A PERCEPTUALLY ENHANCED SCALABLE-TO-LOSSLESS AUDIO CODING SCHEME AND A TRELLIS-BASED APPROACH FOR ITS OPTIMIZATION

Emmanuel Ravelli, Vinay Melkote and Kenneth Rose

Department of Electrical and Computer Engineering,
University of California, Santa Barbara, CA 93106-9560, USA
{ravelli, melkote, rose}@ece.ucsb.edu

ABSTRACT

Scalable-to-Lossless (SLS) audio compression, as standardized by MPEG, provides a lossy base layer compatible with the Advanced Audio Coding (AAC) format, ensuring state-of-the-art quality in the base layer, and additional fine grained enhancements that eventually provide a lossless compressed version of the signal. While SLS offers highly efficient lossless compression, the perceptual quality of its intermediate lossy layers has been observed to be suboptimal. This paper proposes a modified SLS audio coding scheme that provides enhanced perceptual quality at an intermediate bit-rate, at the expense of an additional parameter per frequency band as side-information. This scheme when coupled with a trellis-based optimization algorithm is demonstrated to outperform, in terms of quality at the intermediate bit-rate, both standard SLS and a recent perceptually enhanced variant, with minimal degradation in lossless coding performance.

Index Terms— Scalable audio coding, lossless audio coding, rate-distortion optimization, trellis.

1. INTRODUCTION

Scalable-to-lossless coding of audio [1] provides an integrated audio coding solution for applications such as archiving, that require lossless compression of the signal, as well as for streaming, playback etc., that generally need a low bit-rate (and hence lossy) perceptually coded version. The SLS bitstream as standardized in MPEG-4 has a base-layer of AAC coded signal, and fine grained enhancements that can be used to incrementally improve the coding quality and enable adaptation of the bitstream to application or user specific bit-rates without transcoding. This minimizes the complexity/storage requirements of digital music servers.

However, it has been observed that the perceptual quality at intermediate bit-rates, i.e., of the reconstruction obtained using the AAC base-layer and only some of the fine-grained enhancements, could be much worse than the quality of non-scalable AAC at the same cumulative bit-rate [2]. This is particularly true when the AAC base layer is of low bit-rate. To overcome this problem, perceptually-motivated variants of the SLS standard have been proposed, that improve significantly the quality of the intermediate layers in SLS. All these variants have been published in numerous papers by the same authors (e.g., [2, 3]). The implementation in [2], being the latest of these methods, can be considered state-of-the-art and will henceforth be referred to as PSLS. It is worth noting that [2] is also the only paper that provides sufficient detail for a proper re-implementation. Though motivated by the need for improved perceptual quality in intermediate layers, like SLS, the PSLS format has a rigid structure described later. In other words, there is no room for a smart allocation of the bit resource available for the intermediate layers, between different frequency bands (or scalefactor bands (SFBs), as they are refered to in AAC). Thus improved perceptual quality cannot be guaranteed or optimization in a rate-distortion sense cannot be effected.

Motivated by this observation we propose a modified SLS format that involves sending a single extra parameter for each SFB but provides flexibility in bit-rate allocation between these bands. Run-length encoding of this parameter ensures that its inclusion in the SLS bitstream has minimal impact on the lossless coding performance. A trellis-based optimization procedure is proposed that chooses this parameter, for every SFB, in a rate-distortion optimization setting and guarantees improved perceptual quality at an additional bit-rate, other than that of the base-layer. This is of advantage to applications of SLS such as the one envisioned in [4], where an online music store might be providing bitstreams for three dominant applications: a low bit-rate version (just the base layer) for playback on a device with low storage capability or low bandwidth requirements (such as a mobile phone), a version at a higher (yet lossy) intermediate bit-rate for listening on a more powerful, high fidelity system, and a lossless version for the consumer who wants to archive an exact copy.

We note here that in [3] (a precursor to [2]), a similar modified SLS format that involves sending one extra parameter per SFB is proposed. But no details on the side information coding were given nor any optimization procedure that ensures optimal quality.

2. PRIOR WORK

2.1. MPEG-4 SLS Standard

The basic SLS encoding algorithm, as described in [1], is summarized here. Note that a revised version of SLS involving more sophisticated encoding techniques (e.g., context-based arithmetic coding) is now included in the standard but for the sake of simplicity only the first version of SLS will be considered in this paper.

The encoder transforms a block of $2N$ audio samples to N integer coefficients $c[k]$ using the integer modified discrete cosine transform (IntMDCT), a reversible integer-to-integer transform that closely approximates the modified discrete cosine transform used in AAC. These coefficients are input to the AAC quantization and coding module, producing a base layer compatible with the MPEG AAC standard. Subsequently, an error mapping process calculates the residual coefficients that will be coded into the SLS

The work was supported in part by the NSF under grant CCF-0917230, the University of California MICRO Program, Applied Signal Technology Inc., Cisco Systems Inc., Qualcomm Inc., and Sony Ericsson, Inc..

978-1-4244-3679-8/09/$25.00 ©2009 IEEE

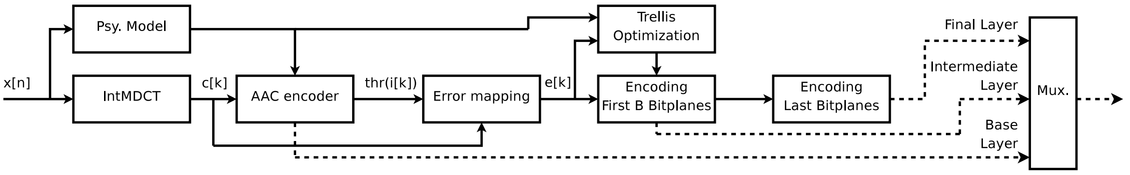

Figure 1: Block diagram of the proposed three stage scalable-to-lossless audio coder.

enhancement layers. This process is performed in each SFB by considering two cases. If all the base-layer quantized coefficients in the SFB s are zeros, the band is then referred to as an explicit band and the mapped error $e[k]$ is simply equal to the original coefficients, i.e., $e[k] = c[k]\ \forall k \in \text{SFB } s$. If not, the band is an implicit band, and the mapped error is

$$e[k] = c[k] - \text{floor}\left(thr\left(i[k]\right)\right)\ \forall k \in \text{SFB } s \qquad (1)$$

where $thr\left(i[k]\right)$ is the boundary closer to zero of the AAC quantizer cell with index $i[k]$ and is given by

$$thr(i[k]) = \text{sgn}\left(i[k]\right)\left[2^{\text{scf}(s)/4}\left|i[k] - 0.4054\right|^{4/3}\right] \qquad (2)$$

for $i[k] \neq 0$, and $thr(0) = 0$. Here, scf(s) denotes the scalefactor for the SFB.

Finally, the mapped error is encoded using Bit-Plane Golomb Coding (BPGC). The magnitude of the error coefficients in SFB s are expressed in $M(s)$-bit binary representation as $|e(k)| = \sum_{j=0}^{j=M(s)} b[k,j]2^j$ where $M(s)$ is the most significant bit (MSB)-plane in the band. The parameter $M(s)$ is deduced from the AAC quantizer cell widths in the case that SFB s is an implicit band, otherwise it is differentially encoded and sent to the decoder as side information. The bits $b[k,j]$ are then encoded using a binary arithmetic coder with the following probability assignment

$$P(b[k,j] = 1) = Q^L(j) = \begin{cases} \frac{1}{1+2^{2^{j-L}}}, & j \geq L \\ \frac{1}{2}, & j < L \end{cases} \qquad (3)$$

and L is a parameter estimated in each SFB with the following rule

$$L = \min\left\{L' \in \mathbb{Z} | 2^{L'+1}N \geq A\right\}\ . \qquad (4)$$

Here, N is the number of error coefficients in the band, A is their absolute sum and $\mathbb{Z}$ is the set of integers. We will denote by $L(s)$ the value of this parameter for SFB s. These parameters are Huffman coded and sent as side information. SLS imposes a specific order in the encoding of the bits $b[k,j]$: first the MSB-plane of all SFBs is coded and then the next bit-plane and so on. Within each bit-plane, coding proceeds from SFBs with lower frequency to higher. This process allows successive refinement of the error and thus fine-grain scalability. If the decoder receives all the bits, then lossless decoding of the error (and thus lossless decoding of the original coefficients) is achieved. Otherwise, the error is only partially decoded (e.g., only T bit-planes are available) and the decoder reconstructs a suitable approximation of the error and a corresponding reconstruction $\hat{c}_T[k]$ of the IntMDCT coefficients using an inverse error mapping process [1]. These are then inverse transformed to give the time-domain reconstruction.

2.2. A perceptually motivated variant of SLS

Ideally bit-plane coding should reduce the quantization noise in each SFB by the same factor whenever a new bit-plane is received. Thus, a distortion metric such as the noise-to-mask ratio (NMR) should improve by the same amount in each SFB enabling perceptual scalability. When the AAC core coder is working at low bit-rates the base-layer distortions in each band may be widely different and the bit-plane coding procedure in SLS cannot guarantee the best quality achievable. In other words, the constraint that a bit-plane of all SFBs has to be coded before moving on to the next bit-plane, may not be the best method to distribute the available bit resource amongst the SFBs. In [2] the authors addressed this problem by proposing PSLS, a perceptually-motivated variant of the SLS standard. PSLS consisted of changing the order in which the bits are encoded, trying to encode first the bits which contribute the most to the perceptual quality (e.g., first two bit-planes from the low-frequency bands, then two bit-planes from the mid-frequency bands and so on). Two pre-defined orders were considered, minimizing the side-information. Depending on the shape of the residual spectrum, one is chosen. While this work seems to improve significantly the quality of the intermediate layers, akin to SLS it imposes a rigid structure in the coding order and a flexible bit distribution to bands, to guarantee optimal quality, is not possible. While our proposed method draws motivation from PSLS, it provides a flexible framework that is amenable to rate-distortion optimization and hence can provide optimal quality at an intermediate bit-rate.

3. PROPOSED METHOD

3.1. Coding scheme

We propose a new three stage coding scheme based on SLS (see Fig. 1). After the IntMDCT and error mapping stages, the error $e[k]$ is encoded with BPGC, but instead of the regular SLS ordering of bits, the first $B(s)$ bit-planes in each band s are coded consecutively. The parameter $B(s)$ for each band can be now chosen to re-distribute bits to different SFBs, thus providing bits to SFBs

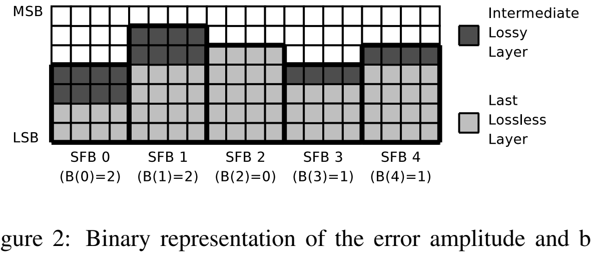

Figure 2: Binary representation of the error amplitude and bit-plane coding: the intermediate stage codes the first $B(s)$ bit-planes in each band s, the final stage codes the remainder.

that really need them. The remaining bit-planes in all the SFBs are subsequently coded in regular SLS order (see Fig. 2). Thus we have an ‘intermediate layer’ that is entirely determined by the set of parameters $B(s)$, which need to be encoded as side information. We have found that run-length encoding was appropriate for encoding these parameters. The bands are grouped into sections, with each band in a section having the same $B(s)$. Side information for each section is encoded using 8 bits - 3 for the parameter value and 5 for the section length (i.e., run-length). Experience shows that this technique produces a rate of approximately 1.5 bits per band, which is half the rate required by raw encoding with 3 bits per band. Side information bit-rate needed up to the intermediate layer is reduced by encoding the parameters $M(s)$ and $L(s)$ only in the case $B(s) \neq 0$. The parameters $M(s)$ and $L(s)$ for bands with $B(s) = 0$ are encoded post intermediate layer bits.

3.2. Rate-distortion optimization

In each frame, the intermediate layer requires a choice of the parameters $\mathbf{B} = \{B(0), \ldots, B(S-1)\}$ where S is the number of SFBs. The “best” set of parameters $\mathbf{B}^*$ is the solution of the following rate-distortion optimization problem

$$\mathbf{B}^* = \arg \min_{\mathbf{B}: R(\mathbf{B}) \leq R_t} D(\mathbf{B}) \tag{5}$$

where $D(\mathbf{B})$ is a per-frame perceptually-relevant distortion measure and $R(\mathbf{B})$ is the bit-rate of this layer. The target rate to be achieved is R_t.

The distortion measure used in the proposed method is the average NMR (ANMR),

$$D(\mathbf{B}) = \frac{1}{S} \sum_{s=0}^{S-1} \frac{d_s}{m_s} \tag{6}$$

where m_s is the masking threshold in SFB s, as given by the psychoacoustic model in AAC, and d_s is the squared error in the band.

$$d_s = \sum_{k \in \text{SFB}s} \left(c[k] - \hat{c}_{B(s)}[k] \right)^2 . \tag{7}$$

The total rate consumed by the intermediate layer is

$$R(\mathbf{B}) = \sum_{s=0}^{S-1} F\left[B(s), B(s-1)\right] + G\left[B(s)\right] + H\left[B(s)\right] \tag{8}$$

where $F\left[B(s), B(s-1)\right]$ is the number of bits required to encode $B(s)$: the run-length encoding produces 8 bits if $B(s) \neq B(s-1)$ and zero bits otherwise. $G\left[B(s)\right]$ is the number of bits required to code $L(s)$ and $M(s)$: it is zero if $B(s) = 0$. The arithmetic coding of the $B(s)$ bit-planes in SFB s contributes $H\left[B(s)\right]$ bits to the bit-stream. Assuming that the arithmetic coder is perfect the cost of encoding bit $b[k,j]$, where $k \in$ SFB s, is $-\log_2 \left[b[k,j] Q^{L(s)}(j) + (1 - b[k,j])(1 - Q^{L(s)}(j)) \right]$. Note that this approximation is, in practice, very close to the real rate in an arithmetic coded bit-stream.

Now we reformulate the problem (5) as the minimization of a Lagrangian cost given by

$$J(\mathbf{B}, \lambda) = D(\mathbf{B}) + \lambda R(\mathbf{B}) \tag{9}$$

where λ is the Lagrangian parameter. Minimization of (9) for a particular value of λ finds the set of parameters $\mathbf{B}^*(\lambda)$ that minimizes the distortion in (6) at the achieved rate $R(\mathbf{B}^*(\lambda))$. By suitably iterating over λ we can find the optimal set $\mathbf{B}^*$, that is the solution to (5). As the parameter $B(s)$ can be one of 8 choices, brute-force minimization of the cost (9) would entail a complexity as high as 8^S, i.e., exponential in the number of SFBs. Instead, we propose a fast, trellis-based algorithm, inspired by a similar method in [5] for AAC parameter selection. A trellis with S stages, as shown in Fig. 3, is constructed. Each stage corresponds to an SFB. The states in each stage represent one particular value of the parameter $B(s)$. Each stage therefore has a maximum of 8 states. Note that if $M(s) \leq 6$ the number of states in stage s will be reduced to $M(s)+2$ (corresponding to $B(s) \in \{1, \ldots, M(s)+1\}$, and an additional state $B(s) = 0$ when no bit-plane of the SFB s is coded in the intermediate layer). A state is populated with distortion associated with the corresponding $B(s)$ (through (7)) and the bit costs $G[B(s)]$ and $H[B(s)]$. Transitions are associated with costs $F[B(s), B(s-1)]$. A Vitterbi algorithm is employed to find the path through the trellis that minimizes (9).

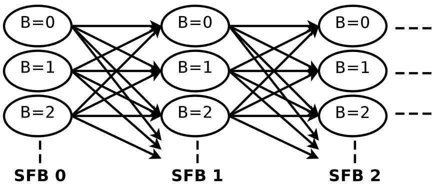

Figure 3: Trellis employed in the proposed method: each stage corresponds to a scale-factor band, and every node represents a value of the proposed parameter $B(s)$.

4. RESULTS

We compared the proposed coding scheme employing trellis-based optimization to standard SLS [1] and PSLS [2]. The coders share the same base-layer, which is an AAC-compliant layer whose parameters have been optimally selected by minimizing the ANMR distortion measure with the trellis-based method in [5]. Mono, 48kHz sampled versions of the same 15 audio files as in [1] have been used in our tests[1] .

4.1. Audio quality of the intermediate layer

The three coding schemes are first compared in terms of their ANMR values at the intermediate layer. Table. 1 compares the distortion in four scenarios: the AAC base-layer rate is 32kbps or 64kbps, and with either base-layer rate the intermediate stage target rate (R_t) is also one of 32kbps or 64kbps. Included in the comparison is a non-scalable AAC coder at the cumulative bit-rate. The proposed scheme is consistently better than SLS or PSLS, in terms of the ANMR metric.

We also provide objective measurements of the intermediate layer quality using PEAQ (ITU-R BS.1387-1), as implemented in the AFsp library [6]. PEAQ gives an objective measure called Objective Difference Grade (ODG) where -4 indicates ‘very annoying’ while 0 means the difference between original and coded versions is imperceptible. Fig. 4(a) compares the ODGs of the competing strategies for each audio sample when both the AAC base-layer and intermediate layers are coded at 32kbps. The comparison includes the non-scalable AAC coder at the cumulative 64kbps.

[1] The authors would like to thank Pierrick Philippe (France Telecom R&D) for providing us the dataset.

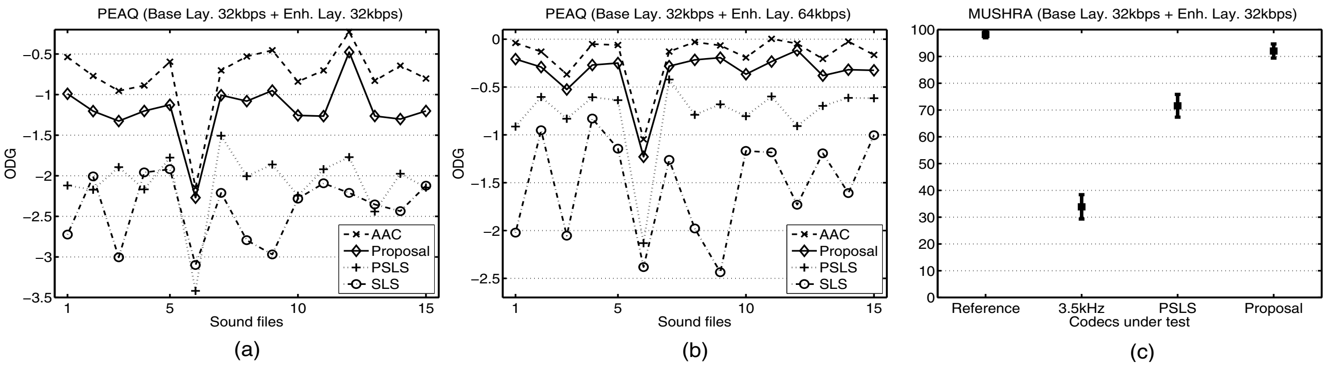

Figure 4: Audio quality of the intermediate layer: (a),(b) ODG measurements given by PEAQ - (c) MUHSRA listening test results.

	32+32	32+64	64+32	64+64
SLS	-2.18	-5.12	-6.68	-9.69
PSLS	-1.38	-5.74	-5.36	-9.76
Proposal	-3.90	-7.78	-8.09	-11.45
Non-scalable AAC	-5.05	-9.48	-9.48	-13.81

Table 1: Audio quality of the intermediate layer: ANMR (in dB) averaged over all frames and all sound files (32+64 indicates a base-layer at 32kbps and intermediate layer at 64kbps).

	SLS	**PSLS**	**Proposal** R_t**=32kbps**	**Proposal** R_t**=64kbps**
AAC base at 32kbps	2.08	2.08 (-0.01%)	2.06 (-0.93%)	2.06 (-0.99%)
AAC base at 64kbps	2.05	2.05 (-0.01%)	2.03 (-1.06%)	2.03 (-1.08%)

Table 2: Compression ratios for the set of 15 samples.

Fig. 4(b) gives similar results when the intermediate layer is instead at 64kbps. These results show that our proposed approach gives better objective audio quality in the intermediate layer.

Finally, a MUSHRA listening test has been conducted to subjectively evaluate the proposed coding scheme and compare it with PSLS. To avoid listener fatigue the material used for testing has been reduced from 15 files to the 6 most critical items. 4 versions of each test sound were evaluated: a hidden reference, a 3.5kHz low-pass anchor, PSLS, and the proposed coder (both coders use two layers at 32kbps+32kbps). The test items were presented in random order to 13 experienced listeners, and scored on a scale of 0 (bad) to 100 (excellent). The average results and the 95% confidence intervals are given in Fig. 4(c). These result confirm the objective measurements, and clearly show that our proposed approach outperforms PSLS.

4.2. Lossless compression performance

The lossless compression ratios achieved by SLS, PSLS and the proposed method are shown in Table. 2. The best compression is achieved by SLS. PSLS adds only one additional bit per frame, so the loss in compression ratio is very small (0.01%). In the proposed method, the additional cost is greater, i.e., equal to the cost of sending the parameters $B(s)$ in each frame, resulting in a loss in compression by about 1%. In a practical application of SLS this is a negligible cost as compared to the improved quality that is provided to a consumer using the intermediate layer.

Note that the trellis-based optimization was measured at approximately 5 times the real-time on a recent processor at 2GHz, which is only 4 times slower than our implementation of PSLS.

5. CONCLUSION

We have proposed a new SLS-based audio coding format that enables an intermediate layer of optimal perceptual quality in a three stage coding scheme. The coding format includes a new parameter for each SFB, whose choice when optimized using a trellis-based algorithm, shapes the quantization noise of the intermediate layer such that the perceived distortion is minimized. This parameter increases the side information but with minimum detriment to the lossless compression ratio. Objective and subjective evaluation of the intermediate layer audio quality show that the proposed method outperforms both SLS and a state-of-the-art perceptually enhanced variant of SLS.

6. REFERENCES

[1] R. Yu, S. Rahardja, L. Xiao, and C. C. Ko, “A fine granular scalable to lossless audio coder,” *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 14, no. 4, pp. 1352–1363, Jul. 2006.

[2] T. Li, S. Rahardja, and S. N. Koh, “Frequency region-based prioritized bit-plane coding for scalable audio,” *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 16, no. 1, pp. 94–105, Jan. 2008.

[3] R. Yu, T. Li, and S. Rahardja, “Perceptually enhanced bit-plane coding for scalable audio,” in *Proc. Int. Conf. Multimedia Expo.*, Jul. 2006, pp. 1153–1156.

[4] T. Li, Y. S. Liew, and S. Rahardja, “A multi-quality audio managing system for internet music store,” in *Proc. IEEE Intl. Conf. on Sig. Proc.*, 2008, pp. 2685–2688.

[5] A. Aggarwal, S. L. Regunathan, and K. Rose, “A trellis-based optimal parameter value selection for audio coding,” *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 14, no. 2, pp. 623–633, Mar. 2006.

[6] P. Kabal, Audio File Programs and Routines, http://www-mmsp.ece.mcgill.ca/Documents/Downloads/AFsp/.