PERCEPTUALLY ENHANCED BIT-PLANE CODING FOR SCALABLE AUDIO

Rongshan Yu, Te Li, Susanto Rahardja

Audio Processing Lab, Institute for Infocomm Research (I^2R)
21 Heng Mui Keng Terrace, Singapore 119613

ABSTRACT

The MPEG-4 *Scalable to Lossless* (SLS) audio coding is recently being developed to provide a unified solution for high - compression perceptual audio coding and high-quality lossless audio coding. SLS provides efficient *Fine Granular Scalable* (FGS) coding from AAC core layer to lossless, and achieves reasonable perceptual quality at its scalable coding range using a sequential bit-plane scanning method, which minimizes the audio distortion according to the spectral shape of the core layer quantization errors. In this paper, it is shown that the perceptual quality performance of SLS at intermediate rates can be further improved by incorporating psychoacoustic model into the bit-plane coding process. In addition, it is also found that such an improvement can be achieved by slightly tweaking the original bit-plane coding process of SLS and hence preserving its nice features such as compatibility to lossless coding and low complexity.

1. INTRODUCTION

Recently, with the advance of network and storage technologies, it is becoming realistic that people will enjoy high sampling rate, high resolution audio contents with lossless quality. Envisioning such a need, the international standardization body MPEG has recently introduced a scalable tool for lossless audio coding, namely, MPEG-4 Audio Scalable to Lossless (SLS) [1] coding. MPEG-4 SLS integrates the functionalities of lossless audio coding, perceptual audio coding, and fine granular scalable audio coding in a single framework; meanwhile it provides backward compatibility to MPEG-4 Advanced Audio Coding (AAC) [2] at the bit-stream level. This new tool, in combination with the existing MPEG audio toolset, provides a universal digital audio format that can be used in a variety of application domains such as professional audio, Internet music, consumer electronics and broadcasting.

The original SLS employs a very simple embedded coding principle to optimize the perceputal quality at its scalable coding range which is done by simple sequential bit-plane coding from *Most Significant Bit* (MSB) to *Least Significant Bit* (LSB) over all frequency bands on the core layer quantization error. In such a way, the spectral shape of the AAC quantization error which has been optimized by the AAC core layer quantizer, is preserved during the bit-plane coding process. This method is only effective when the AAC core layer is working at high rates ($\geq$64kbit/s/ch). However, when AAC core layer is working at lower rates, or in non-core mode operation where the AAC bit-rate is zero, this method doesn't perform very well since the spectral shape of the residual signal is far from the optimal as in these cases noise shaping is usually performed in only limited frequency range due to the bit-rate constraint at the core layer. In this paper, it is shown that this problem can be solved by incorporating human perceptual model in the bit-plane coding process. This can be achieved by simply shifting bit-plane of each frequency band according to its significance with respect to the masking threshold information before the sequential bit-plane coding process. Experimental results show that despite its simplicity, the proposed method dramatically improve the perceptual quality of SLS at its intermediate bit-rates without introducing significance overhead in terms of lossless coding efficiency or complexity.

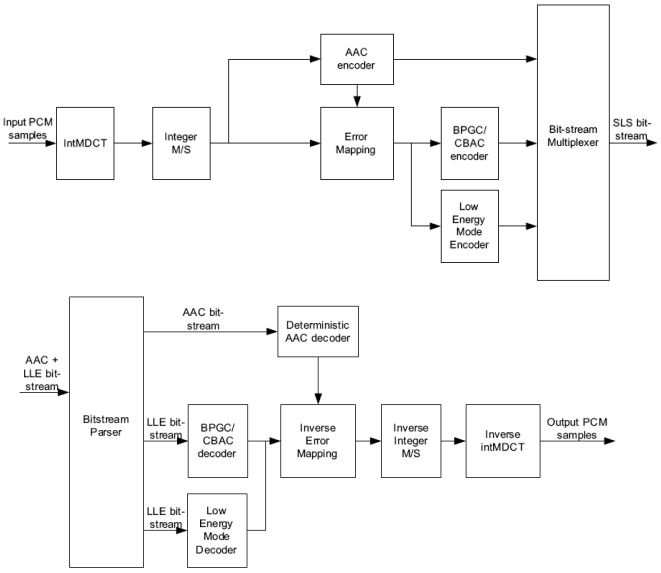

Figure 1: Structure of SLS encoder and decoder.

2. OVERVIEW OF SLS

The system diagram of the SLS codec is given in Figure 1, which comprises of two distinguished layers in both the encoder and decoder, namely, a core layer and an Lossless Enhancement (LLE) layer. In particular, the core layer is simply an MPEG-4 AAC codec. In the SLS encoder, the input audio signal which is integer PCM format is losslessly transformed to the frequency domain by using the *Integer Modified Discrete Cosine Transform* (IntMDCT) [3]. The resulting IntMDCT coefficients $c[k]$, where $k = 1, ...1024$, are passed to the core layer AAC encoder to generate the core layer AAC bit-stream. In the AAC encoder, $c[k]$ are first grouped into scalefactor bands (sfbs), which are then quantized with non-uniform quantizer, usually with different quantization steps in different sfb to shape the quantization noise so that

they can be best masked [2].

In order to efficiently employ the information of the spectral data that has been carried in the core layer bitstream, the following error-mapping procedure is employed to generate the residual spectrum $e[k]$ coded in the LLE layer:

$$e[k] = \begin{cases} c[k] & i[k] = 0 \\ |c[k] - thr[k]| & i[k] \neq 0 \end{cases} \quad (1)$$

Here $thr[k]$ is the closer-to-zero quantization threshold (rounded) for $c[k]$ in the core layer AAC quantizer, and $i(k)$ is the quantized IntMDCT spectral data vector produced by the AAC quantizer. In addition, for the sfb s that contains non-zero value quantized signal (implicit band), side information necessary for the bit-plane decoding process is determined implicitly from the core layer quantizer and hence doesn't need to be transmitted from the encoder. Otherwise, sfb band s is coded as an explict band that requires all the necessary side from the encoder [4].

The IntMDCT residual spectrum output $e[k]$ is then bit-plane coded to generate the scalable LLE layer bitstream. As the first step, the *Most Significant Bit* (MSB) for spectral data from all scalefactor bands is coded. After that, the coding process is progressed to the 2nd MSB, 3rd MSB and so on until it reaches the *Least Significant Bit* (LSB) for all scale factor bands. As a result, the level of the noise spectrum will be decreased progressively during this coding process. Finally, the output of LLE bitstream is multiplexed with the core AAC bit-stream to produce the final lossless bit-stream.

3. IMPROVE SLS WITH PERCEPTUALLY ENHANCED BIT-PLANE CODING

3.1. Perceptually Enhanced Bit-Plane Coding

Consider an input n-dimensional data vector $\mathbf{x} = \{x_0, x_1, \ldots x_{n-1}\}$ for which each element $x_i, i = 0, \ldots, N-1$ is extracted from an independent and identically distributed (i.i.d.) random source of some alphabet $A \subset \Re$. It can be seen that x_i can be represented in a binary format

$$x_i = (2s_i - 1) \cdot \sum_{j=-\infty}^{\infty} b_{i,j} \cdot 2^j \quad (2)$$
$$i = 0, \ldots, n-1$$

that comprises a sign symbol

$$s_i \triangleq \begin{cases} 1 & x_i \geq 0 \\ 0 & x_i < 0 \end{cases},$$

and the bit-plane symbols $b_{i,j} \in \{0, 1\}$. In practice, the bit-plane symbols ususally starts from a maximum bit-plane M that satisfies

$$2^{M-1} \leq \max\{|x_i|\} < 2^M, i = 0, \ldots, n-1.$$

In Bit-Plane Code (BPC), the input data vector is first scanned into sign and bit-plane symbols, usually from MSB to LSB. The resultant binary string is then entropy coded with a properly assigned statistical model [5]. In the decoder, the data flow is reversed where the sign and amplitude symbols are decoded to reconstruct the original data vectors. Clearly, the compressed bitstream result from BPC can be arbitrarily truncated to lower rates which still can be decoded to a coarse reconstruction that comprises partial bit-plane symbols. In such a way, BPC provides

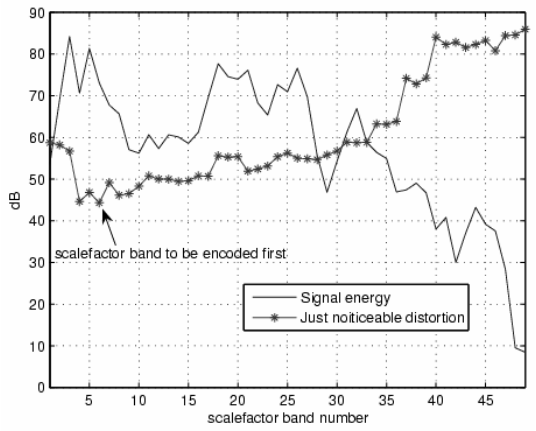

Figure 2: Signal energy vs. Just Noticeable Distortion.

a very convenient way to implement an embedded code with sequentially refined quantization step size

$$\Delta_i = 2^j, j = M-1, M-2, \ldots.$$

In addition, if the original data vectors is an integer one, lossless reconstruction is also possible if all the bit-plane symbols from MSB to LSB are reconstructed.

In most cases (unless for source with very skewed probability distribution), the MSB to LSB bit-plane scanning process provide a step-wise optimal approach in terms of mean square error (MSE). However, in most cases we are more interested in the problem where the distortion function is given by weighted MSE as:

$$d(\mathbf{x}_n, \hat{\mathbf{x}}_n) = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \hat{x}_i)^2 w_i(x_i), \quad (3)$$

where $\mathbf{x}_n$ and $\hat{\mathbf{x}}_n$ are the original data vector and the reconstructed data vector respectively, and the weighting function $w_i(x_i)$ reflects the importance of x_i with respect to the value of x_i and its location in the data vector. Example of this kind of weighted distortion can be found in the context of perceptual audio coding [6] where the transform audio data are coded according to the level of the so-called *Just Noticeable Distortion* (JND). As an illustration, Figure 2 plots signal energy and the JND for a piece of audio frame (stereo, 48kHz sampled, 1024 samples) as functions of frequencies, where lower JND level stands for easily perceptible quantization noise and hence finer representation of the signal is required. In this case the distortion function is given by the masking threshold function T_i as:

$$w_i(x_i) \propto \frac{1}{T_i}, i = 1, \ldots, n \quad (4)$$

where x_i is the tranformed audio signal, and i stands for the frequency location.

It can be observed that the above perceptually weighted dis-

tortion function (3) may be re-written as follows

$$\begin{aligned} d(\mathbf{x}_n, \hat{\mathbf{x}}_n) &= \frac{1}{n}\sum_{i=0}^{n-1}(x_i - \hat{x}_i)^2 w_i(x_i) \\ &= \frac{1}{n}\sum_{i=0}^{n-1}(\sqrt{w_i(x_i)}\cdot x_i - \sqrt{w_i(x_i)}\cdot \hat{x}_i)^2 \\ &= \frac{1}{n}\sum_{i=0}^{n-1}(x'_i - \hat{x}'_i)^2 \end{aligned} \quad (5)$$

where

$$x'_i \triangleq \sqrt{w_i(x_i)}\cdot x_i, i = 0, ..., n-1. \quad (6)$$

Hence the weighted MSE function now becomes the MSE function over data vector $\mathbf{x}'_n$. Therefore, optimal coding of $\mathbf{x}_n$ in terms of weighted MSE can be achieved by simply performing sequential bit-plane coding on the scaled domain $\mathbf{x}'_n$. In the decoder, the decoded data vector $\hat{\mathbf{x}}'$ can be de-scaled to obtain a reconstructed data vector $\hat{\mathbf{x}}$ as follows

$$\hat{x}_i = \frac{1}{\sqrt{w_i(x_i)}}\cdot \hat{x}'_i, i = 0, ..., n-1. \quad (7)$$

The problem can be further simplified if the weighting function is quantized into an even integer power of 2

$$\hat{w_i}(x_i) = 2^{2\tau_i}, i = 0, ..., n-1, \quad (8)$$

where

$$\tau_i = round\left[\frac{1}{2}\log_2\left(w_i(x_i)\right)\right], i = 0, ..., n-1, \quad (9)$$

and $round\,[\bullet]$ is the rounding to nearest integer function. In this way, the scaled data vector can be obtained by performing bit-shifting on original data vector accroding to τ_i as follows:

$$x'_i = 2^{\tau_i} x_i, i = 0, ..., n-1. \quad (10)$$

Clearly, the above simplification not only reduces the implementation complexity but also helps to maintain the compatibility to lossless coding for integer inputs.

3.2. Integration with SLS

The *Perceptually Enhanced* BPC (PE-BPC) described in previous section is integrated into MPEG-4 SLS to further improve its percpeutal performance at lossy rates. In the encoder (Figure 3), the residual spectral data from the error mapping process are firstly bit-plane shifted according to the JND calculated from the psychoacoustic model, which are then sequentially scanned and coded for all the sfbs. The total amounts of the bit-plane shifting for each sfb are also sent to the decoder for correctly decoding. In the decoder, the bit-plane symbols are decoded sequentially and shifted in the reverse manner to reconstruct the original spectral data. Perceptually optimized coding is achieved as the resulted quantization noise is shaped accroding to the JND in this process at lossy rates. In addition, lossless reconstruction is obtained if all the bit-plane symbols are decoded. The detailed bit-plane scanning method is illustrated in Figure 4.

It can be further understood that, due to the noise shaping scheme in the core AAC coder, the level of IntMDCT error coefficients $e[k]$ would tend to be flat with respect to the JND level for implicit bands. This fact can be utilized by letting all the implicit bands to share one common shifting level τ_{common}. Thus, only one parameter will be transferred for all implicit bands.

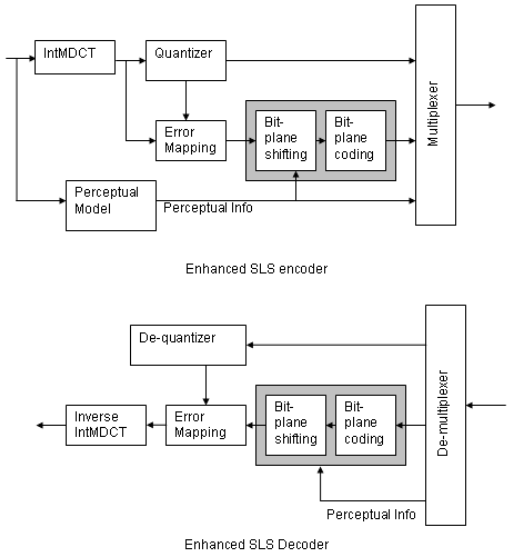

Figure 3: Integration of PE-BPC with SLS

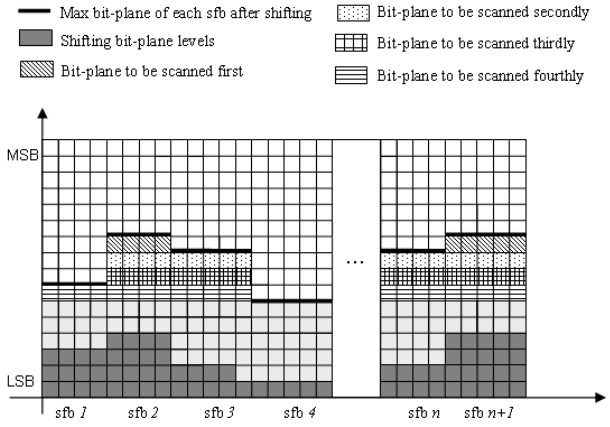

Figure 4: Perceptual Enhanced Bit-Plane Coding in SLS.

4. PERFORMANCE

We compared the perceptual quality of the enhanced SLS with PE-BPC with those of the original SLS and single rate AAC at various intermediate rates by using *noise-to-mask ratio* (NMR) measurement. Compared with SNR, it is well-known NMR is more "perceptually meaningful" as it compares the noise energy with respect to the psychoacoustic mask instead of the signal energy. In addition, as the enhanced SLS uses the same psychoacoustic model as those in the SLS and AAC codecs in our evaluation, the NMR performance difference between them reflects the difference of their ability to adjust the noise level according to a given masking threshold function, which in turn is exactly what the proposed algorithm targets at. In our evaluation, we used the standard MPEG-4 audio test sequences [7], which include 12 stereo music files sampled at 48 kHz, 16 bits/sample. The results are given in Figure 5 for total bit-rate of 128 kbps and Figure 6 for

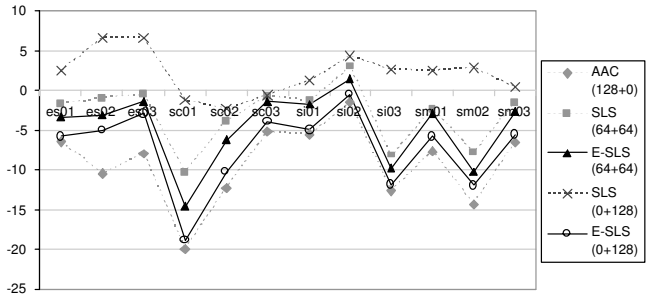

Figure 5: NMR results for AAC, SLS and enhanced SLS at 128 kbps.

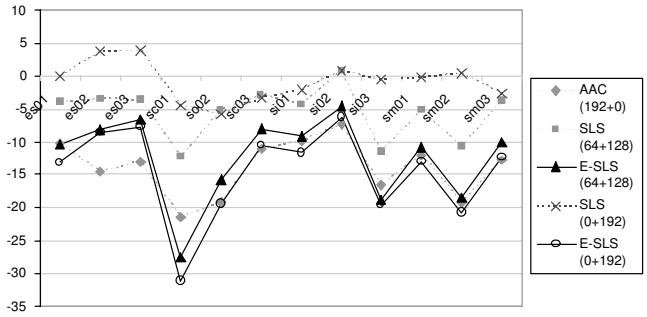

Figure 6: NMR results for AAC, SLS and enhanced SLS at 192 kbps.

total bit-rate of 192 kbps. In these figures, the bitrates for SLS and enhanced SLS are shown in the format of $(m + n)$, which means AAC core encoder is working at m kbps and the bit-rate for the lossless enhancement layer is n kpbs. Here the bit-rate of the lossless enhancement layer includes the rates for the (truncated) bit-plane coded IntMDCT spectral data as well as necessary side informations such as the bit-plane shifting amount as described in previous section.

From these results, it can be seen that in general, enhanced SLS achieves significantly better NMR performance compared with SLS at the same bit-rate combinations in particular for non-core mode operation where the AAC core bit-rate is zero. This is easily understandable since the original SLS totally relies on the AAC core encoder for noise shaping. It is also worth pointing out that for most of the testing sequences the performance of enhanced SLS at non-core mode operation is very close to that of AAC in the 128 kbps test, and even outperforms it in the 192 kbps test, suggesting that enhanced SLS non-core mode could potentially be used as a low-complexity alternative for AAC for perceptual audio coding at these bit-rates. It can also be observed that for enhanced SLS, the presense of an AAC core in fact degrades the perceptual quality compared with non-core mode operation which is clearly due to the overhead of embedding such a bit-stream. However, for use cases where perceptual quality at low rates is of primary concern, SLS with an AAC core is still a preferable choice as it still achieves better perceptual quality at low rates.

The lossless compression performance of the enhanced SLS can be found in Table 1. It can be observed that the enhanced SLS only adds negligible amount of extra overhead into the bitstream.

Table 1: Lossless compression performance of enhanced SLS and SLS.

Items (.wav)	enhanced SLS (AAC core @ 64kb/s/ch)	SLS (AAC core @ 64kb/s/ch)	enhanced SLS (non-core)	SLS (non-core)
avemaria	2.47	2.49	2.56	2.57
clarinet	2.02	2.03	2.09	2.09
cymbal	3.04	3.08	3.21	3.23
etude	2.30	2.31	2.39	2.39
flute	2.39	2.39	2.47	2.46
haffner	1.74	1.74	1.79	1.79
violin	1.98	1.98	2.05	2.04

5. CONCLUSIONS

In this paper, we propose improving the perceptual quality of SLS at intermediate bit-rates by using a new bit-plane coding method, namely, PE-BPC. PE-BPC uses a very straightforward manner to optimize the quality of the bit-plane decoded signal according to a given weighted distortion measure rather than plain MSE as in the plain BPC. This is achieved without introducing significant overhead in terms of computational complexity or lossless coding efficiency. Experimental results show that PE-BPC significantly improves the perceptual quality of MPEG-4 SLS at its scalable coding range. In addition, it also enables SLS to provide reasonable perceptual quality at 128 kbps and above for stereo signal without the need of embedding an AAC core bit-stream.

6. REFERENCES

[1] Rongshan Yu, Ralf Geiger, Susanto Rahardja, Juergen Herre, Xiao Lin, and Haibin Huang, "Mpeg-4 scalable to lossless audio coding," *117th AES Convention Preprint 6183*, 2004.

[2] "Information technology - coding of audiovisual objects, part 3. audio, subpart 4 time/frequency coding," ISO/IEC 14496-3, 1998.

[3] R. Geiger, T. Sporer, J. Koller, and K. Brandenburg, "Audio coding based on integer transform," *111th AES Convention Preprint 5471*, 2001.

[4] "Information technology - coding of audio-visual objects - part 3: Audio, amendment 3: Scalable lossless coding (SLS)," ISO/IEC 14496-3:200X 3RD EDITION/FDAM 3:2005(E), 1998.

[5] R. Yu, C.C. Ko, S. Rahardja, and X. Lin, "Bit-plane golomb code for sources with laplacian distributions," *Proceedings of ICASSP 2003*.

[6] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–513, 2000.

[7] ISO/IEC JTC1/SC29/WG11 MPEG2001/N3793, "Call for proposals for new tools for audio coding," Jan 2001.