Binaural Reproduction of Spatially Squeezed Surround Audio

B. Cheng[1], C.H. Ritz[2], I.S. Burnett[3]
[12]*Whisper Laboratories, University of Wollongong, Wollongong, NSW, Australia*
[3]*School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia*
{[1]bc362, [2]critz}@uow.edu.au, [3]ian.burnett@rmit.edu.au

Abstract

Spatially Squeezed Surround Audio Coding (S^3AC) has been previously proposed as an efficient approach to multi-channel spatial audio coding with stereo/mono backward compatibility. This paper presents a binaural reproduction scheme that exploits source localisation information in the S^3AC squeezed soundfield or S^3AC cues to simulate surround audio scene over headphones. The approach utilises interpolated HRTFs to bring the S^3AC advantages of accurate, localised sound sources to stereo headphone systems. The integration of the HRTF approach to reproduction also exploits human localisation ability to reduce interpolation complexity. Subjective experiments demonstrate that accurate localisation is achieved from binaural, interpolated HRTF playback when compared to multi-channel playback.

1. Introduction

Reproduction of surround auditory scenes has been an area of great interest for decades. For both speaker and headphone reproduction systems, the goal is to provide the listener a virtual sound scene with realistic sound source localisation experience. While the transmission and reproduction of surround sound normally requires a number of full bandwidth audio channels (e.g. ITU 5.1-channel signal [1] and Ambisonics [2]), significant research has been performed on either efficient compression of the original multi-channel audio signals for multi-speaker reproduction [3] or transformation of the soundfield into binaural information for localised headphone playback [4]. These two approaches can be linked by the transaural system described in [5] but additional complexity is required.

Recent development of multi-channel audio coding approaches, e.g. MPEG Surround (MPS) [6], has achieved stereo/mono backward compatible compression of spatial audio signals in which the transmission bit-rate is independent of the number of original channels. In such an approach, a stereo/mono downmix is created from the summation of original channels while the arithmetic relationships (also called spatial cues) between different channels (including level difference, phase difference and coherence) are exploited and transmitted as side information. The surround scene is then recovered on the basis of the downmix and the spatial cues. Binaural rendering has also been introduced into MPS [7], but additional complexity is introduced for driving proper HRTF pair from MPS cues, where inter-channel difference has to be translated into binaural difference.

Spatially Squeezed Surround Audio Coding (S^3AC) is a novel approach to spatial audio coding introduced by the authors in [8]. Rather than exploiting the inter-channel mathematical redundancy as in MPS, S^3AC exploits perceptual localisation redundancy [8, 9]. In the frequency domain, S^3AC analyses the sound field, represented by either multi-channel audio signals (e.g. ITU 5.1 [8]) or a microphone recording (e.g. Ambisonics [10]), to identify virtual sound sources and their location (in general, one virtual source is identified or generated for each frequency band). The S^3AC spatial squeezing process is then used to represent the original surround soundfield with a stereo signal. Due to the limited localisation resolution of human hearing [9], no perceptual loss in localisation is introduced in this process. This approach has the advantage that the source localisation derived in the S^3AC soundfield analysis directly represents the location of virtual sources. This source localisation information can be saved as side information for further bit reduction, and the stereo signal then reduced to a monophonic downmix signal [10].

This paper describes a solution that further exploits source localisation information in S^3AC for binaural reproduction. Human auditory perception of a localised sound source can be described by a pair of head-related transfer functions (HRTFs) [4]. For a given HRTF database (e.g. KEMAR HRTFs [11]), the proposed system (illustrated in Figure 1) derives the

978-1-4244-2179-4/08/$25.00 ©2008 IEEE

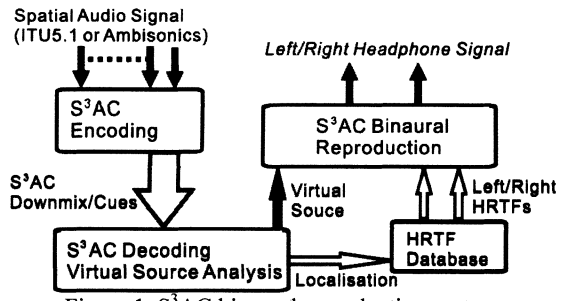

Figure 1. S³AC binaural reproduction system

virtual sound source and its localisation information in each frequency band from a S³AC stereo downmix (or mono downmix and S³AC cues), and filters the virtual source with the HRTFs measured at the same location. Since the S³AC localisation information directly represents the angular property of a sound source [8, 10], no additional analysis is required to translate the S³AC localisation information into angular information for matching to available HRTFs.

This paper is formatted as follows: Section 2 describes the proposed S³AC binaural reproduction system; Section 3 presents the evaluation over the proposed system; Conclusions are drawn in Section 4.

2. System

2.1. S³AC encoding

The proposed binaural reproduction does not require any modification to the S³AC encoding process [8]. For both 5.1-channel and Ambisonics signals, S³AC performs soundfield and virtual source localisation analysis in the frequency domain [8]. The transform between the time and frequency domains can be implemented by existing time-frequency transform methods, e.g. Pseudo-Quadrature Mirror Filterbanks (PQMF) or the Short-Time Fourier Transform (STFT) [12]; here an STFT is used. The resulting frequency bins/bands are re-grouped into perceptual filter bands, e.g. Equivalent Rectangular Bandwidth (ERB) [12], to remove perceptual irrelevancy for further processing.

In 5.1-channel signals, for each frequency band k, one virtual source is derived from a pair of channels a and b, represented by a mono signal:

$$S(k) = \sqrt{A_a^2(k) + A_b^2(k)}.e^{\phi_{ab}} \quad (1)$$

while its localisation is derived by applying inverse amplitude panning algorithms:

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

where $A_a(k)$ and $A_b(k)$ are the magnitude of the two channels, ϕ_{ab} is the phase information, $\varphi_{ab}(k)$ and $\theta_{ab}(k)$ are the channel-bisector azimuth and the source direction, respectively. Channels a and b are selected as the two channels rendering a single source, with multiple sources overlapped in time and frequency represented by S³AC cues [13].

For the representation of first-order Ambisonics signal sets, virtual sources are represented (in the frequency domain) by a scaled version of the W-channel:

$$S(k) = \sqrt{2} \cdot W(k) \quad (3)$$

while the localisation is derived by:

$$\theta(k) = \arctan\left(Y(k)/X(k)\right) \text{ or } \theta(k) = \arctan\left(Y(k)/X(k)\right) + \pi \quad (4)$$

where $W(k)$, $X(k)$ and $Y(k)$ are the three B-Format channels in two dimensional Ambisonics.

For both 5.1-channel and Ambisonics systems, the derived source localisation is mapped to a downmix localisation in a 60° stereo sound field by a unique linear mapping criteria such that $\theta_{DM}(k) = f(\theta(k))$ [8, 10]. The stereo downmix signal is then formed by amplitude panning the mono virtual source to its location in the downmix soundfield:

$$\begin{bmatrix} Y_L(k) \\ Y_R(k) \end{bmatrix} = S(k) \begin{bmatrix} \tan(60°) + \tan(\theta_{DM}(k)) \\ \tan(60°) - \tan(\theta_{DM}(k)) \end{bmatrix} \quad (5)$$

This process effectively creates a stereo soundfield containing localisation information from the original surround soundfield by exploiting the localisation redundancy of human hearing [9]. The stereo downmix signal can be coded by conventional audio coders e.g. AAC [12] to further reduce the bit-rate. A mono downmix can be also created by using the S³AC virtual source with the localisation information saved as side information for surround sound recovery [10].

2.2. S³AC binaural decoding

For binaural reproduction, the S³AC frequency domain virtual sources $S(k)$ and the respective localisation descriptors $\theta(k)$ are required. This can be either derived from one of the two following approaches. In the first, inverse amplitude panning is applied to the two downmix channels $Y_L(k)$ and $Y_R(k)$, which effectively reverses (5) and re-maps the localisation to a 360° soundfield according to $\theta(k) = f^{-1}(\theta_{DM}(k))$ [8]. In the second, the source and localisation information can be directly transmitted in the mono downmix and S³AC cues [10]. For each virtual source and frequency bin, the ear entrance signals can be derived by filtering the virtual source with the HRTFs measured at the same location. In the

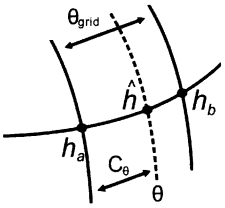

Figure 2. HRTF Interpolation

proposed system, a virtual source in each frequency band is filtered separately, which can be described in the frequency domain, for one frequency band k, by:

$$\begin{bmatrix} B_L(k) \\ B_R(k) \end{bmatrix} = S(k) \begin{bmatrix} H_L(k,\theta(k)) \\ H_R(k,\theta(k)) \end{bmatrix} \quad (6)$$

where $B_L(k)$ and $B_R(k)$ are the left and right ear entrance signals; $\theta(k)$ is the azimuth of virtual source in the kth band; $H_L(k, \theta(k))$ and $H_R(k, \theta(k))$ are the the kth band of the HRTFs measured at $\theta(k)$ for the left and right ear respectively. This method results in one pair of ear entrance signals for each frequency domain virtual source in each perceptual band, while the full bandwidth signal is assembled from the combined signals of all bands. This combined signal is then transformed to time-domain for binaural playback.

2.3. Azimuth interpolation of HRTFs

Due to the limited measurement resolution of the HRTF databases, HRTF interpolation must be performed to generate a database with adequate resolution, especially for dynamic sound sources. For example, in the KEMAR HRTF database, measurements were made for every 5 ° in the listening plane. In [10] , the authors confirmed the results of [9] for S^{3}AC, showing that a resolution of 1° in the frontal plane between ±30° should be employed. Hence the HRTFs need to be interpolated to a resolution of 1° in that region. In this work, the bilinear interpolation method [14] was used and simplified to 2D as:

$$\hat{h}(n) = (1 - \frac{C_\theta}{\theta_{grid}}) h_a(n) + \frac{C_\theta}{\theta_{grid}} h_b(n) \quad (7)$$

where $\hat{h}(n)$ is the time-domain head-related impulse response (HRIR) of a point located in an arbitrary azimuth between two adjacent measurement points a and b, C_θ is the related position and θ_{grid} is the original measurement resolution, as illustrated in Fig. 2. While S^{3}AC spatial cues provide localisation resolution of 1° in the front area [10], this interpolation process improves the resolution of HRTF database in S^{3}AC binaural reproduction system so that sufficient source localisation resolution is achieved. On the sides and in the rear, the perceptual resolution drops significantly to approximately 10° and 30° respectively [10]. Hence, the 5° resolution in the original HRTF database is sufficient for sources in these regions.

Table 1. Listening test mark guidelines

Localisation and Sound Quality Property	Mark
Transparent localisation and quality	100
Very accurate localisation with good quality	80
Accurate localisation with satisfactory quality	60
Little localisation with poor quality	40
No localisation with very poor quality	20
Noise or blank	0

3. Evaluations

The proposed S^{3}AC binaural reproduction system was evaluated by subjective listening tests. A modified MUSHRA [15] approach was employed. The original multi-channel surround audio signals were used as a reference for evaluating binaurally reproduced S^{3}AC coded signals. Since there is no perfect transformation from a multi-channel signal set to the binaural signals in terms of both sound quality and localisation, no hidden reference was used in this modified MUSHRA. Instead, listeners were instructed to directly compare the audio quality and localisation performance of the headphone playback versions against the original multi-channel sounds. A mark out of 100 was then given to each binaural item according to the guidelines given in Table 1. Since the KEMAR HRTFs were recorded without room reflections, this listening test was conducted in an anechoic chamber to ensure reverberant-free playback of the multi-channel references; hence, the multichannel audio and HRTF filtered headphone versions could be compared by listeners. Several S^{3}AC conditions were evaluated:

a. Mono: S^{3}AC mono downmix of original multi-channel audio with S^{3}AC spatial cues quantised to approximately 6kbps [10], and binaural decoded using original KEMAR HRTF database.
b. Stereo: S^{3}AC stereo downmix [10] and binaural decoded using original KEMAR HRTF database.
c. Interpo: S^{3}AC stereo downmix [10] and binaural decoded by interpolated KEMAR HRTF database.

Three additional conditions (AAC Mono, AAC Stereo, AAC Interpo) were also evaluated, where mono/stereo downmixs in condition a, b and c were further coded by AAC in 64kbps per channel. An anchor signal was created using a 3.5kHz low-pass filtering of the mono version of each test signal. Eight test signals listed in Table 2 were used. All signals were 44.1kHz 16-bit PCM signals. Genelec 8020A monitor speakers and Sennheiser HD250 monitor headphones were used for the multi-channel and binaural playback respectively. Eight listeners including both experienced and non-experienced listeners participated in the tests.

Table 2. Listening test signals

No.	Name	Type	No.	Name	Type
1	Airbus	5.1	5	Mosquito	5.1
2	Ambulance	5.1	6	Ring	Ambi
3	Female	5.1	7	Male	5.1
4	Flea	Ambi	8	Taxi	Ambi

The results including mean and 95% confidence intervals are shown in Fig. 3, and are shown separately for each file tested. The results show that, in comparison with multi-channel playback, the proposed S^{3}AC binaural reproduction systems achieves marks over 80 for most of the test files (i.e. very accurate localisation with good quality), and over 60 for the remainder of the test files (i.e. good localisation with satisfactory quality). Ambisonic files show lower average marks compared with other files and larger confidence intervals. It is suggested that this is due to the more complex soundfields represented by these files, which contain multiple sources and fast moving objects, and which led to non-expert listeners finding difficulties in comparison and marking.

Little improvement is introduced by the HRTF interpolations, while no significant degradation is caused by the S^{3}AC quantisation process in the Mono mode. This indicates a lower localisation resolution requirement in binaural reproductions when compared to loudspeaker playback. AAC compression does not introduce significant perceptual quality loss either. This proves the backward compatibility of the proposed approach (by using AAC, S^{3}AC downmixes can be transmitted and stored by any stereo audio system) as well as the bit-rate efficiency. The AAC Mono mode can achieve a bit-rate as low as 70kbps (64kpbs for mono downmix and 6kbps for S^{3}AC cues), while headphone playback with very accurate source localisation and good sound quality can be achieved.

4. Conclusions

S^{3}AC is an efficient spatial audio compression approach based on soundfield analysis and squeezing. A binaural reproduction approach for S^{3}AC has been presented. This approach utilises the S^{3}AC frequency domain virtual source and localisation information for HRTF filtering to generate localised binaural signals from S^{3}AC compressed spatial audio signals. No additional complexity is required to transform S^{3}AC source localisation information to the HRTF domain. Subjective evaluation of the S^{3}AC binaural reproduction system showed that, in comparison with the original multi-channel loudspeaker playback, the approach achieves accurate source localisation as well as good sound quality over headphones. A bit-rate as low as 70kbps can be used by the proposed system for transmission with no noticeable perceptual distortion introduced in comparison with higher bit-rates.

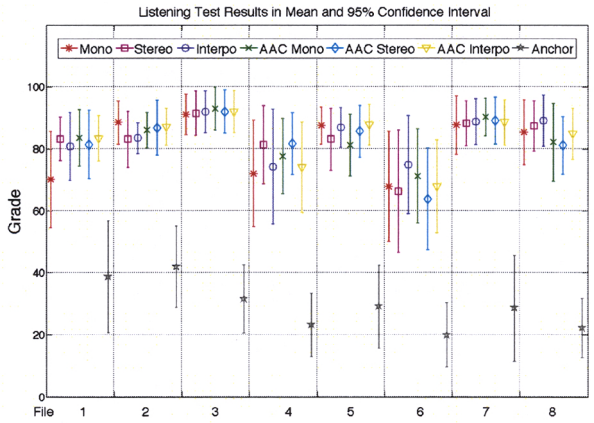

Figure 3. Listening test results for each of the 8 test files.

5. References

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

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

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

[4] Cheng, C.I., Wakefield, G.H., Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency and Space, *J. Audio Eng. Soc.*, 49(4):231-249, Apr. 2001.

[5] Bauck, J., Cooper, D.H., Generalized Transaural Stereo and Applications, *J. Audio Eng. Soc.*, vol.44, No.9, Sep.1996.

[6] Breebaart, J., Faller, C., Spatial Audio Processing: MPEG Surround and Other Applications, *Wiley*, USA, 2007.

[7] Breebaart, J., Analysis and Synthesis of Binaural Parameters for Efficient 3D Audio Rendering in MPEG Surround, *IEEE ICME 2007*, Beijing, China, Jul. 2007.

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

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

[10] Cheng, B., Ritz, C., Burnett, I., A Spatial Squeezing Approach to Ambisonic Audio Compression, *IEEE ICASSP 2008*, Las Vegas, USA, Mar. 2008.

[11] Gardner, B., Martin, K., HRTF Measurements of a KEMAR Dummy-Head Microphone, *MIT Media Lab Perceptual Computing* – Technical Report #280, 1994.

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

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

[14] Begault, D.R., 3D Sound for Virtual Reality and Multimedia, *Academic Press*, Cambridge, MA, USA, 1994.