Psychoacoustic-based quantisation of spatial audio cues

B. Cheng, C.H. Ritz and I.S. Burnett

The derivation of spatial cues representing source localisation information is a typical component of multichannel spatial audio coders. Efficient compression of spatial cues based on psychoacoustic localisation features is investigated. Results show that the proposed quantisation approach for spatial cue compression achieves bit-rates of less than 6 kbit/s while preserving critical source localisation information.

Introduction: Spatial audio coding (SAC) has attracted significant interest in recent years, and aims to provide channel independent compression of multichannel spatial audio signals as well as maintaining backward compatibility to a conventional stereo/mono system. A typical 'downmix + cues' coder is MPEG Surround [1], where all original channels are summed to form a stereo/mono downmix and time/level differences and coherence information between original channels are conveyed by spatial cues. Alternatively, the authors' approach, spatially squeezed surround audio coding (S^3AC) [2] exploits the redundancy of human sound localisation [3] to squeeze multichannel audio (representing a 360° soundfield) into stereo audio (representing a 60° soundfield). This process is based on a source localisation oriented soundfield analysis and requires no side information since the spatial information (i.e. localisation cues) can be derived from the downmix for surround sound reproduction. In addition, side information was introduced into S^3AC [4] for efficient coding of complex sound environments where sources overlap in time and frequency, as well as to allow for a mono downmix to achieve lower bit-rate than stereo downmix [5].

To minimise bit-rates, the spatial cue side information must be efficiently quantised for transmission. While scalar quantisation is used in [1], this Letter presents an alternative quantisation solution composed of two stages: a psychoacoustical codebook exploiting human source location perception; and frame-wise differential coding to further reduce the spatial cue quantisation bit-rate.

Derivation of spatial cues: This Letter investigates the quantisation of spatial cues derived using S^3AC. Five standard 44.1 kHz/16-bit ITU 5.1 multichannel spatial audio signals, including immersive audio, live recording and movie sound track, are used for evaluation. Here, the LFE (.1 channel) is ignored. For each channel, a 50% overlapped 1024-point short-time-Fourier-transform (STFT) is applied, with the frequency bins further grouped into 20 bands equivalent to double ERB frequencies [6] for each frame. For each band, spatial cues representing source locations are derived through S^3AC amplitude analysis [5]. Using 16 bits per cue results in a total bit-rate of 27.56 kbit/s for all channels, if no further quantisation is performed.

Psychoacoustical codebook design: Human sound localisation precision is highly dependent on source location. In an ideal listening environment, the precision is approximately 1° in front of a listener and reduces to less than 10° at the sides and rear [3]. Recent subjective experiments evaluating the perception of spatially panned sound sources in practical reproduction environments [5] have shown that using a lower precision in coded spatial audio does not introduce perceivable distortion in localisation. In particular, in the frontal region between ±30°, reducing the localisation resolution from 1° to 2° gives no perceptual distortion, which suggests that 30 discrete azimuths are adequate for quantising frontal sources. Similarly, for the sides and rear, resolutions of 5° and 35° (32 and 4 discrete azimuths, respectively), have been found satisfactory [5]. Based on these results, a 6-bit quantisation codebook described in Table 1, with 64 discrete azimuths non-uniformly distributed between regions in the 360° circle was derived. A 5-bit codebook giving further bit-rate reduction is also described in Table 1. These codebooks result in a fixed bit-rate of 10.36 and 8.61 kbit/s, respectively. According to the subjective evaluation in [5], these codebook designs will result in no significant localisation distortion in the vital frontal region, while the degradation in other regions is less than ten MUSHRA [7] scores when compared to multichannel reference where source locations are not quantised.

Table 1: Psychoacoustical codebook design

Regions	Linear azimuth resolution		Number of azimuths in region	
	6-bit	5-bit	6-bit	5-bit
Front [−30°, 30°]	2°	3°	31	21
Left (30°, 110°]	6.67°	20°	13	4
Right (−30°, −110°]	6.67°	20°	13	4
Rear left (110°, 180°]	17.5°	35°	4	2
Rear right (−180°, −110°)	17.5°	35°	3	1
Total number of azimuths			64	32

Differential coding of spatial cues: For further bit-rate reductions, lossless differential coding of the codebook quantised spatial cues can be used. Owing to the band perception property of the human auditory system [6], each frequency band can be assumed to represent a single sound source. Hence, it is expected that the location of the source varies smoothly over time, resulting in highly correlated cues. In contrast, spatial cues between adjacent frequency bands represent different sound sources and hence will show less correlation. Therefore, the redundancy remaining in the spatial cues of one frequency band can be further removed using frame-wise differential coding. In this approach, the difference between spatial cue codebook indices derived for the same frequency band k between two adjacent frames is derived as

$$d_n^k = C_n^k - C_{n-1}^k \quad k = 1, 2, \ldots 20; n = 2, 3, \ldots N$$

where C_n^k and C_{n-1}^k are the codebook indices for the kth frequency band of the nth and $(n-1)$th time frame. Here, N represents the differential prediction length; hence, a sequence of quantised spatial cue indices is represented by the anchor index (the first quantised spatial cue index) followed by $N - 1$ differential values.

Spatial cues derived from the five tests are used to evaluate this differential coding approach. Fig. 1 shows the histogram of both differential values d_n^k and the original codebook azimuth indices C_n^k using a 6-bit codebook. It is shown that, while the probability of the quantised azimuths is evenly distributed over all indices, the differential coded result has a highly centralised distribution, which can be exploited by entropy coding. In this Letter, standard Rice coding [8] is utilised.

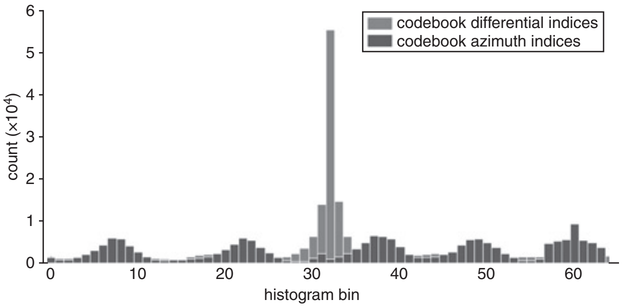

Fig. 1 *Histogram of codebook azimuth indices (ranging from 0 to 63) and codebook differential indices (ranging from −31 to 31) for 175 000 cues derived from five test files*

Results: The proposed codebook quantisation and differential coding approaches were evaluated using the five test files. Several differential prediction lengths N from five frames to the same as the total number of frames in a file are examined. The resulting cue bit-rates for 6- and 5-bit codebooks are given in Fig. 2. For the 6-bit and 5-bit codebooks, respectively, using five frames for the prediction length reduces the bit-rates from 10.36 and 8.61 kbit/s to approximately 6.4 and 6 kbit/s, while using 50 frames further reduces the bit-rates to 5.5 and 5.4 kbit/s. Table 2 gives the resulting azimuth error from the overall approach evaluated using:

$$E_n^k = \left| C_n^k - \hat{C}_n^k \right|$$

where C_n^k and $\hat{C}_n^k$ are the original and quantised azimuths. Errors in different regions are also described separately. File 1, which is a recording of concert hall applause having widespread distributed localisation, shows the highest error, while the error in other signals is limited. The vital front region has least error, while more errors are caused by the lower codebook precision in the sides and rear. However, according to results in [5], the perceptual distortion on localisation caused by these errors is limited owing to the perceptual localisation oriented codebook design.

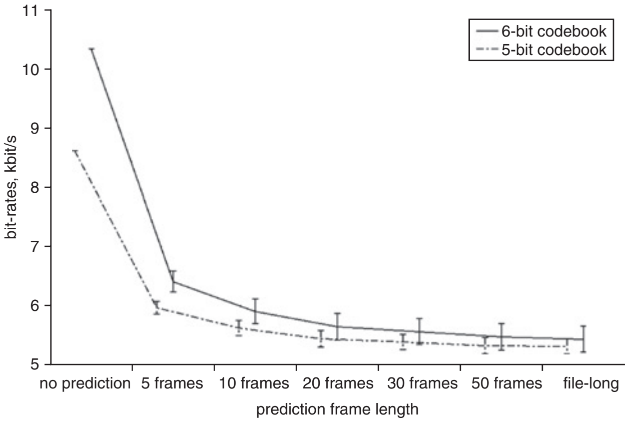

Fig. 2 *Average cue bit-rate (in kbit/s) and 95% confidence intervals of test files for different prediction lengths using 6-bit and 5-bit codebooks*

Table 2: Azimuth error (in degrees) compared with original signal

Region	Average		Front		Sides		Rear	
Codebook	6-bit	5-bit	6-bit	5-bit	6-bit	5-bit	6-bit	5-bit
File 1	5.9	10.4	0.5	0.7	1.6	5.0	12.7	27.0
File 2	0.5	1.2	0.3	0.4	1.4	4.4	N/A	N/A
File 3	1.6	2.1	0.4	0.6	3.0	5.7	5.2	11.9
File 4	2.3	3.4	0.5	0.6	1.2	4.1	7.2	15.2
File 5	1.8	2.6	0.3	0.4	3.3	5.9	6.0	13.9

Conclusions: An efficient quantisation and compression method for spatial cues has been presented. The approach contains two parts: psychoacoustical codebook quantisation and lossless differential coding. This approach achieves less than 6 kbit/s bit-rate for compressing S^{3}AC spatial cues, while location dependent codebook design limits the perceptual localisation degradation. This quantisation approach not only fits the S^{3}AC cues but any SAC technique that contains spatial cues representing source localisation information. When combining with a mono-downmix quantised at 64 kbit/s (e.g. using the Advance Audio coder (AAC) as described in [1]), this approach provides full surround sound coding at around 70 kbit/s. An investigation into alternative lossless entropy coding techniques may provide further bit-rate reduction.

28 April 2008
Electronics Letters online no: 20081199
doi: 10.1049/el:20081199

B. Cheng, C.H. Ritz and I.S. Burnett (*Whisper Labs, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia*)

References

1 Breebaart, J., and Faller, C.: ‘Spatial audio processing: MPEG surround and other applications’ (Wiley, New York, 2008)
2 Cheng, B., Ritz, C., and Burnett, I.: ‘Principles and analysis of the squeezing approach to low bit rate spatial audio coding’. Proc. IEEE ICASSP 2007, Honolulu, USA, April 2007
3 Blauert, J.: ‘Spatial hearing: the psychophysics of human sound localisation’ (MIT Press, Cambridge, MA, 1996)
4 Cheng, B., Ritz, C., and Burnett, I.: ‘Encoding independent sources in spatially squeezed surround audio coding’. Proc. PCM 2007, Hong Kong, People’s Republic of China, December 2007
5 Cheng, B., Ritz, C., and Burnett, I.: ‘A spatial squeezing approach to ambisonic audio compression’. Proc. IEEE ICASSP 2008, Las Vegas, USA, March 2008
6 Glasberg, B.R., and Moore, B.C.J.: ‘Derivation of auditory filter shapes from notched-noise data’, *Hear. Res.*, 1990, **47**, pp. 103–138
7 ITU-R BS. 1534 ‘Method for the subjective assessment of intermediate quality level of coding system (MUSHRA)’, 2001
8 Rice, R.F.: ‘Some practical universal noiseless coding technique’, *JPL Publ.*, 1979, pp. 79–22