JOINT OPTIMIZATION OF THE PERCEPTUAL CORE AND LOSSLESS COMPRESSION LAYERS IN SCALABLE AUDIO CODING

Emmanuel Ravelli, Vinay Melkote, Tejaswi Nanjundaswamy, and Kenneth Rose

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

ABSTRACT

MPEG-4 High-Definition Advanced Audio Coding (HD-AAC) enables scalable-to-lossless (SLS) audio coding with an Advanced Audio Coding (AAC) base layer, and fine-grained enhancements based on the MPEG SLS standard. While the AAC core offers better perceptual quality at lossy bit-rates, its inclusion has been observed to compromise the ultimate lossless compression performance as compared to the SLS 'non-core' (i.e., without an AAC base layer) codec. In contrast, the latter provides excellent lossless compression but with significantly degraded audio quality at low bit-rates. We propose a trellis-based approach to directly optimize the trade-off between the quality of the AAC core and the lossless compression performance of SLS. Simulations to test the effectiveness of the approach demonstrate the capability to adjust the trade-off to match application specific needs. Moreover, such optimization can in fact achieve an AAC core of superior perceptual quality while maintaining state-of-the-art (and surprisingly sometimes even better) lossless compression, all this in compliance with the HD-AAC standard.

Index Terms— Audio coding, lossless coding, AAC, SLS, rate-distortion optimization.

1. INTRODUCTION

MPEG-4 HD-AAC [1] is a recent standard for scalable-to-lossless audio coding that combines a lossy base layer of AAC [2] with fine-grained enhancements via MPEG-4 SLS [3]. While the AAC core ensures state-of-the-art perceptual quality and backward compatibility with legacy decoders, the SLS layers offers the capability to achieve bit-exact reconstruction. The *lossless* compression performance of HD-AAC is shown in recent evaluations (e.g., [4]) to underperform the SLS 'non-core' (NC) codec (i.e., a standalone SLS codec without an AAC base layer). However, the lack of a perceptually coded base layer in the SLS NC bitstream results in poor perceptual quality at intermediate (lossy) bit-rates. Thus, a compromise between perceptual quality and lossless compression seems inevitable in most practical applications.

In HD-AAC, the residual signal after base-layer coding, and in turn its lossless compression by SLS, is largely determined by the choice of AAC encoding parameters. Despite this fact, prior work on HD-AAC has not considered optimizing the AAC encoder, while explicitly accounting for its effects on subsequent SLS encoding. Motivated by this observation, we propose a novel, joint optimization of coding parameters for the AAC core and SLS enhancements within the HD-AAC bitstream. A trellis-based algorithm chooses the encoding parameters to optimally control the trade-off between perceptual (and lossy) coding performance at the AAC core, and overall lossless compression by SLS. Simulations provide evidence that careful optimization achieves "the best of both worlds", namely, an AAC core of very good perceptual quality, and lossless coding performance comparable to (and in some cases surpassing) that of the SLS NC codec. We emphasize that the optimization is applied to the encoder decisions and generates a standard-compliant bitstream.

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

2. THE MPEG-4 HD-AAC STANDARD

This section introduces notation and gives a brief description of the two standards, MPEG-4 AAC [2], and MPEG-4 SLS [3], which have been combined in HD-AAC (Fig.1).

2.1. MPEG-4 AAC

The AAC encoder segments the audio signal into 50% overlapped frames of $2K$ samples each ($K = 1024$). Let an audio file consist of N such frames. A modified discrete cosine transform (MDCT) is employed to produce K transform coefficients per frame, which are subsequently grouped into frequency bands referred to as scalefactor bands (SFBs). We denote by $c_n[k],\ 0 \le n < N,\ 0 \le k < K$ the k^{th} transform coefficient of frame n. All the coefficients in an SFB of a frame are quantized using the same quantizer - a generic AAC quantizer scaled by a parameter called the scalefactor (SF), and subsequently encoded with the same Huffman codebook (HCB). Thus, in addition to the quantized and entropy coded transform coefficients, the bitstream for each AAC frame contains side information specifying the SF and HCB, where the former is differentially encoded and the latter run-length encoded, across SFBs. The SF, s_n^l, and HCB, h_n^l, for each SFB l ($0 \le l < L$) are selected from a standard-specified set. We denote by $\mathbf{p}_n = (\mathbf{s}_n, \mathbf{h}_n)$ the encoding parameters for frame n, with $\mathbf{s}_n = \{s_n^0, \ldots, s_n^{L-1}\}$ and $\mathbf{h}_n = \{h_n^0, \ldots, h_n^{L-1}\}$. The AAC encoding parameters for the entire signal (or audio file) are subsumed in the compact notation $\mathcal{P} = \{\mathbf{p}_0, \ldots, \mathbf{p}_{N-1}\}$. Note that for simplicity we will exclude other optional tools available in AAC, such as block switching, temporal noise shaping, etc.

Let $\mathcal{R}_b(\mathcal{P})$ denote the base layer bit-rate needed to encode the signal with parameters $\mathcal{P}$, and let $\mathcal{D}_b(\mathcal{P})$ denote the corresponding distortion. The AAC encoder's objective may either be expressed as a rate-constrained problem

$$\mathcal{P}^* = \arg \min_{\mathcal{P}:\mathcal{R}_b(\mathcal{P}) \le \mathcal{R}_t} \mathcal{D}_b(\mathcal{P}) \,, \tag{1}$$

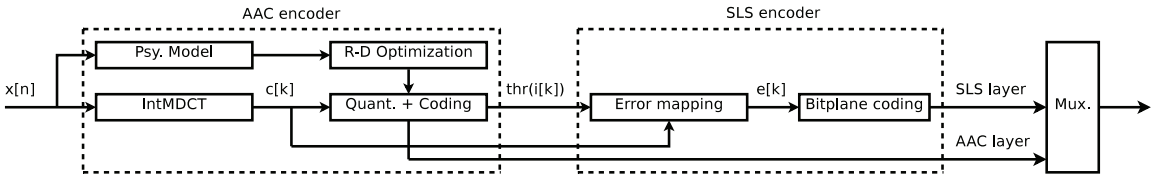

Fig. 1. Block diagram of a MPEG-4 HD-AAC encoder.

or as a distortion-constrained problem

$$\mathcal{P}^* = \arg \min_{\mathcal{P}:\mathcal{D}_b(\mathcal{P})\leq\mathcal{D}_t} \mathcal{R}_b(\mathcal{P}) \,, \tag{2}$$

where $\mathcal{R}_t$ and $\mathcal{D}_t$ denote the target rate and distortion, respectively. The distortion $\mathcal{D}_b(\mathcal{P})$ is typically based on the Noise-to-Mask Ratio (NMR), which is calculated for each SFB as the ratio of quantization noise to a noise masking threshold provided by a psychoacoustic model. The NMR or distortion of the SFB depends on the SF value and will be denoted as $d_n^l(s_n^l)$. The overall distortion $\mathcal{D}_b(\mathcal{P})$ may then be calculated by averaging or maximizing over SFBs and/or frames. Typical examples are the Average-Average NMR (AANMR) given by

$$\mathcal{D}_b^A(\mathcal{P}) = \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{L}\sum_{l=0}^{L-1} d_n^l(s_n^l) \,, \tag{3}$$

or the Maximum-Maximum NMR (MMNMR),

$$\mathcal{D}_b^M(\mathcal{P}) = \max_{0\leq n<N}\max_{0\leq l<L} d_n^l(s_n^l). \tag{4}$$

Given the huge parameter space, encoders generally employ a truncated search and obtain a corresponding sub-optimal $\mathcal{P}$. A frequently used approach is to process each frame independently and employ a sub-optimal method (e.g., the two-loop search (TLS) [2]) to find the frame-specific parameters. In contrast, a trellis-based algorithm is employed in [5] to solve the rate-constrained problem optimally for individual frames, and in [6] this trellis approach was extended to jointly optimize all frames (1). These trellis-based approaches serve as building blocks for the method proposed herein.

2.2. MPEG-4 SLS

The SLS layer provides fine-grained enhancements all the way to lossless reconstruction. The lossless compression feature necessitates the use of a reversible integer transform known as IntMDCT in the base layer encoder, that closely approximates regular MDCT. The integer coefficients are input to the AAC quantization and coding module as usual, producing a base layer compatible with the AAC standard. Subsequently, an error mapping process calculates the residual coefficients that will be coded into the SLS enhancement layer. If all the base layer quantized coefficients in an SFB are zeros, the band is referred to as an explicit band and the mapped error $e_n[k]$ is simply equal to the original coefficients, i.e., $e_n[k] = c_n[k]$. If not, the band is said to be an implicit band, and the mapped error is

$$e_n[k] = c_n[k] - \text{floor}\left(thr\left(i_n[k]\right)\right) \tag{5}$$

where $i_n[k]$ denotes the base layer quantization index for the coefficient and $thr\left(i_n[k]\right)$ is the boundary closer to zero of the AAC quantizer cell with index $i_n[k]$.

$$thr(i_n[k]) = \text{sgn}\left(i_n[k]\right)\left(2^{s_n^l/4}\left|i_n[k] - 0.4054\right|^{4/3}\right) \tag{6}$$

for $i_n[k] \neq 0$, and $thr(0) = 0$ (in the above $k \in SFB\ l$). Finally, the mapped error is encoded using either Bit-Plane Golomb Codes (BPGC) or Low Energy Mode Codes (LEMC) [3], both of which are bit-plane arithmetic codes (Context-based Arithmetic Coding, another standardized bit-plane code [3], is not permitted in HD-AAC [1]). The magnitude of the error coefficients in SFB l are expressed in $\mathcal{M}_n^l$-bit binary representation as $|e_n[k]| = \sum_{j=0}^{j=\mathcal{M}_n^l} b[k,j]2^j$, where $\mathcal{M}_n^l$ is the most significant bit (MSB)-plane in SFB l. The parameter $\mathcal{M}_n^l$ is deduced from the AAC quantizer cell widths in the case that SFB l is an implicit band, otherwise it is differentially encoded and sent to the decoder as side information. Note that the mapped error $e_n[k]$ and the MSB-planes $\mathcal{M}_n^l$ are determined by the AAC core. The only free encoding parameters in the SLS enhancement layer are the so called 'lazy bit-plane' parameters $\mathcal{L}_n^l$ [7] that characterize the binary distribution of bits in each SFB, and control the arithmetic coding operation in the bit-plane coders. The bits $b[k,j]$ in each SFB are encoded using BPGC in the case $\mathcal{M}_n^l - \mathcal{L}_n^l > 0$, or with LEMC otherwise. The parameter $\mathcal{L}_n^l$ for each SFB is encoded using a Huffman code, and can be derived using a suitable algorithm (e.g., as suggested in [7]). Alternatively, since the standard restricts $\mathcal{L}_n^l$ to take one of 3 values for each SFB [3], even an exhaustive search to find the $\mathcal{L}_n^l$ that minimizes the enhancement layer bit-rate is not computationally prohibitive. We denote by $\mathcal{R}_e(\mathcal{P})$ this minimized enhancement layer bit-rate, when the base layer parameters are fixed at $\mathcal{P}$.

3. JOINT OPTIMIZATION OF AAC AND SLS

As already explained, current HD-AAC encoders are "myopic": the AAC parameters are chosen (either by solving (1) or (2) exactly, or more often by a sub-optimal method), but irrespective of their impact on the SLS performance. Subsequently the SLS enhancement layer is encoded given the base layer residue. In contrast, we propose an approach where the cost of choosing a particular set of AAC encoding parameters includes the effect these parameters have on the subsequent SLS encoding process.

3.1. Problem settings

We modify (2) to include a penalty term that takes into account the cost of encoding the SLS enhancement layer given a particular choice of base layer parameters. Thus,

$$\mathcal{P}^* = \arg \min_{\mathcal{P}:\mathcal{D}_b(\mathcal{P})\leq\mathcal{D}_t} \mathcal{R}_b(\mathcal{P}) + \alpha\mathcal{R}_e(\mathcal{P}) \,. \tag{7}$$

The parameter α, $0 \leq \alpha \leq 1$ controls the AAC-SLS performance trade-off. When $\alpha = 0$, (7) degenerates to the regular AAC optimization problem (2), while $\alpha = 1$ is the other extreme where base layer rate is ignored and AAC parameters are chosen to minimize the total rate, i.e., provide the best lossless compression. Note that

irrespective of α the distortion at the base layer is bound, thus ensuring the required level of perceptual quality. Alternatively, one can fix the base layer rate and optimize the trade-off between base layer distortion and lossless rate. It should be noted that although (7) is a minimization over the set of AAC parameters, our definition of $\mathcal{R}_e(\mathcal{P})$ subsumes an optimization over SLS parameters too, thus effecting a jointly optimized selection of all the HD-AAC encoding parameters. We propose here, trellis-based approaches to solve the above optimization problem for the two distortion measures MMNMR (4) and AANMR (3).

3.2. MMNMR solution

The maximization involved in MMNMR(4) translates the distortion constraint in (7) into the equivalent form $d_n^l(s_n^l) \leq \mathcal{D}_t$. Since the bit-rate (both AAC and SLS) for coding a particular frame depends only on the choice of encoding parameters for that frame, the rate-cost in (7) can be decomposed across frames. Thus, the overall minimization problem in (7) is equivalent to N per frame minimizations of the form:

$$\mathbf{p}_n^* = \arg \min_{\mathbf{p}: d_n^l(s^l) \leq \mathcal{D}_t} \mathbf{R}_b^n(\mathbf{p}) + \alpha \mathbf{R}_e^n(\mathbf{p}) \tag{8}$$

where $\mathbf{R}_b^n(\mathbf{p})$ and $\mathbf{R}_e^n(\mathbf{p})$ denote, respectively, the base and enhancement layer bit-rates for frame n, when the AAC parameter set is $\mathbf{p}$. A computationally efficient, trellis-based algorithm has been proposed in [5] that solves the above problem for the case $\alpha = 0$. A trellis, with stages corresponding to SFBs of the frame, and nodes in each stage corresponding to different pairs (s^l, h^l) of SF and HCB values, is constructed. Each node is populated with the distortion $d_n^l(s^l)$ corresponding to that state, and bits $Q_n^l(s^l, h^l)$ needed to entropy code the quantized AAC spectral data in the SFB. Transitions are associated with bit costs $E(s^{l-1}, s^l)$ and $F(h^{l-1}, h^l)$, respectively, required to differentially encode SFs and run-length encode HCB values (for more details refer [5]). Only nodes that satisfy the distortion constraint in (8) are retained. A Viterbi algorithm is used to find the path through this trellis (equivalently, a particular set $\mathbf{p}$) that minimizes the total bit-rate along the path. The optimal path corresponds to $\mathbf{p}_n^*$.

We modify this trellis to include the cost of encoding the enhancement layer. Note that since the SF in each node of the trellis is fixed, the corresponding base layer quantized spectral data, and hence the mapped error for the SLS layer in the SFB, is known. Thus, each node of this trellis also has a corresponding value of the MSB-plane $\mathcal{M}_n^l$. We now perform an exhaustive search over the lazy bit-planes $\mathcal{L}^l$ to obtain the least number of bits in the enhancement layer (bits $G(s^l, \mathcal{L}^l)$ needed to bit-plane code the mapped error in the node, and bits $H(\mathcal{L}^l)$ needed to Huffman code the lazy bit-planes). Since in case of either BPGC or LEMC an arithmetic coder is involved, an exact estimate of $G(s^l, \mathcal{L}^l)$ cannot be obtained. Instead we use the following approximation: if the arithmetic coder is ideal, the rate for encoding bit $b[k,j]$ will be given by $-\log_2 [b[k,j]Q^x(j) + (1 - b[k,j])(1 - Q^x(j))]$, with $Q^x(j)$ being the probability assignment of the BPGC (or LEMC) coder for the j^{th} bit-plane when $\mathcal{L}^l = x$. Thus, we associate with each node the corresponding optimal value of $\mathcal{L}^l$, and additional bit costs $\alpha G(s^l, \mathcal{L}^l)$ and $\alpha H(\mathcal{L}^l)$. The bit-rate for differentially encoding the MSB-planes $\mathcal{M}_n^l$ (for explicit bands) is included as a transition cost in the trellis (similar to the cost of the SFs). The same Viterbi algorithm as before is used to find the path that minimizes the total cost in (8). The algorithm is repeated for each frame.

3.3. AANMR solution

In the case of AANMR (3), the constrained problem is solved via the Lagrangian formulation (similar to [6]). Unconstrained minimization is performed on the Lagrangian cost

$$\mathcal{J}(\mathcal{P}, \lambda) = \mathcal{R}_b(\mathcal{P}) + \alpha \mathcal{R}_e(\mathcal{P}) + \lambda \mathcal{D}_b(\mathcal{P}) \,, \tag{9}$$

where λ is a Lagrangian parameter. Given a particular value of λ, minimizing (9) yields the set of parameters $\mathcal{P}^*(\lambda)$ that minimizes $\mathcal{R}_b + \alpha \mathcal{R}_e$ while maintaining distortion $\mathcal{D}_b(\mathcal{P}^*(\lambda))$. By adjusting λ we can find the optimal set $\mathcal{P}^*$ for the desired distortion level, or the solution to (7). The cost $\mathcal{J}(\mathcal{P}, \lambda)$ can be again split into costs for individual frames:

$$\mathbf{J}_n(\mathbf{p}_n, \lambda) = \mathbf{R}_b^n(\mathbf{p}_n) + \alpha \mathbf{R}_e^n(\mathbf{p}_n) + \lambda \big(\frac{1}{L} \sum_{l=0}^{L-1} d_n^l(s_n^l) \big) \,. \tag{10}$$

We use the same trellis as in Sec. 3.2 but the optimal path is chosen as the one that minimizes the above cost. Note that we now retain all the nodes in the trellis (since there is no per frame distortion constraint as in Sec. 3.2). The above minimization is performed for each frame. If the constraint on the overall distortion is not met, λ is adjusted and the minimization re-done for all frames. Unlike in the MMNMR case, multiple passes of the file (for different values of λ) are necessitated. Since the distortion and rate costs in each node of the trellis (for every frame) is independent of λ, complexity can be minimized by running the algorithm in parallel for multiple values of λ, while sharing the values of these costs.

4. RESULTS

The following codecs are compared in our experiments: *(i)* The proposed AAC+SLS codec with trellis-based optimization; *(ii)* SLS NC: for fairness, the $\mathcal{L}_n^l$ are obtained by exhaustive search, i.e., the coder provides the optimal non-core performance; *(iii)* TLS based AAC+SLS codec: same as the MPEG-4 HD-AAC reference software, except that $\mathcal{L}_n^l$ are obtained by exhaustive search. The experiments were performed on mono, 48kHz sampled versions of the 15 audio files[1] used in [4, 7].

In Fig. 2, the core bit-rate and the total (lossless) bit-rate for different constraints on the distortion (measured as MMNMR) have been compared. The bit-rates at the same distortion have been averaged over the test files. It is interesting to note that the core bit-rate is relatively insensitive to α, compared to the total bit-rate. This suggests that with minimal sacrifice in optimality of the AAC layer (see curves for $\alpha = 0.8$ or $\alpha = 0.9$), *considerable gains in lossless compression can be obtained by accounting for SLS impact during AAC encoding*. As is evident, the lossless compression can be improved even beyond that of the optimal SLS NC coder. Similar results are shown with the AANMR measure in Fig. 3. Note also that careful optimization using the trellis-based methods results in substantial performance improvements over the TLS-based reference software.

In Table.1 we compare, using the Objective Difference Grade (ODG) of the PEAQ method (ITU-R BS.1387-1) implemented in the AFsp library [8], the perceptual quality of the core produced by the different codecs, the core bit-rate being 64kbps. The results have again been averaged over the test files. In case of SLS NC, the fine-grained bitstream is truncated to obtain a partial reconstruction at 64kbps. An ODG of -4 indicates 'very annoying' perceptual quality, while 0 signifies that the difference between the original

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

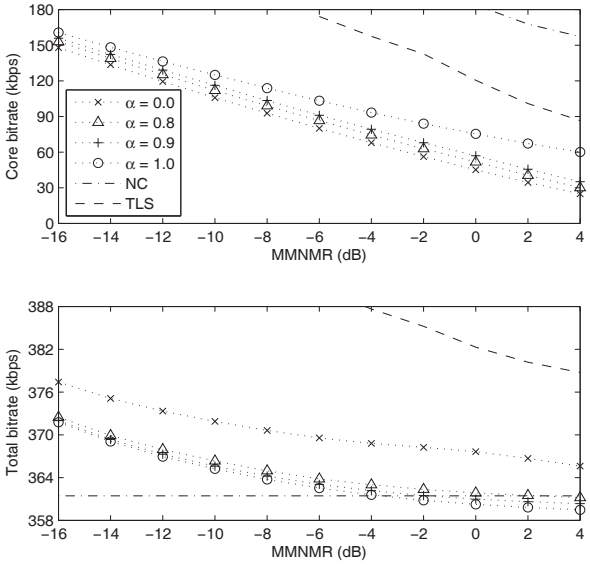

Fig. 2. MMNMR minimization: core bit-rate and total bit-rate for different distortion constraints and different values of α.

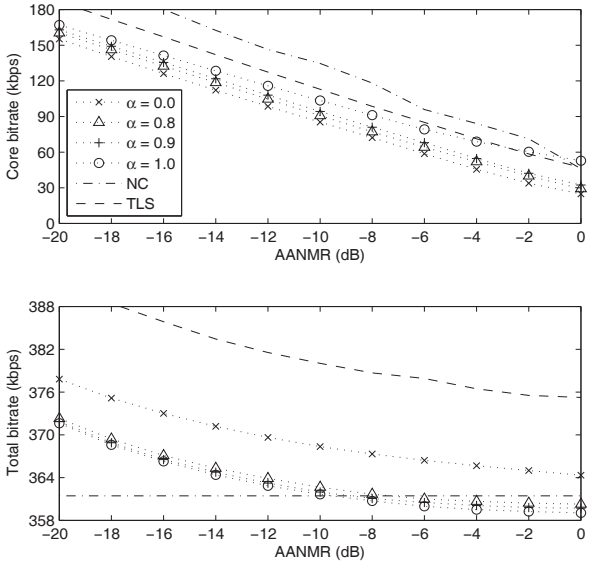

Fig. 3. AANMR minimization: core bit-rate and total bit-rate for different distortion constraints and different values of α.

and coded versions is imperceptible. Also provided in the table is the lossless performance in terms of the total bit-rate (in kbps), and the compression ratio. As expected, the best ODG measurements (-0.943 with MMNMR optimization and -0.885 with AANMR), were obtained using the trellis-based approach with $\alpha = 0$, either optimizations outperforming the HD-AAC reference and SLS NC. Choosing $\alpha = 1$ provides lossless performance even better than that of SLS NC, but the perceptual quality is significantly impaired. A better trade-off is obtained when $\alpha = 0.92$ (with MMNMR) or $\alpha = 0.79$ (with AANMR), providing good perceived quality in the core layer (ODG close to -1) as well as lossless performance comparable to SLS NC (compression ratio $= 2.125$).

	PEAQ ODG	Tot. Bit-rate	Comp. Ratio
MMNMR 0.00	-0.942	368.7	2.083
MMNMR 0.92	-0.957	361.3	2.125
MMNMR 1.00	-1.919	360.0	2.133
AANMR 0.00	-0.885	367.0	2.093
AANMR 0.79	-1.125	361.4	2.125
AANMR 1.00	-2.060	359.7	2.135
TLS	-1.676	376.3	2.041
SLS noncore	-1.923	361.4	2.125

Table 1. Core at 64kbps: ODG given by PEAQ, total bit-rate and compression ratio.

5. CONCLUSION

A novel trellis-based algorithm for optimization of the HD-AAC encoding process is proposed. The problem of choosing the coding parameters is formulated as a distortion constrained optimization, that allows a trade-off between the perceptual quality of the AAC core and the lossless compression of SLS. Two candidate distortion measures, MMNMR and AANMR, have been considered. The results demonstrate that significant improvement in lossless compression can be achieved with minimal detriment to the perceptual quality of the AAC layer, by careful selection of the AAC encoding parameters. Contrary to popular opinion, the results indicate that the presence of a superior AAC core does not preclude excellent lossless compression performance.

6. REFERENCES

[1] ISO/IEC 14496-3:2005/Amd.10:2008, "Information technology - Coding of audio-visual objects - Part 3: Audio - Amd. 10: HD-AAC profile," 2008.

[2] ISO/IEC 14496-3:2005, "Information technology - Coding of audio-visual objects - Part 3: Audio - Subpart 4: General audio coding (GA)," 2005.

[3] ISO/IEC 14496-3:2005/Amd.3:2006, "Information technology - Coding of audio-visual objects - Part 3: Audio - Amd. 3: Scalable lossless coding (SLS)," 2006.

[4] 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," *J. Audio Eng. Soc.*, vol. 55, no. 1/2, pp. 27–43, Jan./Feb. 2007.

[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] V. Melkote and K. Rose, "Trellis-based approaches to rate-distortion optimized audio encoding," *To appear in IEEE Trans. Audio, Speech, and Lang. Proc.*, Feb. 2010.

[7] 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.

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