A NOVEL SCALABLE AUDIO CODING SCHEME

Huan Zhou, Haiyan Shu, Rongshan Yu, Haibin Huang and Susanto Rahardja

Signal Processing Department
Institute for InfoComm Research
Singapore

ABSTRACT

A new scalable audio coding scheme is introduced in this paper. Its core idea is to create one additional scalability dimension during the encoding process for the purpose of generating a plural of scalable sub-bitstreams. Based on the multiple sub-streams, a smart truncator is designed that can truncate these sub-bitstreams with optimal rate-distortion (R-D) trade-off. Benefited from the flexible R-D trade-off, the proposed new scheme could, within a wide bitrate range, outperform those traditional scalable coding schemes, which usually provides a fixed R-D relationship designed at a specified bitrate. To verify the performance, the proposed scheme is further implemented based on a prior art scalable audio codec. Significant quality improvement is observed from the new codec via a series of subjective listening tests.

Index Terms— scalable coding, transcoding, truncator, coding priority

1. INTRODUCTION

Scalable audio coding is well-known preferred in real time streaming applications because of its flexibility in bitrate adaptation. However, this flexibility also brings quality degradation problem that stems from the transcoding process (the procedure to provide ending users with different decoding bitstreams from the scalable bitstream). To reveal the problem, firstly, two state-of-the-art scalable audio coding schemes are analyzed as follows.

The 1st scheme is BSAC (Bit-Sliced Arithmetic Coding), one MPEG standardized scalable audio coding scheme [1], its scalable bitstream has layered structure, consisting of one base layer (16 kbps) and several enhanced layers (up to 64 kbps). Its transcoding process is to combine the base layer (including the most important information) with additional enhanced layer (including less important information) to meet the network bitrate.

The 2nd scheme is SLS (Scalable to Lossless Coding), another MPEG standardized scalable audio coding scheme [2], its scalable bitstream is structured as a whole, where the encoding information is prioritized from the beginning to the ending of the scalable bitstream. Its transcoding process is to truncate the lossless bitstream from the end of it, and discard the less important low bit-plane coding information.

Note that both of above scalable coding schemes share one common feature: the coding priority order in their scalable bitstream is fixed in the encoding process and shared in all possible transcoding processes at different network bitrates. From R-D point of view, it means that if this scalable bitstream is optimized for one target bitrate, only the decoding bitstream with the same bitrate can provide optimized high quality output, the other decoding bitstreams with different bitrates have to, more or less, suffer quality loss.

Inspired by the above observation, a new scalable audio coding scheme is proposed that can diminish such quality loss efficiently. For simplicity, the discussion in this paper is particularly relevant to the SLS-type scalable coding scheme, although the same idea can also be applied in the BSAC-type coding scheme.

The paper is organized as follows. First the new scalable coding scheme is introduced. Second, an illustrative scalable coding system is described which incorporates the new scheme in a SLS system. Finally, experimental results are presented to illustrate the performance of the new system.

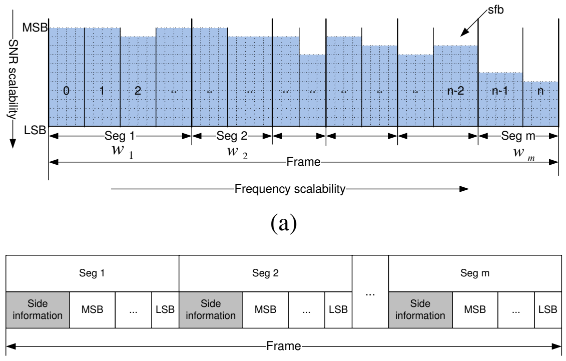
2. PROPOSED SCALABLE CODING SCHEME

The core idea of the proposed scheme is to create one additional dimensional scalability during the encoding process so that it is possible to generate flexible coding priority order at the truncator. One possible approach to realize it is to divide a frame spectrum into a plural of segments along the frequency direction and generate a plural of scalable sub-bitstreams, with each sub-bitstream associated with one subset of the frame spectral information. By introducing the plural sub-bitstreams, it becomes possible to design a truncator which can, with optimal R-D design, truncate those sub-bitstreams smartly to achieve a good trade-off between the sound quality and given network bandwidth.

Such a new scalable coding scheme is illustrated in Fig. 1 (a), where one audio spectrum is divided into m segments with each one covering a specified frequency range, e.g., including a couple of scale-factor bands (SFB). All m segments

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

are parallel encoded, possibly with different segmental coding priority orders for high quality perceptual coding. The parallel coding generates m sub-bitstreams, which are independently packed from LF sub-bitstream to HF sub-bitstream, to construct the final scalable bitstream. Again, like the original SLS bitstream, each sub-bitstream is structured in an order of segmental side information, segmental data information (from perceptually most important information to less important information, e.g., MSB to LSB) (as illustrated in Fig. 1(b)).



(a)

(b)

Fig. 1. Proposed coding scheme. (a) Frequency segments; (b) Bit stream structure.

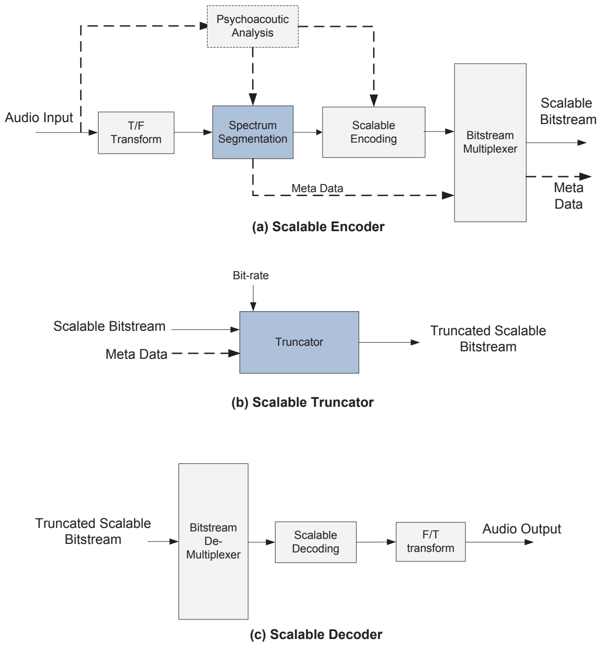
Based on the above design, the subsequent truncator has the freedom to truncate those sub-bitstreams at different truncation points. By taking into account the segmental R-D properties, a series of globally optimized truncation points could be located that is supposed to provide the best possible sound quality.

Take note that although the above segmental truncation information needs to be transmitted to the decoder as additional side information, its amount is generally ignorable. For example, if the segmental coding is based on bit-plane coding, like SLS, the truncation information could be conveniently expressed as the number of encoded bit-plane layers, and the number can be further converted to bytes representation for more economic bit assumption.

Comparing to the other traditional schemes, the above proposed scheme has two advantages on the final streaming quality. One is that, with a segment-based encoding, the original frame side information is divided into a plural of shorter segmental side information, which are dispatched to different positions in the overall scalable bitstream, therefore, much less side information is left at the beginning of the scalable bitstream. The other is that, with a smart truncating, a better overall coding priority order can be chosen by combining those sub-bitstreams with different lengths. Therefore, it can be generally envisioned that, the new scheme could outperform those traditional scalable coding schemes by providing higher quality output, especially at low to middle bitrate scenarios, where the overall coding priority order is more crucial to sound quality.

3. ENHANCED SLS SYSTEM IMPLEMENTATION

To experimentally testify whether the proposed new scheme can provide the quality improvement as what expected, the scheme is implemented on the basis of SLS system. The resulting system is called Enhanced SLS (ESLS) in this paper. A general framework of ESLS is illustrated in Fig. 2 with detailed description as follows.



(a) Scalable Encoder

(b) Scalable Truncator

(c) Scalable Decoder

Fig. 2. The Proposed Scalable Audio Codec Structure

3.1. ESLS encoder implementation

As shown in the Fig. 2(a), the ESLS encoder firstly converted an audio input into its frequency presentation (e.g., by integer MDCT transform [2]). Meanwhile, an optional psychoacoustic analysis is optionally performed on the original input. The transformed signal spectrum is subsequently divided into several segments, possibly with the fixed segmentation boundaries to avoid additional side information.

All spectral segments are then independently encoded. The segmental encoding process can be operated alone, like the bit-plane coding adopted in SLS (with sequential coding order from low frequency SFB to high frequency SFB); or guided by some external information, like the perceptually enhanced bit-plane coding [3] (with leaping order from SFB with higher PSNR (perceptual noise-to-mask ratio) to SFB with lower PSNR). The resulting segmental sub-bitstream is arranged in the order of possible side information (like maximum bit-plane information, perceptual information), possible meta data (like segmental boundary information) and the bit-plane coding data.

Lastly, the bitstream multiplexer packs those independent sub-bitstreams from corresponding low frequency segment to high frequency segment to construct the overall ESLS bitstream.

3.2. ESLS truncator implementation

When the ESLS truncator receives a streaming requirement, according to the available network bandwidth, it calculates the optimal sub-bitstream truncation points then accordingly truncate ESLS bitstream so that the length of truncated bitstream is close to bit budget for one frame coding (as illustrated in Fig. 2(b)).

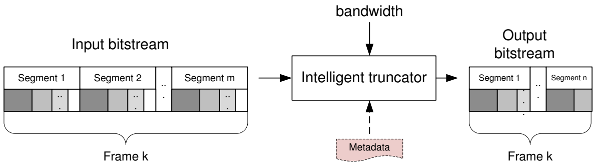


Fig. 3. Truncation process.

Targeted at the sound quality, a R-D based optimal truncation approach is proposed herein (with further illustration in Fig. 3). It includes the following 3 steps:

- *Step 1: to decide bandwidth boundary under a given bitrate* γ. It is usually decided by rule of thumb, assuming there are n segments (segment set $\Theta = \{S_1, S_2, \ldots, S_n\}$) are chosen from the m segments with $n < m$.

- *Step 2: to establish R-D tables for all involved* n *segments*. For each segment $S_l \in \Theta$, establish a R-D table composed of a series R-D data pairs. The 'R' refers to the bit consumption number, which can be easily extracted by counting the sub-bitstream length after each bit-plane coding and recorded as a non-decreasing sequence $\mathbb{R}_l = \{R_l^1, R_l^2, \ldots, R_l^{v_l}\}$ (assuming the segment is losslessly coded with v_l bit plane coding). The 'D' refers to the perceptual distortion, or the mean NMR of the segment, recorded as another non-increasing sequence $\mathbb{D}_l = \{D_l^1, D_l^2, \ldots, D_l^{v_l}\}$. The sequence pair $\{\mathbb{R}_l, \mathbb{D}_l\}$ describes the R-D feature of the l^{th} segment.

- *Step 3: to perform joint R-D optimization to decide the optimal truncation points for all* n *sub-bitstreams*. The joint R-D optimization problem can be expressed as:

$$\min_{k_l} \sum_{l=1}^{n} D_l^{k_l} \tag{1}$$

subject to

$$\sum_{l=1}^{n} R_l^{k_l} \quad \leq \quad R \tag{2}$$

$$1 < k_l \quad < \quad v_l \tag{3}$$

for $l = 1, 2, \cdots, n$, simultaneously, where R in Eq.(2) refers to the frame bit budget with $R \propto \gamma$.

Table 1. Listening test conditions

Subjects No.	9
Headphone	STAX SRM-717
Systems under Test	1. SLS non-core scheme 2.Proposed new scheme 3.Hidden reference 1 (full bandwidth) 4.Hidden reference 2 (3.5kHz bandwidth)
Test Items	5 audio items (mono)
Test bitrate	32kbps/ 48kbps/64kbps/96kbps

It usually needs high computational complexity to solve above joint optimization problem. In our implementation, a binary integer programming method is adopted for low complexity.

In the end of the truncation, the above truncating information, remaining segment number n and sub-bitstream truncation points k_l (for $l = 1, \cdots, n$) are packed (at the very beginning of truncated scalable bit stream) and transmitted to the decoder.

3.3. ESLS decoder implementation

As shown in Fig. 2(c), on the decoding side, after demultiplexer, each received sub-bitstream is independently decoded. In addition, some post-processing techniques, like fill-element and noise-filling, are adopted to avoid spectral holes. The final outputs are combined in frequency domain before converting back to time domain.

4. SUBJECTIVE LISTENING TEST

4.1. Test setup

To verify the effects of the above ESLS system, a series of subjective listening tests has been carried out based on the MUSHRA method [4]. The test conditions and test samples are listed in Table 1 and Table 2, respectively.

Table 2. Test material in the listening test

Test item	Content description	Duration (s)
es03	English speech	7.52
sc03	Contemporary pop music	11.47
si01	Harpsichord	7.92
si02	Castanets	7.64
sm02	Glockenspiel	10.01

4.2. Test results

The subjective tests were performed by 9 experienced listeners. Their raw results are post-processed via statistical analysis. The final results are presented from Fig. 4 to Fig. 7 for different testing bitrates, where the performances of two different coding schemes are measured in way of statistical means (the top end of score bars), compared with the hidden reference (original item) and the bandwidth-limited anchors (3.5kHz low-passed item). And each statistical mean has an associated 95% confidence interval (the line segment surrounding the top of bars).

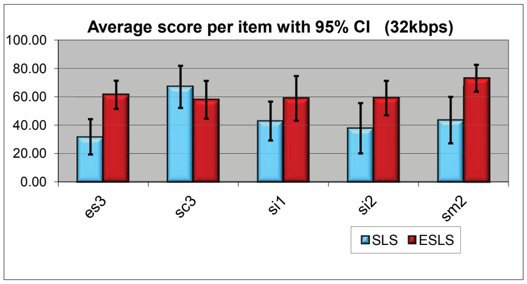


Fig. 4. The average score per item @32kbps

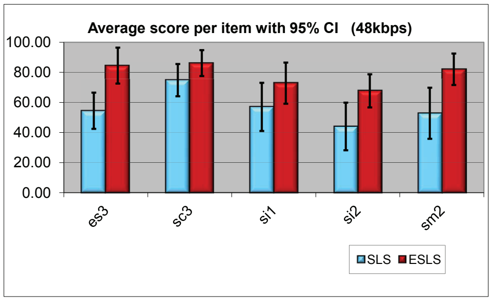


Fig. 5. The average score per item @48kbps

As can be seen from the above comparisons, the proposed ESLS system clearly improves the perceptual sound quality by a noticeable margin statistically at the low to middle bitrate range (i.e. 32 $\sim$ 64 kbits/sec). Even at comparatively high bitrate, say 96 kbits/sec, the proposed scheme is still statistically better than SLS. Such results are consistent to the theoretical analysis in Sec. 2.

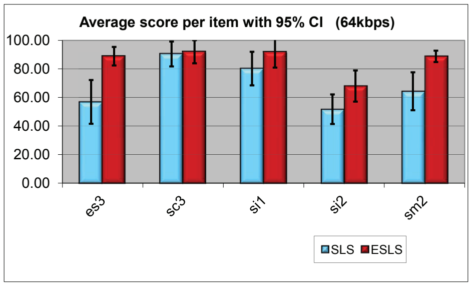


Fig. 6. The average score per item @64kbps

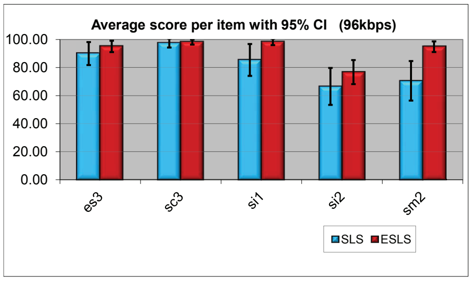


Fig. 7. The average score per item @96kbps

Table 3. Overall performance of two coding systems

Testing bitrate	SLS	ESLS
32kbps	44.6	62.1
48kbps	56.6	78.6
64kbps	68.7	85.9
96kbps	79.6	92.7

Finally, to gain an overall view about the performance of proposed scheme, the above comparison results are statistically processed to remove the their dependency on testing items. The processing results are listed in table 3 which shows a clear quality improvement trend over a wide bitrate range.

5. CONCLUSION

This paper presents a new scalable audio coding scheme. Essentially, this scheme proposes three new technologies to be embedded into the SLS Non-Core. They are: spectrum segmentation, adaptive scanning order and R-D based segmentation truncation. The quality improvement from the new scheme has been confirmed by a series of subjective listening tests.

6. REFERENCES

[1] ISO/IEC 14496-3:2009, "Information technology – Coding of audio-visual objects – Part 3: Audio," Oct. 2009.

[2] ISO/IEC 14496-3:2005/Amd 3:2006, "Scalable Lossless Coding (SLS)," June 2006.

[3] Rongshan Yu, Te Li, and Susanto Rahardja, "Perceptually enahnced bit-plane coding for scalable audio," *ICME*, pp. 1153–1156, 2006.

[4] European Broadcasting Union, "MUSHRA-EBU method for subjective listening tests of intermediate audio quality," *Draft Technical Recommendation BMC 607 (Rev.1) B/AIM 022 (Rev.9), Technical Department*, Jan,2000.