A TWO-LAYERED TRELLIS APPROACH TO AUDIO ENCODING

Vinay Melkote and Kenneth Rose

ECE Department, University of California Santa Barbara, CA - 93106, USA

ABSTRACT

The fact that audio compression for streaming or storage is usually performed offline alleviates traditional constraints on encoding delay. We propose a rate-distortion optimized approach, within the MPEG Advanced Audio Coding framework, to trade delay for optimal window switching, resource allocation and selection of quantization and coding parameters for the entire audio file using a two-layered trellis. Stages of the outer trellis correspond to audio frames, nodes represent window choices, and branches implement transition constraints. The inner trellis operates within each node of the outer layer and has stages corresponding to scalefactor bands and nodes representing combinations of quantization and coding parameters. A suitable cost, comprising bit consumption and psychoacoustic distortion, is optimized via multiple passes through the two-layered trellis to achieve the desired bitrate. The procedure thus optimizes most of the encoding decisions involved in audio compression. Objective and subjective tests show considerable performance gains.

Index Terms— Audio coding, trellis optimization, AAC, window switching

1. INTRODUCTION

Audio streaming and storage are popular applications of audio compression techniques. For such purposes audio is compressed in advance, i.e., offline. This alleviates constraints on encoding delay which are imposed on traditional encoders. This observation forms the motivation for our proposed rate-distortion (RD) optimized approach to audio compression. The approach trades delay for improved encoding decisions across all frames of the audio file and also within each frame. We demonstrate this principle by jointly optimizing window switching decisions and bit resource allocation across frames and the choice of quantization and coding parameters within each frame in the MPEG Advanced Audio Coding (AAC) [1] framework.

In AAC, the audio file is divided into overlapping frames where time to frequency mapping is obtained via the modified discrete cosine transform (MDCT). MDCT coefficients are grouped into frequency bands called scalefactor bands (SFBs). A psychoacoustic model is used to find the masking thresholds in each SFB. A typical quantization and coding module employs a two loop search (TLS) to find encoding parameters called *scalefactors* (SFs) and *huffman code books* (HCBs) for each SFB, such that a bit budget is met while attempting to minimize quantization noise relative to the thresholds. AAC allows a finite choice of HCBs and SFs for each SFB. Side information specifying these parameters consumes a portion of the bit budget. Stationary frames are encoded using “LONG” windows (1024 new samples) and non-stationary with 8 “SHORT” windows (128 new samples each). “START” and “STOP” windows of suitable shapes are used to transition between the above configurations. The corresponding decision procedure is called *Window Switching*. In addition, the “SHORT” windows within a frame can be “grouped” for coding efficiency. A *bit reservoir* allows saving bits unused by past frames to encode future frames that require more bits. A detailed description can be found in [1].

1.1. Prior Work

In [2] we proposed a trellis based approach to jointly optimize window switching decisions and bit distribution across frames and demonstrated the suboptimality of the bit reservoir technique and heuristic window switching decisions. A trellis (the “outer” trellis shown in Fig. 1) is constructed where stages correspond to frames of audio, nodes correspond to the four window choices and transitions are allowed only between compatible window types. A suitable cost measure accounting for psychoacoustic distortion per frame and bit consumption, was minimized via multiple passes through this trellis.

In [2] the focus was on optimization across frames in terms of window switching and bitrate allocation. Standard TLS was employed for determining the encoding/quantization parameters of each frame. On the other hand, a trellis based optimal selection of internal parameters for an audio frame was also proposed by our lab in [3] and [4], and shown to outperform TLS. In [4] a trellis (the “inner” trellis shown in inset of Fig. 1) is constructed with SFBs as stages and combinations of SFs and HCBs as nodes within each stage. Each node (SF and HCB pair) that is traversed determines the distortion involved in quantizing the spectral values of the corresponding SFB and the number of bits needed to encode

This work was supported in part by the University of California MICRO Program, Applied Signal Technology, Inc., Cisco Systems, Inc., Dolby Laboratories, Inc., and Qualcomm, Inc.

them. Transitions between nodes determine the number of bits needed to differentially or run-length encode the choice of SFs and HCBs, respectively. A Lagrangian cost measure involving distortion and bit consumption, suitably tailored to reduce the number of paths at each stage of this trellis to one survivor per node, was optimized by muliple traversals of this trellis. The objective of the iteration was to minimize frame distortion subject to the frame bitrate constraint. The above method drastically reduces the complexity involved in searching through all combinations of SFs and HCBs for the SFBs of the frame to find the optimal set (MPEG AAC allows a choice of about 120 SF values and 12 HCBs for each SFB).

The present work is an enhancement of our work in [2] by embedding, after suitable modifications, the intraframe trellis of [4] within the interframe trellis of [2]. The aim is thus to globally optimize the encoding decisions involved. Important related results in RD optimal encoding of audio include [5] which proposes a MILP technique for parameter selection within a frame, [6] on time segmentations of audio frames and [7] on RD optimization based bit reservoir control.

2. PROPOSED METHOD

2.1. Distortion Measure

The distortion measure used in the proposed RD optimization algorithm is the “total noise to mask ratio” (TNMR). This measure is defined for every frame of audio. It is the sum of noise to mask ratios (NMRs) of all SFBs in the frame. In the SHORT configuration it is the sum of SFB NMRs over 8 short frames. TNMR (and bit consumption) of a frame depend on the choice of quantization and coding parameters and window configuration for that frame. The measure “average noise to mask ratio” (ANMR) used in [4] is the TNMR of a frame divided by the number of scalefactor bands, which depends on the window configuration. This implies that the ANMR minimization method of [4] can be adopted to minimize TNMR within each frame. The SFBs for LONG and SHORT windows have different lengths in ‘bark’ units. Thus, ANMR which is average distortion per SFB does not yield fair comparison between different window types, hence our preference for TNMR.

2.2. Problem Statement

Consider an audio file of N frames. $D_k(w_k, S_k, H_k)$ is a perceptually relevant distortion measure (here TNMR) for the k^{th} frame. This distortion and corresponding bit cost b_k depend on the window decision w_k, the combination of SFs $S_k = (s_1, \ldots, s_L)$ and HCBs $H_k = (h_1, \ldots, h_L)$, where L is the number of SFBs of the frame. w_k is one of four possible window configurations: LONG, SHORT, START or STOP. S_k and H_k are L-tuples drawn from the sets $\mathcal{S}^L$ and $\mathcal{H}^L$, respectively, where $\mathcal{S}$ and $\mathcal{H}$ are the finite sets of possible SFs and HCBs in the AAC framework. b_k includes bits needed for the quantized spectral values, differentially encoded SFs, run-length encoded HCBs and window configuration and grouping. Let $\mathcal{W}$ be the set of all admissible sequences $W = (w_1, \ldots, w_N)$, i.e., sequences of window decisions where START and STOP windows are appropriately used between LONG and SHORT windows. $\mathcal{P} = (\mathcal{S}^L \times \mathcal{H}^L)^N$ is the set of all possible parameter selections $P = ((S_1, H_1), \ldots, (S_N, H_N))$ for the N audio frames. For each P and W the corresponding bit costs across frames is $B = (b_1, \ldots, b_N)$. Given an average bitrate R we formulate two constrained minimization problems, depending on how one defines an *overall* distortion given the frame-specific distortions. Find W^* and P^* (and the corresponding B^*) such that:

$$(W^*, P^*) = \arg \min_{\mathcal{W} \times \mathcal{P}} \frac{1}{N} \sum_{k=1}^{N} D_k(w_k, S_k, H_k)$$

called the “$Average\ TNMR\ (ATNMR)\ problem$”
or

$$(W^*, P^*) = \arg \min_{\mathcal{W} \times \mathcal{P}} \max_{k=1}^{N} D_k(w_k, S_k, H_k)$$

called the “$Maximum\ TNMR\ (MTNMR)\ problem$” subject to the rate constraint (in either case) $\frac{1}{N} \sum_{k=1}^{N} b_k^* \leq R$.

2.3. Two-layered trellis solution

2.3.1. ATNMR problem

An outer trellis is constructed across frames, as shown in Fig. 1. The ATNMR problem is solved using a Lagrangian minimization approach, i.e, by finding the path through the outer trellis which minimizes the cost

$$\sum_{k=1}^{k=N} \left(D_k(w_k, S_k, H_k) + \lambda b_k(w_k, S_k, H_k) \right)$$

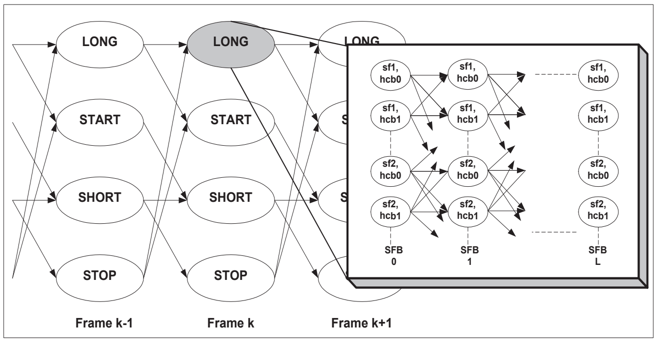

Fig. 1. A two-layered trellis with outer layer spanning frames with nodes as window choices. Each outer layer node contains an inner trellis (shown in inset) where stages are frame SFBs and where states represent SF and HCB choices.

where λ is the Lagrangian parameter. Since the distortion and corresponding bit consumption for a frame in a particular window configuration w_k is only determined by the choice of (S_k, H_k) (and is independent of the choice (S_i, H_i) for $i \neq k$) the above minimization implies that at each node (i.e., window choice w_k for frame k) of the outer trellis, the pair (S_k^*, H_k^*) which minimizes $D_k(w_k, S_k, H_k)+\lambda b_k(w_k, S_k, H_k)$ needs to be found. This task is performed by the complexity minimizing trellis based optimization method of [4]. At each outer layer node we traverse once an inner trellis spanning the SFBs (inset in Fig. 1). The path through the inner trellis which minimizes the above cost is found (detailed description of such minimization can be found in [4]). The corresponding values, $D_k^*(w_k)$ and $b_k^*(w_k)$ for that frame, are then stored in the node of the outer trellis. After performing this step at each outer layer node, the "best" path through the outer trellis (or equivalently best sequence of window decisions) is determined. The above definition of the cost measure allows us to compare at each stage, the different paths ending in the same outer node, and choose one survivor. The complexity of comparing all paths through the trellis is thus reduced to comparing four survivor paths (one ending in each node) at each stage and ultimately at the end of the trellis. If the rate constraint is not met λ is altered suitably and the process repeated.

2.3.2. *MTNMR problem*

Here too, the two layer trellis is as in Fig. 1. The MTNMR problem is solved by fixing a parameter γ, which controls the maximum frame distortion across the entire audio file. The aim is to find a path (i.e. $w_1, \ldots, w_N$) through the outer trellis which minimizes $\sum_{k=1}^{k=N} b_k(w_k, S_k, H_k)$ with distortion $D_k(w_k, S_k, H_k) \leq \gamma \ \forall k$ and then vary γ until the desired average rate is achieved. This implies that, at each outer layer node we need to find the parameter sets S_k^* and H_k^* which minimize the frame bit cost $b_k(w_k, S_k, H_k)$ subject to the distortion constraint, γ. In [4], the trellis based Lagrangian minimization technique finds the parameters that achieve the minimum ANMR given a rate constraint for the frame. A similar procedure holds if a distortion constraint, as here, is imposed. Accordingly, we traverse the inner trellis multiple times to find the required S_k^* and H_k^*. This fixes $b_k^*(w_k)$ for the frame which is stored in the node of the outer trellis. The path through the outer trellis which minimizes the total number of bits consumed is found. γ is varied and the trellis is retraversed if the rate constraint is not met. Note that here too we can compare different paths ending in the same node of the outer trellis and choose a survivor.

We re-emphasize that the outer trellis imposes delayed decisions regarding each frame so as to achieve effective bit distributions and window switching. This delay is incurred only at the encoder.

3. EXPERIMENTS AND RESULTS

The base AAC encoder or the reference model (RM) utilizes TLS of the MPEG verification model, transient detection based window switching and conventional bit reservoir techniques. The models L2_MTNMR and L2_ATNMR (which address the MTNMR and ATNMR problems respectively) employ the two-layered trellis to make encoding decisions. For comparison we also include the corresponding models, L1_MTNMR and L1_ATNMR, which have only the outer trellis, i.e., within a frame TLS determines SFs and HCBs. These are the models proposed in [2]. The model RM_TB is RM with TLS replaced by the trellis based parameter selection within each frame (i.e., no outer trellis). This model is similar to the one proposed in [4] but with the addition of standard window switching and bit reservoir. (The trellis based method in [4] was implemented to be used without "Window Switching". We have extended it to cater to all window choices and "SHORT" groupings.) In our experiments we chose the best among 8 possible "SHORT" window grouping choices.

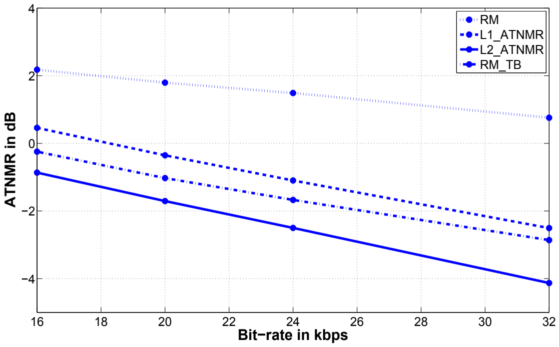

Fig. 2. Average TNMR vs bitrate produced by: the reference model(RM), the single trellis schemes, RM_TB and L1_ATNMR, and the two-layered trellis scheme L2_ATNMR

Fig. 2 shows ATNMR versus bitrate results obtained by the RM, RM_TB, L1_ATNMR and L2_ATNMR models. The two-layered trellis model, L2_ATNMR, outperforms its nearest competitor by 0.7-1.4 dB. The distortion values have been averaged over 6 audio samples with different characteristics drawn from the EBUSQAM database. Fig. 3 shows similar results with respect to the distortion measure MTNMR. Here, the two-layered trellis scheme, L2_MTNMR, performs better than the nearest competitor by 1.5-2 dB. We observed in [2] that minimizing the maximum distortion (MTNMR) yields better subjective quality than when the average distortion (ATNMR) is minimized. Thus MUSHRA tests comparing the MTNMR optimization schemes with the reference RM have been conducted. Fig. 4 compares average MUSHRA test scores of audio encoded at 16kbps with these schemes.

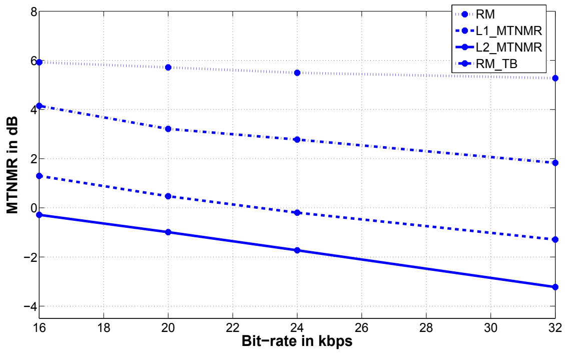

Fig. 3. Maximum TNMR vs bitrate produced by: the reference model(RM), the single trellis schemes, RM_TB and L1_MTNMR, and the two-layered trellis scheme L2_MTNMR

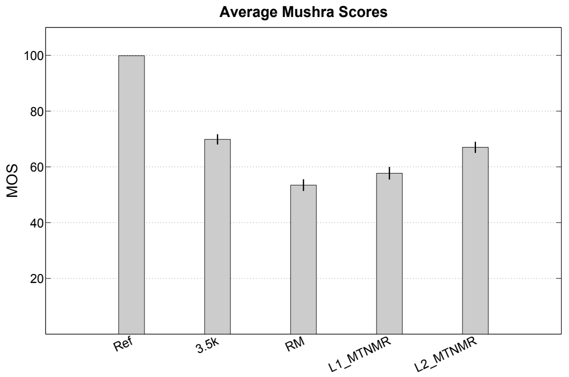

Fig. 4. MUSHRA scores for audio encoded at 16kbps using the different schemes. "Ref" is the original and "3.5k" is the low pass anchor.

The single layer trellis scheme performs better than RM and the two-layered trellis scheme beats both of them. The quality gains are higher for pieces like german speech, harpsichord and drums than for stationary audio like accordion.

The complexity of traversing the outer trellis (and with it the inner trellis at each outer node) multiple times can be considerably reduced at some cost in memory. Multiple outer trellises working for a range of λ (or γ) can be maintained to reuse computations. For the ATNMR problem, the λ for the Lagrangian of the inner trellis is the same as that of the outer trellis, and the best paths through the inner trellis for this range of λ can be simultaneously found to minimize computation. For the MTNMR problem, though, at each outer node the inner trellis has to be traversed multiple times for each γ of the range, each traversal producing a different RD point for the frame. The RD points found when working the inner trellis for one γ can be used to subsequently reduce the search region for other γ converging faster. The RD curve (with distortion on the log scale) for each frame is near linear and RD points can be interpolated for an approximate Lagrangian parameter for the inner trellis satisfying the γ constraint. This results in about 0.2 dB decrease in gain, while reducing run time by about a factor of 4. The majority of the computational complexity is due to populating the inner trellis in the different window configurations and groupings for each frame.

4. CONCLUSIONS

Two-layered trellis based schemes that utilize delayed decision to optimally compress audio and improve its objective and subjective quality have been demonstrated. The schemes though complex generate standard compatible bit streams and the end user is unaffected by this encoding delay. Improvements in distortion measure and better comparison between the different window configurations are expected to substantially improve the subjective quality.

5. REFERENCES

[1] ISO/IEC std, "Information technology - generic coding of moving pictures and associated audio," *ISO/IEC JTC1/SC29 13818-7:1997(E)*, 1997.

[2] V. Melkote and K. Rose, "Trellis based approach for joint optimization of window switching decisions and bit resource allocation," in *To appear in Proc. 123rd AES convention, New York*, Oct 2007, Preprint 7216.

[3] A. Aggarwal, S.L. Regunathan, and K. Rose, "Trellis-based optimization of MPEG-4 advanced audio coding," in *Proc. IEEE Workshop on Speech Coding*, 2000, pp. 142–144.

[4] A. Aggarwal, Shankar L. Regunathan, and K. Rose, "A trellis-based optimal parameter value selection for audio coding," *IEEE Transactions on Audio, Speech and Laguage Processing*, pp. 623–633, Mar 2006.

[5] C. Bauer and M. Vinton, "Joint optimization of scale factors and huffman codebooks for MPEG4 AAC," in *Proc. IEEE Workshop. MMSP*, 2004, pp. 111–114.

[6] 0.A. Niamut and R. Heudsens, "Optimal time segmentation for overlap-add systems with variable amount of window overlap," *IEEE Sig. Proc. Lett.*, vol. 12, pp. 665–668, Oct 2005.

[7] E. Camberlein and P. Philippe, "Optimal bit-reservoir control for audio coding," in *Proc. IEEE Workshop. App. of Sig. Proc. to Audio and Acoust.*, 2005, pp. 251–254.