defined by:

$$S(k) = \sqrt{L^2(k)+R^2(k)} \quad (2)$$

This processed is followed by the S^{3}AC azimuth squeezing, as illustrated in Fig. 1, which assigns a new azimuth for the virtual source in a squeezed soundfield rendered by a stereo downmix signal. This can be modelled by:

$$\theta_s(k) = f\left(\theta(k)\right) \quad (3)$$

where function f is a linear mapping between the original and squeezed soundfield and $\theta_s(k)$ is the source azimuth in the squeezed soundfield with a size of $2\times\varepsilon_d$ in degrees. A tangent law re-panning is performed to generate the stereo downmix signal by:

$$L_S(k) = \frac{S(k)\cdot\left(\tan\varepsilon_d + \tan\theta_s(k)\right)}{\sqrt{2\tan^2\varepsilon_d + 2\tan^2\theta_s(k)}}$$
$$R_S(k) = \frac{S(k)\cdot\left(\tan\varepsilon_d - \tan\theta_s(k)\right)}{\sqrt{2\tan^2\varepsilon_d + 2\tan^2\theta_s(k)}} \quad (4)$$

where $L_S(k)$ and $R_S(k)$ are frequency components of the left and right downmix channel. This stereo signal can be either quantised by conventional perceptual audio coding techniques, as used in MP3/AAC [8] or converted to the time-domain for transmission. This effectively results in using the same bandwidth as the conventional stereo system to transmit a multichannel surround soundfield without any accompanying side information.

At the decoder, inverse tangent panning is applied again in the frequency domain to derive the squeezed azimuth of the virtual source:

$$\theta_d(k) = \arctan\left(\frac{L_s(k)-R_s(k)}{L_s(k)+R_s(k)} \cdot \tan\varepsilon_d\right) \quad (5)$$

so that the original azimuth of this virtual source can be recovered by:

$$\theta'(k) = f^{-1}\left(\theta_d(k)\right) \quad (6)$$

The spectral information of the virtual source can be derived from the two downmix channels by:

$$S'(k) = \sqrt{{L_S}^2(k)+{R_S}^2(k)} \quad (7)$$

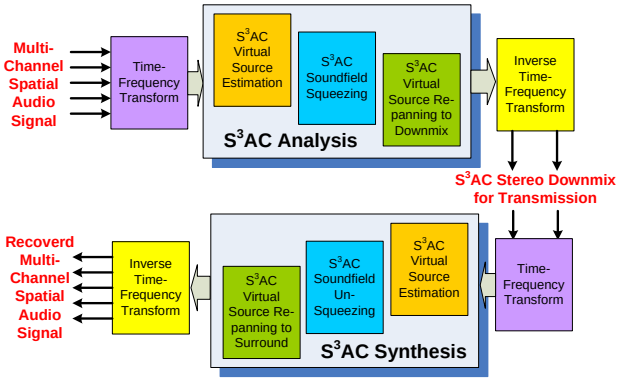

Fig. 2 S^{3}AC Encoding/Decoding Process

and it is amplitude-panned to the relative loudspeakers in the multi-channel playback system according to the loudspeaker setup and $\theta'(k)$. After a frequency-time transform, a decoded surround soundfield is recovered replicating the original surround sound scene.

2.2 Frequency dependent localisation resolution

In practice, limited signal quantisation precision is applied during the S^{3}AC encoding/decoding process, which has an impact on the localisation resolution capability of the S^{3}AC encoded signal. In other words, the number of azimuths that the downmix signal can be used to manipulate is limited. Assuming that a 16-bit integer quantisation is applied for the transmission of the encoded signal, i.e. the stereo downmix has an integer value in [0 2^{16}-1], Eq. (5) is re-written as

$$\theta'_d(k) = \arctan\left(\frac{\|L_s(k)\|-\|R_s(k)\|}{\|L_s(k)\|+\|R_s(k)\|} \cdot \tan\varepsilon_d\right) \quad (8)$$

where $\|X\|$ stands for rounding to the nearest integer. While the quantised values of $\|L_s(k)\|$ and $\|R_s(k)\|$ are also restricted by the overall energy $S_s(k)$, such that:

$$0 \le \|L_s(k)\| \le \|S(k)\|$$
$$\|R_s(k)\| = \left\|\sqrt{S^2(k)-{L_S}^2(k)}\right\| \quad (9)$$

This results in having a localisation resolution capability of $\|S(k)\|+1$ for each frequency, since each pair of $\|L_s(k)\|$ and $\|R_s(k)\|$ defined by Eq. (9) can be effectively utilized at the decoder to derive one virtual source azimuth. This further suggests that, for each frequency, the number of azimuths that can be stored in an S^{3}AC downmix is quantified by the signal amplitude.

In addition to the localisation resolution being limited by the virtual source energy, localisation loss is also introduced by the spatial squeezing step in the S^{3}AC analysis, which is modelled by Eq. (3). Assuming that, for frequency *k*, an S^{3}AC virtual source is derived from a channel pair with angular separation of $2\times\varphi(k)$, as defined in Eq. (1), and squeezed into the downmix soundfield with a size of $2\times\varepsilon_d$, the spatial squeezing process is effectively equivalent to limiting the range of $L_S(k)$ and $R_S(k)$. The resulting localisation precision in the S^{3}AC squeezed soundfield can be modelled by a proportional function between the size of the

channel pair $2\times\varphi(k)$, from which the source is derived, and the size of the squeezed soundfield $2\times\varepsilon_d$, which is:

$$\rho_S(k)=\left(\left\|S(k)\right\|+1\right)\cdot\frac{\varepsilon_d}{\varphi(k)} \quad (10)$$

Recalling the localisation blur theory [4] that indicates a perceptual localisation resolution limit of approximately 1°, the perceptual redundancy of the S^{3}AC downmix signal for transmitting an original soundfield with a size of ψ degrees is:

$$r(k)=\rho_S(k)-\psi=\left(\left\|S(k)\right\|+1\right)\cdot\frac{\varepsilon_d}{\varphi(k)}-\psi \quad (11)$$

2.3 Localisation resolution of the low amplitude components and perceptual relevancy

In order to maintain perceptual lossless in localisation, a non-negative redundancy value is required in Eq. (11). For instance, in a given set of geometrical parameters in the S^{3}AC analysis, i.e. defining the size of the original and squeezed soundfields as ψ and ε_d, respectively and the loudspeaker layout as $\varphi(k)$, the minimum spectral amplitude required to maintain non-negative localisation redundancy can be derived by setting Eq. (11) to zero:

$$\left\|S(k)\right\|^{\min}=\frac{\psi\cdot\varphi(k)}{\varepsilon_d}-1 \quad (12)$$

In the typical S^{3}AC coding of ITU 5.1-channel signal, where a 360° surround soundfield is squeezed into a 60° downmix soundfield, a condition is given as:

$$\psi=360^o,\ \varphi(k)\in\{15,30,40,55,70\}^\circ,\ \varepsilon_d=60^o \quad (13)$$

Note that the value of $\varphi(k)$ is defined according to the ITU recommended 5.1 loudspeaker setup [5]. By substituting the maximum value of $\varphi(k)=70^o$ into Eq. (12), it gives a minimum spectral amplitude of $S(k)=419$ to ensure no localisation loss for all cases. While this indicates that, for some low amplitude components, the localisation redundancy can be negative resulting in insufficient localisation resolution, these low amplitude components tend to be below the threshold of hearing or masked simultaneously or temporally by other nearby spectral components. Defining the lowest point of the absolute hearing threshold as 0dB [9] with amplitude equivalent to the minimum integer value of 1 in the 16-bit format, Fig. 3 shows a case where the minimum masking curve of a 65 dB white noise signal ($1/10$ of the maximum energy of a 16-bit audio signal) is at least 15 dB above the level of $S(k)=419$, which is 47dB.

In addition, the derivation here further indicates that,

(1) While the spatial resolution is given in proportion to the amplitude of the source, louder sources with higher perceptual importance inherently attract higher localisation resolution.

(2) When there is no priori knowledge of the pre-amplifying and post-amplifying in the actual playback system, quantifying a minimum size of the squeezed soundfield is only an estimation of the ideal condition. For a given target size of the squeezed soundfield, a post-amplifying process can be applied on the encoded signal to ensure no perceptual localisation loss.

(3) As can be found in Eq. (10), increasing the size of the squeezed soundfield can also improve the localisation resolution if pre-amplifying and post-amplifying of the signal is not desired.

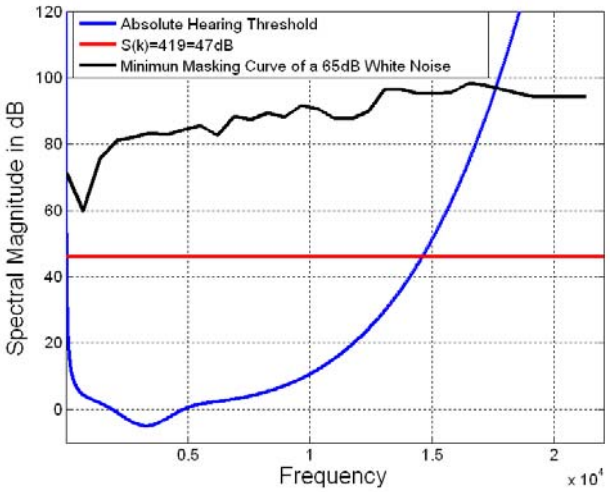

Fig. 3 Absolute hearing threshold, minimum spectral amplitude *S(k)=419* for no localisation loss, and the minimum masking curve of a 65dB white noise.

2.4 S^{3}AC Squeezing Limitation

As discussed in the last section, the size of the S^{3}AC downmix soundfield has an impact on the localisation resolution. Earlier experiments have shown that the standard S^{3}AC 360°-to-60° analysis-synthesis process results in no perceived localisation distortion [3, 6]. Here, it is suggested that the S^{3}AC squeezed can be performed in a more intensive way, i.e. squeeze the surround soundfield into a smaller soundfield than 60°. By setting the redundancy formula of Eq. (11) to zero, the smallest size of the squeezed soundfield in degrees, without causing localisation loss, can be derived as a function of the spectral energy:

$$\varepsilon^{\min}(k)=\frac{\psi\cdot\varphi(k)}{\left\|S(k)\right\|+1} \quad (14)$$

In this equation, while the numerator is defined by the size of the soundfield and loudspeaker layout, the spectral energy component, $S(k)$, is signal dependent and varies over time-frequency. In order to further quantify the minimum size of the S^{3}AC downmix soundfield, subjective analysis is required. In the following section of this paper, a perceptual experiment is presented, where the impact of decreasing sizes of the S^{3}AC squeezed soundfield on the localisation distortion are perceptually evaluated.

3. EXPERIMENTAL EVALUATION

This section presents the subjective evaluation on the perceptual impact of using different sizes for the downmix soundfield in the S^{3}AC spatial squeezed process. The aim is to find the smallest size of the S^{3}AC squeezed soundfield that does not introduce perceptual localisation loss. In the experiment, standard ITU 5.1-channel files with 16-bit quantisation precision and 44.1kHz sampling rate were used. Five files with different types of immersive surround sound scenes, including dynamic sound scenes of an aeroplane, car siren, mosquito and localised speech recording, were coded with S^{3}AC. Various S^{3}AC squeezing approaches were applied to generate several coding conditions for evaluation, including S^{3}AC squeezing from 360° to 60°, 40°, 20°, 10°, 5° and 1°, respectively, i.e. the size of the squeezed soundfield ($2\times\varepsilon_d$ in degree as described in Section 2) varies between these given values to generate different S^{3}AC downmix signal and decoded respectively. This can be

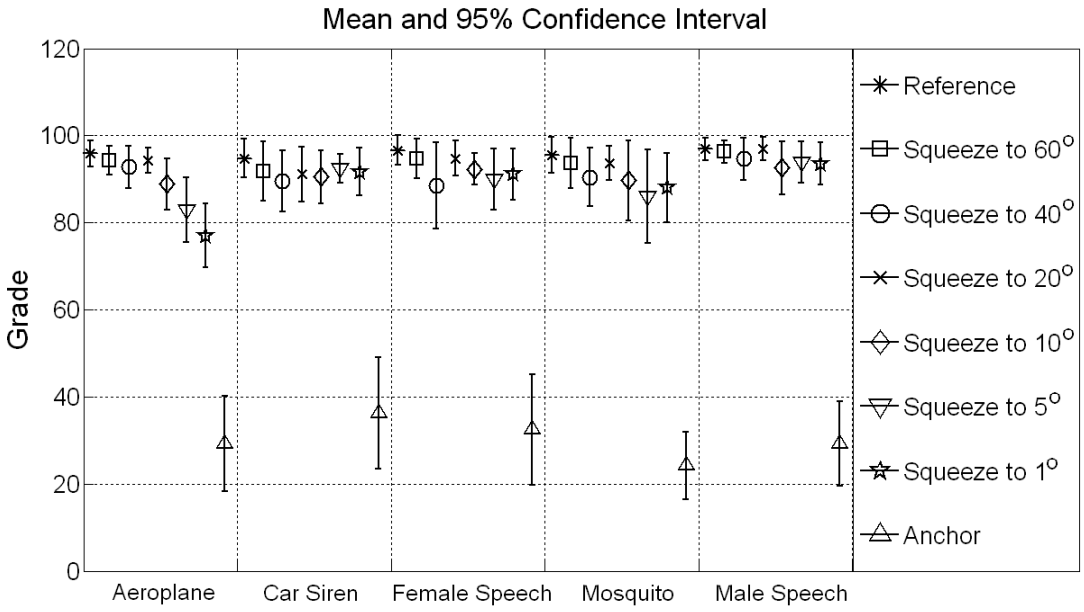

Fig. 4 Listening test results comparing different size of the S^{3}AC downmix soundfield

achieved by modifying Eq. (3) and Eq. (6) during the S^{3}AC encoding/decoding process. All the S^{3}AC encoded files were decoded to the standard ITU 5.1-channel format and played back on a 5.1-multichannel system using GENELEC loudspeakers. The volume of the playback systems was adjusted such that a relative -6dB white noise (by defining the maximum energy in 16-bit format as 0 dB) was played back in an absolute sound pressure of 90dB in every loudspeaker. The experiment was based on the MUSHRA methodology [10], where the six coding conditions described above were randomly mixed with a hidden reference and a 3.5kHz low-pass-filtered anchor, which is also un-localised by equally mixing a mono signal to each channel. Ten listeners took part in the experiment including both experienced and in-experienced listeners, while they were instructed to compare both the perceptual quality and sound source localisation precision between the reference and candidate signals.

The average results with 95% confidence intervals are illustrated in Fig. 4. It is shown that, for all the test files, different S^{3}AC spatial squeezing types from 360° to 60°, 40°, 20° and 10° do not introduce any distortion from a statistical point of view, when compared to the un-coded reference. More intensive squeezing types from 360° to 5° and 1° only cause degradation of less than 10 MUSHRA marks for the aeroplane sound scene, while no statistical difference is found for other test files. The results indicate that while the standard 360°-to-60° S^{3}AC soundfield squeezing does not introduce perceivable distortion, a more intensive squeezing approach such as 360°-to-10° also maintains perceptual and localisation equivalence between the original and coded signals. This is further investigated in the next section.

4. SPATIAL SQUEEZING OF MORE THAN ONE SOUNDFIELD

Since the S^{3}AC squeezing can be performed in a more intensive way than the standard 360°-surround to 60°-stereo squeezing, it is suggested that a stereo downmix can be utilized to save the localisation information of more than one surround soundfield. For instance, Fig. 5 illustrates an approach that the localisation information of two distinct surround soundfields is saved in one

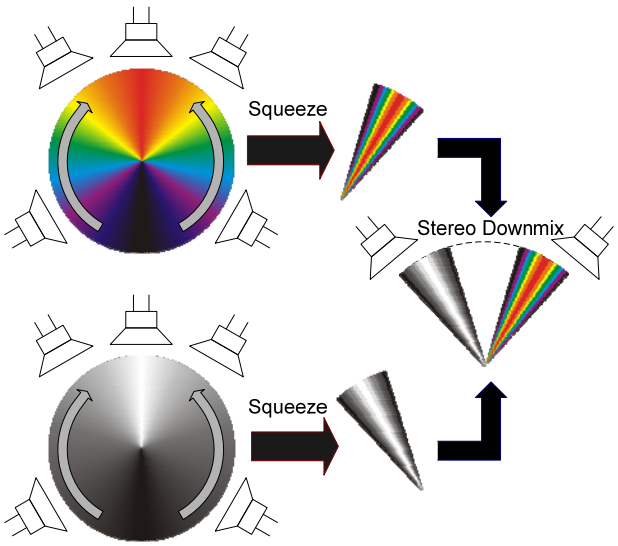

Fig. 5 Squeezing two soundfields into one downmix

single stereo downmix by exploiting the S^{3}AC 360°-to-10° spatial squeezing approach. By utilizing the standard S^{3}AC virtual source estimation procedure, virtual sound source and localisation information can be derived for each original soundfield. The S^{3}AC squeezing step defined by Eq. (3) is modified respectively for the two original soundfields such that the squeezing illustrated in Fig. 5 is achieved, i.e. each original soundfield is squeezed into 10° and the two squeezed soundfields are saved in the [30°, 20°] and [-20°, -30°] regions in the squeezed soundfield, with a 40° empty region for discriminating purpose. During decoding, the two squeezed soundfields with a size of 10° are derived from the stereo downmix and recovered to two 360° surround soundfields by separately applying the inverse-squeezing approach.

For this approach, it is discovered that sound sources with overlapping time-frequency components, described in [11], can cause distortion in localisation and perceptual quality. Compared

with the standard S^{3}AC approach to compressing one spatial audio file, more overlapping time-frequency components can be introduced by the inter-soundfield interference, while two distinct files are compressed into one S^{3}AC downmix. However, this scheme is proposed as an efficient approach for compressing multiple surround speech scenes, especially for teleconferencing application, e.g. teleconference application for multiple distributed sites described in [7]. In a multi-site teleconference scenario, the distortion caused by overlapping time-frequency components is less significant as it only occurs when multiple participants speak concurrently.

This proposed S^{3}AC two-soundfield-to-one-downmix approach was implemented on two multi-channel speech recordings, which are the male and female speech files used in the listening test presented in the last section. The files coded with this scheme were inserted into the same MUSHRA test as one of the evaluation conditions. The average MUSHRA marks and 95% confidence intervals for these two coded speech files compared with the hidden reference and anchor are shown in Fig. 6. It is shown that the MUSHRA marks for the two coded condition are in the region between 60 and 90, which refers to good and fair quality according to MUSHRA recommendation [10]. In the experiment, listeners claimed that, compared with the reference, there is occasional audible distortion in these two coded speech files. This is caused by the overlapping time-frequency components described above. However, listeners also claimed that the quality of these two speech files is suitable for teleconference application; in particular, high localisation accuracy is preserved when compared with the reference file.

In addition, since the stereo downmix containing information of two soundfields can be further compressed by the existing MP3/AAC coders, this results in an approach that uses the same bandwidth as a conventional stereo MP3/AAC bit-stream for transmitting two multi-channel surround sound scenes. In the example application described in [10] with three participating sites, a bandwidth of only 128kbps is required for transmitting localised speech content of the two remote conference sites to each participating site. At each site, speech scenes can be flexibly spatialised to disambiguate each speaker in the teleconference.

5. CONCLUSIONS AND FURTHER WORK

Based on psychoacoustic principles, analysis has been performed on the S^{3}AC spatial audio compression technique to evaluate the localisation resolution and the limitation of the spatial squeezing approach. The derivation has shown that, the sound source localisation resolution is dependent on the spectral energy of each frequency domain virtual source. It is also shown that, the minimum size of the S^{3}AC squeezed soundfield is frequency dependent. Subjective evaluation has been performed to further investigate the theoretical analysis. Although the low spectral energy components have insufficient localisation resolution according to the derivation, it is not discovered in the listening test. This concludes that, while localisation resolution in the S^{3}AC analysis/synthesis is dependent on the frequency domain virtual source amplitude, audible sound sources with perceptual significance have sufficient localisation resolution. Furthermore, it is shown by the perceptual evaluation that, in addition to the standard S^{3}AC 360°-to-60° spatial squeezing, other squeezing approaches, such as 360°-to-10°, does not introduce perceptual localisation distortion.

Based on these reults, an approach to compressing multiple teleconferencing surround sound scenes has been proposed. By utilizing this approach, significant bit-rate efficiency is achieved as only the bandwidth for transmitting a stereo signal is required for transmitting two full surround soundfields. Subjective evaluation shows that good quality and precise localisation is achieved for the recovered surround sound scenes. This approach will be further exploited so that three or more soundfields can be squeezed into one stereo downmix, while spectral and temporal shaping algorithms can be employed to resolve the distortion caused by time-frequency overlapping.

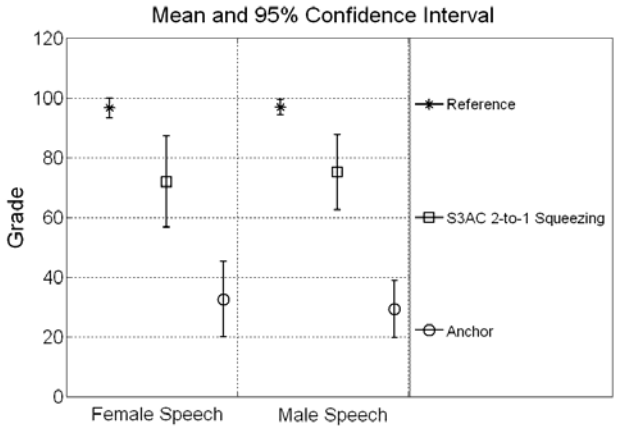

Fig. 6 Listening test results for S^{3}AC two-soundfields-to-one downmix compression of multiple speech spatial scenes

REFERENCES

[1] L. Villemoes, et al., "MPEG Surround: the forthcoming ISO standard for spatial audio coding", in *Proc. AES 28th International Conference*, Pitea, Sweden, June 2006.

[2] V. Pulkki, "Spatial sound reproduction with directional audio coding", *J. Audio Eng. Soc.*, vol.55, no. 6, pp. 503-516, June 2007.

[3] B. Cheng, C. Ritz, and I. Burnett, "Principles and Analysis of the Squeezing Approach to Low Bit Rate Spatial Audio Coding", in *Proc. IEEE International Conf. on Acoustic, Speech and Signal Processing, ICASSP 2007*, Honolulu, USA, Apr. 2007.

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

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

[6] B. Cheng, C. Ritz, and I. Burnett, "A Spatial Squeezing Approach to Ambisonics Audio Compression", in *Proc. IEEE International Conf. on Acoustic, Speech and Signal Processing, ICASSP 2008*, Las Vegas, USA, Mar. 2008.

[7] E. Cheng, B. Cheng, C. Ritz, and I. Burnett, "Spatialised Teleconferencing: Recording and 'Squeezed' Rendering of Multiple Distributed Sites", in *Proc. Australian Telecommunication. Networks and Applications Conf., ATNAC 2008*, Adelaide, Australia, Dec. 2008.

[8] M. Bosi, R.E. Goldberg, "Introduction to digital audio coding and standards", Springer Science+Business Media, New York, USA, 2002.

[9] T. Painter, A. Spanias, "Perceptual Coding of Digital Audio", in Proc. Of the IEEE, v. 88-4, p. 451-515, Apr. 2000.

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

[11] B. Cheng, C. Ritz, and I. Burnett, "Encoding Independent Sources in Spatially Squeezed Surround Audio Coding", in *Proc. Pacific-Rim Conference on Multi-Media PCM 2007*, Hong Kong, China, Dec. 2007.