NEAR-OPTIMAL SELECTION OF ENCODING PARAMETERS FOR AUDIO CODING

Ashish Aggarwal, Shankar L. Regunathan and Kenneth Rose

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106, USA
Email:[ashish,otto,rose]@scl.ece.ucsb.edu

ABSTRACT

We address the issue of optimizing side information rate for efficient audio coding. In coders such as the MPEG-4 AAC, at rates around 16kbps to 48kbps, the side information rate forms a substantial part of the total rate. The parameter search procedure in the Verification Model optimizes each band separately and results in poor performance at low rates. We propose to jointly optimize the encoding parameter of all the bands. The near-optimal solution using a brute force search has drastic computational complexity. However, the same solution is obtained at a much reduced complexity using a Viterbi search through a trellis. The search procedure is developed and evaluated for two objective measures, the average and the maximum noise-mask ratio. For both the measures, the trellis-based search yields substantially better solutions. In particular, trellis-based optimization of maximum noise-mask ratio greatly improves the performance of AAC at low rates. The resulting bit stream is standard-compatible, and the additional complexity due to the proposed optimization is only incurred at the encoder.

1. INTRODUCTION

Low rate coding of audio finds wide variety of applications including the Internet and wireless applications. The MPEG-4 Advanced Audio Coding (AAC) [1] achieves transparent quality above 64kbps. However, it suffers from relatively poor compression efficiency at low bit rates (from 16kbps to 48kbps). Further, if a multi-layer coder employs low rate AAC at its base layer, then its performance is substantially compromised. In order to mitigate this problem MPEG-4 currently uses a different coding scheme (TwinVQ) at low bit rates. We propose to directly attack the problem of poor performance of AAC at low bit rate.

The key step in audio coding is the dynamic bit allocation for the transform coefficients, which is used to exploit perceptual redundancies in the signal. The bit allocation information must be transmitted to the decoder as side information. In AAC, side information consists of the quantizer step size and the choice of Huffman codebook for each band. The side information parameters associated with adjacent bands are coded differentially or by run-length coding. Therefore, the side information rate and the total rate vary based on inter-band relationships of the parameters. This makes it difficult to select encoding parameters so as to minimize the objective measure for a specified total rate constraint. This optimization problem is further complicated by the existence of two popular objective measures, the average and the maximum noise-mask ratio, which can be used for performing bit allocation [2][3].

The standard optimization procedure [4] used in the Verification Model (VM) [5] of the AAC ignores the effect of inter-band dependencies on side information. Under this convenient assumption, the optimal solution is identical for both the objective measures, and parameters for each frequency band can be optimized *independently* using a two-loop search (TLS) [4]. However, the assumption is invalid and the search is ineffective in controlling the side information rate resulting in poor performance of VM at low rates.

We propose to *jointly* optimize the encoding parameters for all the frequency bands for the prescribed *total* rate. Note that the total bit rate accounts for quantizer indices and the side information. The encoding parameters are chosen to minimize the objective measure while taking into account the effect of inter-band dependencies. The joint optimization problem is formulated as a search through a trellis, and dynamic programming is proposed to drastically reduce the search complexity. The resulting quantization scheme greatly improves the low bit rate performance of AAC and, consequently, the performance of a multi-layer AAC system. It is important to emphasize that the bit stream syntax and the decoder are left unchanged, and the resulting bit stream is standard-compatible. In a recent independent paper [6] Prandoni and Vetterli outline a similar approach to optimize side information for a bank of quantizers.

The organization of the paper is as follows: Section 2 gives an overview of the AAC quantization scheme. The objective measures are outlined in Section 3. The Viterbi search for determining the optimal encoding parameters is explained in Section 4. In Section 5 we discuss the difference in output of the trellis-based coding for the two objective measures and in Section 6 we summarize the simulation results.

2. OVERVIEW OF AAC QUANTIZATION

A simplified, high-level block diagram of the AAC [1] encoder is shown in Figure 1. The quantization and coding (QC) module, which is central to this work, is shown in more detail. Fixed length modified discrete cosine transform converts a frame of 1024 time domain samples into the spectral domain. The time domain data is also input

This work is supported in part by the NSF under grants no. MIP-9707764, EIA-9986057, the University of California MICRO Program, Conexant Systems, Inc., Lernout & Hauspie Speech Products, Lucent Technologies, Inc., Medio Stream, Inc., and Qualcomm, Inc.

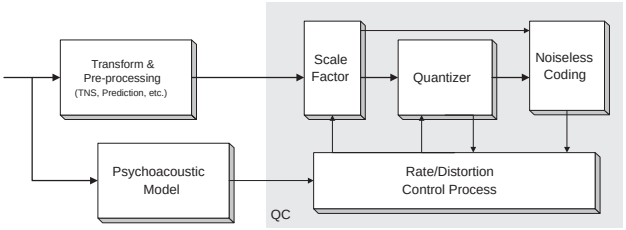

Figure 1: *Block diagram of the AAC encoder*

to the psychoacoustic model, whose output is the masking threshold for the spectral coefficients. These 1024 spectral coefficients are grouped into 49 scale factor bands (SFB).

All transform coefficients within a given band are quantized using the same non-uniform scalar quantizer. The scale factor (SF) parameter controls the quantizer step size and, consequently, the quantization noise level in the SFB. The quantized coefficients in a SFB are entropy coded using a Huffman codebook (HCB) which is chosen from a bank of 12 pre-designed HCBs. SFs are differentially encoded using a fixed Huffman code, and the choice of HCBs is encoded using a run-length code. The SF and HCB parameters are transmitted as side information along with the quantizer indices.

3. OBJECTIVE MEASURES

Objective measures used in rate-distortion optimization are designed in an effort to model the subjective perceptual distortion criteria. At low rates, the most widely used objective measures are based on the noise to mask ratio (NMR) [2]. For a given band, NMR is the ratio of noise energy in the band to its masking threshold. NMR may equivalently be viewed as a weighted squared error measure, $w_i d_i$, where d_i is the quantization noise energy (squared error) in band i and the weight, w_i, is chosen as the inverse of the masking threshold in that band.

NMR values obtained from different bands are usually combined to yield a scalar metric in order to perform bit allocation for the whole frame. Two common criteria are: *Average NMR (ANMR)*, which is the NMR averaged over all bands in the frame ($\sum_i w_i d_i$), and *Maximum NMR (MNMR)*, which is the maximum NMR in a frame ($\max_i w_i d_i$) [3].

In minimizing these distortion measures, it is typical to assume that the total rate can be expressed as sum of the rate used for coding individual bands. Note that this assumption implies that inter-band dependencies do not affect the total rate. Further, the side information rate required to transmit the bit allocation information to the decoder is assumed to be constant or insignificant. Under these conditions, high-resolution quantization theory [7] can be used to show that the condition for optimality is identical for both distortion measures, and states: $w_i d_i = const, \forall i$. Thus, parameter optimization becomes a simple water-filling algorithm in which the NMR is constant throughout the frame and this constant is adjusted to meet the target bit rate. The TLS [4] algorithm of AAC-VM is based on this result. However, the above assumptions are invalid and are the main cause of the VM's poor performance at low rates.

AAC employs run-length coding of HCBs and differential coding of SFs which introduce significant inter-band dependencies in coding of the side information. Also, the side information rate varies widely based on the choice of encoding parameters. *In principle, optimal encoding of bands may be performed independently (hence, $w_i d_i = const$) only if band dependencies are negligible.* At high rates, side information accounts for a small percentage of the total rate and the implicit assumptions used by TLS do not cause substantial damage. However, at rates around 16kbps, side information may be using as much as 50% of the total rate [8]. Thus, it is imperative to optimize the *total* bit rate taking into account the effects of inter-band dependencies. A straight-forward joint optimization of parameters for all the bands is prohibitively complex. In AAC, there are 49 bands, 60 SFs and 12 HCBs, and the complexity of a brute force search is of $O((60 * 12)^{49})$. In the next section, we outline a dynamic programming approach to find the optimal encoding parameters at reasonable complexity.

4. TRELLIS-BASED OPTIMIZATION OF AAC

4.1. Preliminaries and Notation

Let B be the target bit rate for the frame, and let b_i be the number of bits required to encode the spectral coefficients of the ith SFB where the SF and HCB in use are specified by s_i and h_i, respectively. The function $\mathcal{F}$ determines the number of side information bits needed to encode the SF as a function of s_i and (through differential coding) s_{i-1}. Similarly, $\mathcal{G}$ represents the number of bits needed to encode the HCB as a function of h_i and h_{i-1} (the run-length coding of HCB produces a fixed number of bits (9) when $h_i \neq h_{i-1}$).

4.2. ANMR

The trellis based optimization of ANMR is discussed in detail in [8]. We provide a brief summary here. The constrained optimization problem,

$$\min \sum_i w_i d_i \qquad \text{s.t}$$

$$\sum_i (b_i + \mathcal{F}(s_i - s_{i-1}) + \mathcal{G}(h_i - h_{i-1})) \leq B$$

is reformulated as the minimization of the Lagrangian

$$C = \sum_i w_i d_i + \lambda(b_i + \mathcal{F}(s_i - s_{i-1}) + \mathcal{G}(h_i - h_{i-1})). \quad (1)$$

Note that : *a)* C consists of non-negative terms; and *b)* the contribution of s_i and h_i to C only depends on "past" decisions s_{i-1} and h_{i-1}. Therefore, C can be minimized by a dynamic programming (Viterbi search) procedure.

4.3. MNMR

The problem of minimizing MNMR subject to total bit rate constraint can be stated as

$$\min \max_i w_i d_i \qquad \text{s.t}$$

$$\sum_i (b_i + \mathcal{F}(s_i - s_{i-1}) + \mathcal{G}(h_i - h_{i-1})) \leq B.$$

We make the usual observations about the constraint: *a)* it is the sum of non-negative terms; and *b)* the contribution of s_i and h_i to the constraint only depends on "past" decisions s_{i-1} and h_{i-1}. It is easy to see that a dynamic programming (Viterbi search) procedure is again applicable to find the optimal solution.

We construct a trellis where each stage corresponds to a SFB (total of 49 stages). The valid states at stage j represent all combinations of possible choices of s_j and h_j for which the NMR in the band is $\leq$ some constant (say γ). If the system passes through state k in stage j (denoted $\Upsilon_{k,j}$) then it employs the kth pair of parameters for the jth SFB: $(s_j, h_j) = (SF, HCB)_k$. Clearly, every path on the trellis represents a particular choice of parameters for the frame. The Viterbi search can be used to find the path through this trellis that achieves the global minimum bit rate for a given γ. The value of γ that achieves the target bit rate constraint can be searched using an iterative procedure.

We define state-transition cost $T_{k\rightarrow l,j}$ as the bit rate for a transition from $\Upsilon_{k,j-1}$ to $\Upsilon_{l,j}$. This cost is:

$$T_{k\rightarrow l,j} = \mathcal{F}(s_{l-j}, s_{k,j-1}) + \mathcal{G}(h_{l,j} - h_{k,j-1}) \qquad (2)$$

The number of bits required to transmit the quantized spectral coefficients using $(SF, HCB)_k$ for stage j is denoted by $b_{k,j}$. Finally, we denote by $C_{k,j}$ the cost of the minimum cost (partial) path that ends at $\Upsilon_{k,j}$. This is also referred to as the *metric* of $\Upsilon_{k,j}$. The search algorithm is outlined next.

1. *Initialize* γ.
2. *Find Valid States.* A state $\Upsilon_{i,j}, \forall i$ is a valid state if $w_j d_i \leq \gamma$, where d_i is the distortion in the band using the state parameters.
3. *Initialize* metrics $C_{k,0} = 0, \forall k$, and $j = 1$.
4. *Search.* $\forall l$ s.t $\Upsilon_{l,j}$ is a valid state find the best path leading to $\Upsilon_{l,j}$ by computing the metric
$$C_{l,j} = \min_k \{C_{k,j-1} + b_{l,j} + T_{k\rightarrow l,j}\} \qquad (3)$$
5. If $j \leq 49$: $j \leftarrow j + 1$, go to 2.
6. *Adjust rate.* For the optimal $\mathbf{s}$ and $\mathbf{h}$, compare total bit rate to prescribed rate. If the constraint is not met adjust γ and go to 2.

For rate savings AAC allows any set of SF and HCB values to be assigned to a band that is below the masking threshold. This is incorporated in the trellis by splitting every state into two - one where quantization is performed using the assigned SF and HCB values, and the other where all quantized coefficients are set to zero.

5. TRELLIS BASED ANMR AND MNMR

It is interesting to compare the band NMR achieved by the trellis-based optimization on the two objective measures. The ANMR criteria allows for much variation in the band NMR and hence the result differs significantly from the high resolution result ($w_i d_i = const$). This may sometimes result in subjective quality degradation. On the other hand, trellis-based scheme for minimizing MNMR guarantees a band NMR smaller than that obtained by independently optimizing each band in the TLS scheme.

6. SIMULATION RESULTS

In this section, we summarize the experimental setup including implementation details, and provides the simulation results. For clearer comparison the TLS of MPEG-4 VM of AAC (VM-TLS) is used with some minor modification as follows. The SF values are chosen such that the NMR in each band is constant (say K). The total bit rate required for encoding the frame given these SFs is computed, and the value of constant K is adjusted so as to meet the constant target rate constraint. The psychoacoustic model was implemented from [9][1] with minor modifications and simplifications. The spreading function and the prediction to find the tonality factor was directly applied to the MDCT coefficients. For the test set, eight audio files of sampling rate 44.1kHz were taken (seven from the MPEG SQAM [5] database), which included tonal signals, castanets, two singing files and two speech files, one with a male-german speaker.

The trellis-based minimization of ANMR (TB-ANMR) and MNMR (TB-MNMR) is implemented as explained in sections 4.2 and 4.3 respectively. The total number of states account for all combinations of 60 SF and 12 HCB values, i.e., the total number of states was $60 * 12 * 2 = 1440$. To further reduce complexity, the transition at each state was restricted to the four nearest HCB values. Hence, the search complexity for the trellis-based scheme is of $O(60{*}4{*}2{*}49 = 480 * 49)$.

6.1. Objective results for a single-layer coder

We compared the performance of TB-MNMR, TB-ANMR and VM-TLS on the test set. Figures 2 and 3 depict the distortion-rate curve of a single-layer coder for the test set. Figures 2 shows the performance of the three schemes evaluated using the ANMR measure. Note here that TB-MNMR is optimized for the MNMR measure but evaluated using ANMR. TB-ANMR outperforms the standard VM-TLS technique. Also of interest is the fact that the TB-MNMR scheme performs better than VM-TLS even when evaluated using ANMR as a distortion criteria.

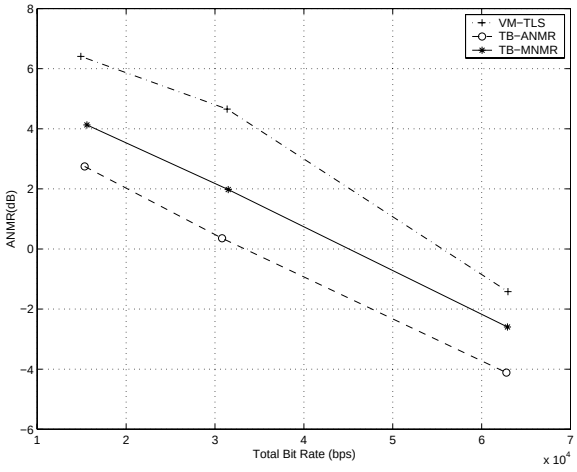

Figure 2: *Single-layer coder: ANMR vs. bit rate for VM-TLS, TB-ANMR and TB-MNMR.*

Figures 3 shows the performance of three schemes evaluated using the MNMR measure. At higher rates, we see that savings in bit rate for TB-ANMR scheme occur at the expense of very high NMR in some bands.

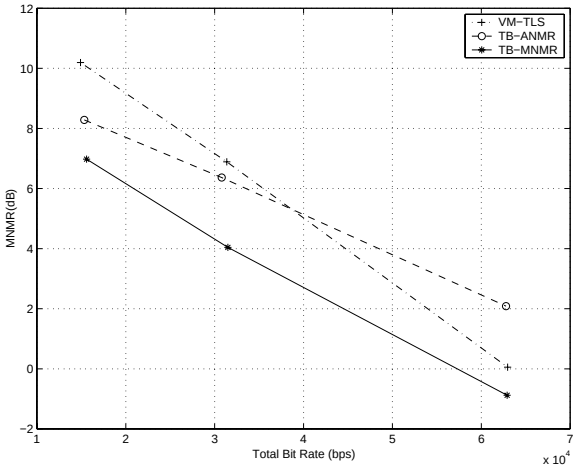

Figure 3: *Single-layer coder: MNMR vs. bit rate for VM-TLS, TB-ANMR and TB-MNMR.*

6.2. Subjective results for a single-layer coder

The three systems under test were evaluated at 16kbps using the ITU 5 grade ACR scheme to get the MOS scores. The listening test was performed using 8 listeners (including several trained listeners). Figure 4 shows the performance of the three schemes.

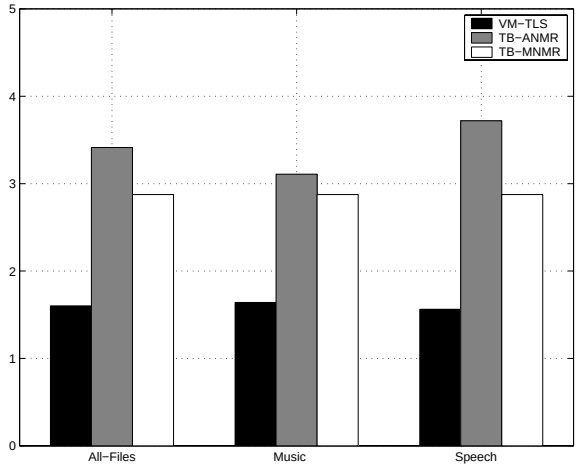

Figure 4: *MOS scores for VM-TLS, TB-ANMR and TB-MNMR for a test set of 4 music and 4 speech files*

Within the margin of error, TB-ANMR and TB-MNMR yield almost the same quality for music signals. For speech signals however, TB-ANMR has better performance than TB-MNMR. Both, TB-ANMR and TB-MNMR are significantly better than VM-TLS.

TB-ANMR produces a low-pass effect while TB-MNMR output contains more high frequency components and yields a crisper sound. However, TB-MNMR has audible artifacts in the 5-10kHz region which explain its poor performance on speech signals. We believe that these artifacts are due to the psychoacoustic model and can be removed.

7. CONCLUSION

In this paper, we derived a trellis-based optimization scheme for AAC for minimizing two different objective measures; average NMR and maximum NMR. The scheme substantially enhances performance at low bit rates. We showed that if the bands are assumed independent then the two objective measures yield an identical solution. However, the solution exhibits poor performance at low rates. The key step was to reformulate the parameter optimization problem at the encoder to account for inter-band dependencies in encoding side information, and utilize dynamic programming to obtain the solution at manageable complexity. The bit stream is standard-compatible, and the additional computational complexity is incurred only at the encoder.

8. REFERENCES

[1] ISO/IEC JTC1/SC29, "Information technology - very low bitrate audio-visual coding," *ISO/IEC IS-14496 (Part 3, Audio)*, 1998.

[2] R. Beaton, *et al.*, "Objective perceptual measurement of audio quality," in *Collected Papers on Digital Audio Bit-Rate Reduction* (N. Gilchrist and C. Grewin, eds.), pp. 126–152, Audio Engineering Society, 1996.

[3] H. Najafzadeh and P. Kabal, "Perceptual bit allocation for low rate coding of narrowband audio," *Proc. of ICASSP*, vol. 2, pp. 893–896, 2000.

[4] M. Bosi, *et al.*, "ISO/IEC MPEG-2 advanced audio coding," *Journal of Audio Engineerng Society*, vol. 45, pp. 789–814, October 1997.

[5] "The MPEG audio web page." http://www.tnt.uni-hannover.de/project/mpeg/audio/.

[6] P. Prandoni and M. Vetterli, "Optimal bit allocation with side information," *Proc. of ICASSP*, pp. 2411–14, 1999.

[7] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, ch. 8, pp. 226–228. Kluwer Academic, 1992.

[8] A. Aggarwal, *et al.*, "Trellis-based optimization of MPEG-4 advance audio coding," *Proc. IEEE Workshop on Speech Coding*, pp. 142–4, 2000.

[9] J. D. Jonhston, "Transform coding of audio signals using perecptual noise criteria," *IEEE J. Selected Areas in Comm.*, vol. 6, pp. 314–323, February 1988.