A FULLY SCALABLE AUDIO CODING STRUCTURE WITH EMBEDDED PSYCHOACOUSTIC MODEL

Te Li[1,2], *Susanto Rahardja*[1,2]

[1] Institute for Infocomm Research (I^{2}R)
21 Heng Mui Keng Terrace, Singapore 119613

Soo Ngee Koh [2]

[2] Nanyang Technological University
50 Nanyang Avenue, Singapore 639798

ABSTRACT

A fully scalable audio coding structure based on a novel combination of the non-core MPEG-4 scalable lossless audio coding (SLS), the state-of-the-art psychoacoustic model, joint stereo coding and the perceptually prioritized bit-plane coding is presented in this paper. The psychoacoustic information is implicitly embedded in the scalable bitstream with negligible amount of side information and trivial modification to the standardized SLS decoder. Results of extensive evaluation show that the subjective quality of scalable audio is improved significantly.

Index Terms— Audio coding, Joint stereo coding, Bit-plane coding.

1. INTRODUCTION

Scalable audio coding allows an encoder to compress data at a high/lossless bitrate and a receiver to decode the compressed signal at different fidelity levels according to the specific applications and user context requirements as well as available device capabilities. This scalability is very important for the co- and inter-operability of today's myriad of multimedia applications spanning across various media platforms including cell phone networks, computer networks and broadcasting networks where each one generally requires a different range of coding bitrate.

MPEG-4 scalable lossless (SLS) coding was released as a standard audio coding tool in June 2006. It allows the scaling up of a perceptually coded representation such as MPEG-4 advanced audio coding (AAC) [1] to a lossless representation with a wide range of intermediate bitrate representations. For some cases where full scalability or near-full scalability are desired, a non-core mode of SLS is available where only the enhancement layers are presented. The general structure of non-core SLS encoder is shown in Fig. 1. The input audio in integer PCM format is losslessly transformed into the frequency domain by using the integer modified discrete cosine transform (IntMDCT) [2]. The spectrum is then coded using bit-plane golomb code (BPGC) [3] combined with context-based arithmetic coding (CBAC) and low energy mode coding (LEMC) [4] to generate the scalable bitstream. More details of the SLS can be found in [5][6].

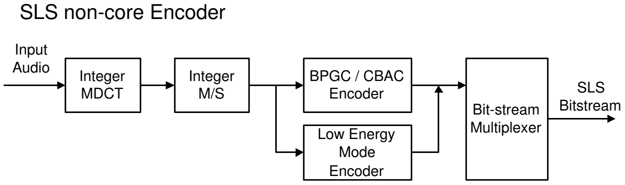


Fig. 1. Structure of SLS non-core encoder and decoder.

Due to the lack of perceptual coder, the audio quality delivered by non-core SLS at common lossy bitrate is far worse than the quality delivered by the state-of-the-art perceptual audio coders (such as AAC) at the same bitrate. A low-complexity enhancement design on the bit-plane coding of non-core SLS is proposed in [7]. By using as little as one bit per frame of the side information, the bit-plane coding order is switched according the energy distribution in different regions of the spectrum. Besides the signal energy, no actual "perceptual information" such as the state-of-the-art psychoacoustic mask is applied in the enhancement. The results show that this method has successfully enhanced the perceptual quality for most of the test sequences by different levels. The advantages of this design include the extremely low complexity and waiver of significant change on the bitstream structure. However, the quality enhancement is thus limited without adopting the psychoacoustic information.

In this paper, a more complicated enhancement scheme with psychoacoustic model is proposed. An important feature of this structure is that the added complexity only distributes to the encoder side, where the extra side information amount and the decoder structure remain the same as the previous scheme in [7]. This is achieved by implicitly embedding the psychoacoustic information in the scalable coding process. The detailed proposed structure is described in the next section, with the elaboration on the two main components: the bitrate allocation to the stereo channels and the spectrum division for bit-plane coding. The performance of the truncated audio at different bitrates are evaluated in Section 3 and in Section 4 the conclusions are given accordingly.

1-4244-1484-9/08/$25.00 ©2008 IEEE

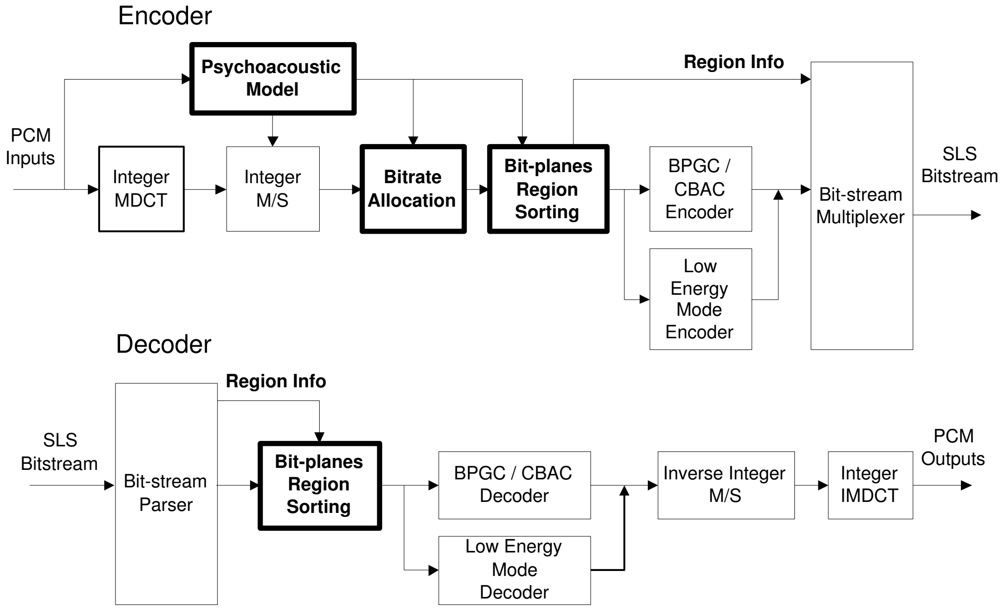


Fig. 2. Structure of enhanced SLS non-core encoder and decoder with embedded psychoacoustic model.

2. SLS NON-CORE STRUCTURE WITH EMBEDDED PSYCHOACOUSTIC MODEL

The proposed SLS non-core structure with embedded psychoacoustic model is depicted in Fig. 2, with the newly added blocks highlighted with bold borders. Based on the signal to mask ratio (SMR) calculated from the psychoacoustic model, the bits available for each frame are allocated to the two channels, either M/S or L/R. Subsequently, the different regions of the spectrum are sorted with perceptual priorities according to the SMR. The bit-plane coding starts from the region with the highest priority, followed by the regions with succeeding priorities until the bits available are used up or the all the regions are coded. The coding sequence of the spectrum regions are coded as side info in the bitstream for the corresponding decoding process. The detailed procedures are elaborated in the following two subsections.

2.1. Bitrate Allocation for Stereo Channels

As one of the most popular joint stereo encoding techniques, mid/side (M/S) stereo coding is widely used in many audio codecs such as MPEG-1 Layer 3 (MP3) [8] and MPEG-4 advanced audio coding (AAC) [1]. M/S coding transforms the left (L) and right (R) channels into a mid (M) channel and a side (S) channel. In SLS [6], M and S channels are computed by an integer operation using Givens rotation

$$\begin{bmatrix} M \\ S \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} L \\ R \end{bmatrix}$$

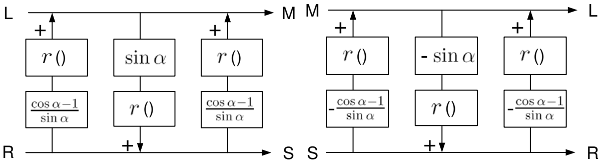
where $\alpha = -\pi/4$. The Givens rotation can be factorized using a lifting scheme [9] as shown in Fig. 3, where r is defined as a rounding operation $r : \mathbb{R} \rightarrow \mathbb{Z}$. Re-arranging the data into M and S channels usually results in a situation where the M channel has much larger value than the S channel if the L and R channels are highly correlated. In this case, the S channel can then be accurately encoded using fewer bits and more resources can then be employed efficiently on the M channel.

In the proposed structure, the algorithm in informative encoder of MPEG-4 AAC [1] is adopted to decide if M/S coding is switched on. the masking thresholds of L and R channels are firstly calculated using the psychoacoustic model. The masking thresholds of M and S channels are then calculated based on the basic thresholds of M/S which can be obtained using the same model of L/R, together with masking level difference (MLD) [10]. By quantizing the channel coefficients according to the thresholds, the M/S coding is switched on if the number of bits actually required to code M/S is less than the number of bits required to code L/R.

For each frame, the bits are evenly distributed to the L and R channels if M/S coding is not switched on. Otherwise, the bits are allocated to the M channel is computed proportional to the *perceptual weight* (PW) which is similar as the perceptual entropy [11]. Specifically, the PW for a channel is defined as

$$PW = \sum_{i=0}^{I-1} \left[\left(10 \log_{10} \frac{E_i}{T_i} \right) \cdot (O_{i+1} - O_i) \right] \qquad (1)$$

where I is the total number of scalefactor bands (sfb) and O_i is the offset spectrum line of the sfb i. E^i and T^i are the signal


L
M
M
L
+
+
+
+
$r()$
$\sin\alpha$
$r()$
$r()$
$-\sin\alpha$
$r()$
$\frac{\cos\alpha-1}{\sin\alpha}$
$r()$
$\frac{\cos\alpha-1}{\sin\alpha}$
$-\frac{\cos\alpha-1}{\sin\alpha}$
$r()$
$-\frac{\cos\alpha-1}{\sin\alpha}$
R
+
S
S
+
R

(a) (b)

Fig. 3. (a) Integer M/S (b) Inverse integer M/S.

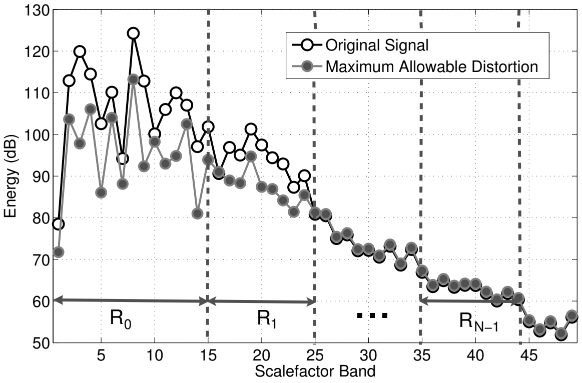


Fig. 4. Spectrum plot for one frame from haffner.wav.

energy and the masking threshold of sfb i, respectively. The bits allocated to M channel is computed as

$$B_M = B \cdot \frac{PW^M}{PW^M + PW^S} \tag{2}$$

where B is the total bits for the frame and B_M is the bits allocated to M channel. PW^M and PW^S are the PW of M and S channels, respectively. The remaining bits of th e frame are then allocated to the S channel.

2.2. Prioritized Bit-plane Coding

The concept of prioritized bit-plane coding implies a combination of two procedures. Firstly, the different spectrum regions of bit-planes are prioritized according to a certain criteria. It is followed by a coding process that the priorities are applied (to the non-lazy [3] bit-planes in SLS). The major difference between the structure proposed in this paper and the previous one in [7] is that instead of considering the energy distribution as the only criteria, the masking threshold is applied as well. The division algorithm of the regions is also modified.

Fig. 4 shows the spectrum plot of one frame from an audio sequence named "haffner" (48kHz/16bit). The masking threshold is computed using the psychoacoustic model in MPEG-4 AAC reference encoder. In [7] it is shown that for a spectrum with 49 sfbs, the last 5 sfbs (44 - 48) are always assigned with the lowest priority as the coefficients in this region always fall into LEMC coding. This is also applied in the current structure. The remaining sfbs from 0 - 43 are then divided into N regions

$$\mathcal{R} = \{\mathbf{R}_0, \mathbf{R}_1, ..., \mathbf{R}_n, ..., \mathbf{R}_{N-1}\}, \;\; 2 \le N \le 44 \tag{3}$$

where each region $\mathbf{R}_n$ contains sfbs start from $\widehat{O}_n$ to $\widehat{O}_{n+1}-1$ with $\widehat{O}_n$ indicating the offset sfb of the region. The PW_n for $\mathbf{R}_n$ can be computed as

$$PW_n = \sum_{i=\widehat{O}_n}^{\widehat{O}_{n+1}-1} \left[\left(10 \log_{10} \frac{E_i}{T_i} \right) \cdot (O_{i+1} - O_i) \right] \tag{4}$$

$\mathcal{R}$ is then sorted according to the perceptual weights of the regions by

$$P(\mathbf{R}_n) > P(\mathbf{R}_m) \text{ if } \begin{cases} PW_n > PW_m, \text{ or} \\ PW_n = PW_m \;\&\; n < m \end{cases} \tag{5}$$

where $P(\mathbf{R}_n)$ indicates the *perceptual priority* of the region. The bit-plane coding can then start from the region with the highest priority, followed by the regions with the subsequent priorities. A side information of $\lceil \log_2(N!) \rceil$ bits/frame is transmitted to indicate the sequence of decoding.

With the above general definition, a simpler design is adopted in the proposed structure. The N regions are further divided into two *groups*

$$\begin{aligned} \mathcal{G}_0 &= \{\mathbf{R}_0, \mathbf{R}_1, ..., \mathbf{R}_K\}, & (6) \\ \mathcal{G}_1 &= \{\mathbf{R}_{K+1}, \mathbf{R}_{K+2}, ..., \mathbf{R}_{N-1}\} & (7) \end{aligned}$$

where

$$PW_K > 0, \sum_{n=K+1}^{N-1} PW_n = 0 \tag{8}$$

The bit-plane coding then takes the sequence that

1. The coding starts from the first maximum bit plane (MSB) till all the non-lazy bit-planes in $\mathbf{R}_0$ are finished. It is then followed by $\mathbf{R}_1$,...,$\mathbf{R}_K$ till all the non-lazy bit-planes of regions in $\mathcal{G}_0$ are coded.
2. Apply the same coding manner as in 1 to $\mathcal{G}_1$.
3. Code the lazy bit-planes in $\mathcal{G}_0$, followed by $\mathcal{G}_1$.
4. Code the region of the last 5 sfbs by LEMC.

In this way, the side information to be coded is then reduced to $\lceil \log_2 N \rceil$ bits/frame. Another important issue of this design is choosing the values for parameters including N and the boundary position of each region. Through the training of a large database of audio sequences, the optimized value of N is fixed at 4 and the region boundaries are fixed at $\widehat{O}_1 = 18, \widehat{O}_2 = 24$ and $\widehat{O}_3 = 31$ in our implementation.

Table 1. Test items for MPEG-4 SLS (48kHz/16bit stereo)

no.	Items (.wav)	no.	Items (.wav)	no.	Items (.wav)
1	avemaria	6	cymbal	11	haffner
2	blackandtan	7	dcymbals	12	mfv
3	broadway	8	etude	13	unfo
4	cherokee	9	flute	14	violin
5	clarinet	10	fouronsix	15	waltz

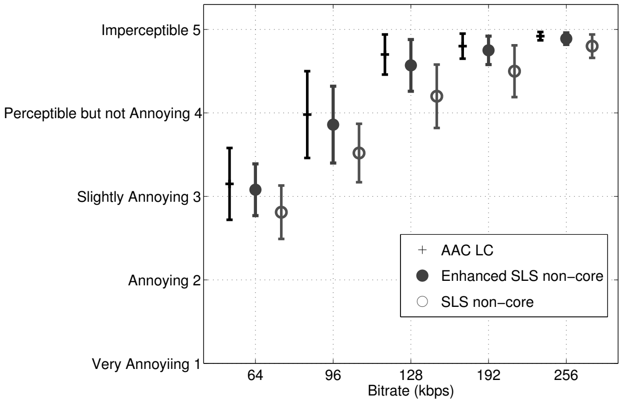


Fig. 5. Subjective test results of enhanced SLS non-core with comparison of SLS non-core and AAC LC at variable lossy bitrates.

3. PERFORMANCE

The proposed structure is denoted by ESLS (for enhanced SLS). The perceptual quality of lossy audio reconstruction by ESLS is evaluated based on the subjective test ITU-R BS.1116 [12] and compared with the results from original non-core SLS and MPEG-4 AAC LC reference codecs (informative version, not optimized). The test sequences are listed in Table 1.

The summarized results are plotted in Fig. 5, with the grading scale ranges from 1 ("Very Annoying") to 5 ("Imperceptible"). It can be observed that for the bitrate ranges from 64 to 256 kbps which are most commonly used bitrates for lossy audio coding, the performance of ESLS is almost comparable to that of the AAC LC. In addition, as only one bit/frame is added (as another bit can be implemented by the reserved bit in SLS structure) compared to the original SLS bitstream, the lossless compression performance is not affected.

4. CONCLUSIONS

Due to the lack of perceptual core, the perceived quality of audio coded by non-core mode of SLS at common lossy bitrates are usually not as good as the quality obtained by the state-of-the-art perceptual audio coders. An enhanced SLS non-core structure with an implicitly embedded psychoacoustic model is proposed in this paper. With trivial modification to the standardized decoder structure, the quality performance of non-core SLS at common lossy bitrates is comparable to that of the AAC LC reference codecs.

5. REFERENCES

[1] "ISO/IEC 14496-3, information Technology - coding of audiovisual objects, part 3. audio," 1998.

[2] R. Geiger, T. Sporer, J. Koller, and K. Brandenburg, "Audio coding based on integer transform," *111th AES Conv.*, , no. 5471, 2001.

[3] R. Yu, C.C. Koh, S. Rahardja, and X. Lin, "Bit-plane golomb code for sources with laplacian distributions," *Proc. Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, vol. 4, pp. 277–280, Apr. 2003.

[4] R. Yu, S. Rahardja, X. Lin, and C.C. Koh, "Improving coding effciency for mpeg-4 audio scalable lossless coding," *Proc. ICASSP*, vol. 3, pp. 169–172, Mar. 2005.

[5] R. Geiger, R. Yu, J. Herre, S. Rahardja, S.W. Kim, X. Lin, and M. Schmidt, "Iso/iec mpeg-4 high-definition scalable advanced audio coding," *120th AES conv.*, May 2006.

[6] "ISO/IEC 14496-3:2005/amd 3, scalable lossless coding (SLS)," 2006.

[7] T. Li, S. Rahardja, and S.N. Koh, "Perceptual enhancement for fully scalable audio," *Int. Workshop on Multimedia Signal Process.*, Oct. 2007.

[8] "ISO/IEC 11172-3, coding of moving pictures and associated audio for digital storage media at up to 1.5mbit/s, part 3. audio," 1993.

[9] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *Tech. Rep., Bell Laboratories*, 1996.

[10] J.D. Johnston and A. Ferreira, "Sum-difference stereo transform coding," *Proc. ICASSP*, vol. 2, pp. 569–572, Mar. 1992.

[11] J.D. Johnston, "Estimation of perceptual entropy using noise masking criteria," *Proc. ICASSP*, vol. 5, pp. 2524–2527, Apr. 1988.

[12] "ITU-R BS.1116, Methods for the subjective assessment of small impairments in audio systems including multi - channel sound system," .