Trellis-Based Approaches to Rate-Distortion Optimized Audio Encoding

Vinay Melkote, *Student Member, IEEE*, and Kenneth Rose, *Fellow, IEEE*

***Abstract*—Many important audio coding applications, such as streaming and playback of stored audio, involve offline compression. In such scenarios, encoding delays no longer represent a major concern. Despite this fact, most current audio encoders constrain delay by making encoding decisions on a per frame basis. This paper is concerned with delayed-decision approaches to optimize the encoding operation for the entire audio file. Trellis-based dynamic programming is used for efficient search in the parameter space. A two-layered trellis effectively optimizes the choice of quantization and coding parameters within a frame, as well as window decisions and bit distribution across frames, while minimizing a psychoacoustically relevant distortion measure under a prescribed bit-rate constraint. The bitstream thus produced is standard compatible and there is no additional decoding delay. Objective and subjective results indicate substantial gains over the reference encoder.**

***Index Terms*—Audio compression, optimization, rate-distortion, trellis, window switching.**

I. Introduction

Audio compression has been fundamental to the success of many applications including streaming of music over the internet and handheld music playback devices. Digital radio and gaming audio are other relatively new applications utilizing compressed audio. Most current audio coding techniques use psychoacoustic criteria to discard perceptually irrelevant information in the audio signal and achieve better compression. MPEG's Advanced Audio Coder (AAC) [1], [2], Sony's Adaptive Transform Acoustic Coder (ATRAC) [3], Lucent Technologies' Perceptual Audio Coder (PAC) [4], and Dolby's AC3 [5] are a few well known audio codecs. Descriptions of these coding techniques and general information regarding audio coding can be found in [6]. These techniques usually analyze the audio signal one frame or a small group of frames at a time and make encoding decisions on them, independently of other frames or frame-groups, thereby restricting encoding delay. Restricted encoding delay enables real-time audio coding, but for the majority of audio coding applications, including those previously mentioned, compression is performed offline. Hence, the end user decodes pre-compressed audio and is not affected by any encoding delays. Moreover, encoding is a one time procedure while the coded audio is typically decoded many times. Thus, we propose here a coding technique that exploits encoding delay to make optimal decisions over the entire audio file, rather than processing each frame independently. The generated bitstream is standard compatible and decodable by standard decoder at no additional decoding delay.

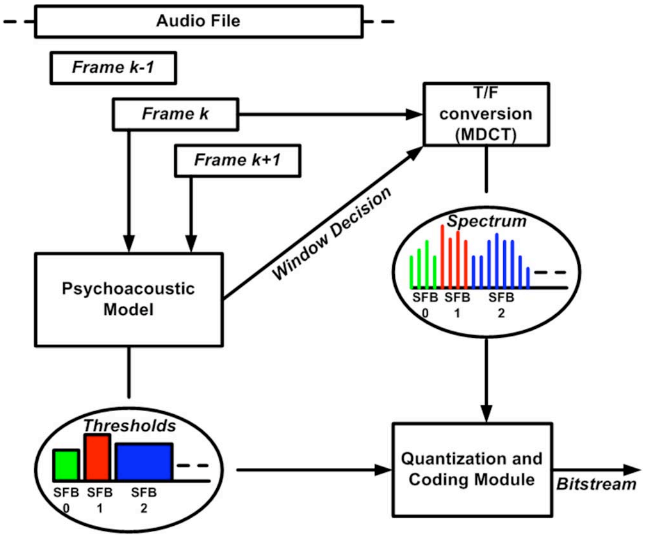

Fig. 1. Schematic of a simple AAC encoder.

As an example consider AAC (Fig. 1). The audio signal is split into overlapping frames. Depending on the stationarity of the signal, the framing is switched between a LONG window of 2048 samples and 8 SHORT windows of 256 samples each. Transition frames of suitable shape act as bridge windows between these configurations and this "window switching" decision induces a one frame encoding delay. Subsequently, a time to frequency transformation is performed on the frame. The frequency-domain coefficients are grouped into bands of unequal bandwidths to emulate the critical band structure of the human auditory system [7]. A psychoacoustic model provides masking thresholds for each of these bands, which determine the threshold of audibility of quantization noise in the bands. In AAC, a generic quantizer scaled by a parameter called the scale factor (SF) is used to quantize all the coefficients in the same band, and hence these bands are named scale factor bands (SFBs). The quantized coefficients in each SFB are then losslessly encoded using one of a prescribed set of Huffman code

Manuscript received February 05, 2009; revised June 18, 2009. First published July 24, 2009; current version published November 20, 2009. This work was supported in part by the National Science Foundation (NSF) under Grant CCF-0917230, the University of California MICRO Program, Applied Signal Technology, Inc., Cisco Systems, Inc., Dolby Laboratories, Inc., Qualcomm, Inc., and Sony Ericsson, Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gaël Richard.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: melkote@ece.ucsb.edu; rose@ece.ucsb.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2028373

books (HCBs). Encoders try to find a set of SFs and HCBs that minimize a psychoacoustic distortion measure while satisfying a bit-rate constraint for the frame. Though the target to be achieved may be a particular mean bit-rate (average across frames) or file size, the instantaneous bit-rate, i.e., for individual frames, can fluctuate around this mean. This feature is generally implemented using a bit-reservoir technique wherein rate unused by frames of low demand is "saved" for use in later frames. Optional tools such as Temporal Noise Shaping and Perceptual Noise Substitution are not discussed here.

The point to note is that the encoding procedure as described above makes decisions regarding each frame almost independently, with few minor exceptions: Due to window switching, the encoder encounters a delay of one frame to decide about transition windows. The bit-reservoir, in a limited sense, makes the encoding process dependent on past frames, but this encoding scheme, due to its constrained delay, cannot foresee the demand for bits in future frames and deliberately save bits at some cost to the current frame. The drawbacks of this encoding procedure will be discussed in detail. For now, suffice it to say that constraining the encoding delay produces a bitstream of suboptimal quality.

Thus, there is merit in increasing encoding delay to search exhaustively over all combinations of encoding parameters, and choose the optimal set, but this may be computationally daunting. AAC, for example, provides a choice of 12 HCBs and nearly 60 SFs for each SFB. There are usually 49 SFBs in the LONG configuration and 56 SFBs for the eight SHORT windows, although the exact number depends on other parameters such as sampling rate and SHORT window grouping decisions [1], [2]. Including the choice of window configurations for each frame, a conservative estimate of such complexity would be $(2 \times (60 \times 12)^{49})^N$ for an audio file of N frames, i.e., exponential in the number of SFBs and frames. So it is desirable to pursue a dynamic programming [8] based approach with a corresponding trellis to search through these choices.

It is obvious that the search for the "optimal" encoding parameters presupposes a criterion or distortion measure to compare the effects of various choices of these parameters. The most commonly used audio distortion measure is the noise-to-mask ratio (NMR) [9]–[12]—the ratio of quantization noise to masking threshold in each coding band (SFB in AAC). The distortion for a frame of audio and subsequently for the entire audio file is usually derived from the NMR. It should be noted that our methods are fairly general and could accommodate any additive distortion measure.

The problem of finding the optimal SFs and HCBs within an AAC frame (i.e., minimizing the frame distortion given a bit budget constraint) has been previously addressed in earlier work of our research group [13] and [14], under the assumption of fixed bit-rate per frame, and that all frames were in the LONG configuration. Thus, no decisions were delayed beyond the given frame. A low-complexity suboptimal alternative was proposed in [15]. A mixed integer linear programming-based solution to the same problem was proposed by Bauer and Vinton in [16] and was extended to compare window decisions per frame in [17], where window decisions were independently performed for each frame, while neglecting dependence through transition windows. Bit-reservoir optimization, using a tree structured search, was proposed in [18], without optimization of window decisions or quantization and coding parameters. Rate-distortion optimal time segmentation of audio frames have been proposed in [19]–[21] without optimization of parameters within a frame or distribution of bits across all frames.

We emphasize that we are, in fact, optimizing *all* the encoding decisions (window choice, SFs, and HCBs as well as bit budget per frame) of the aforementioned simplistic AAC encoder. The eventual results show that there are significant gains over the reference encoder in terms of both objective metrics and subjective measures such as MOS scores within the MUSHRA test framework [22], and for a variety of audio samples drawn from the EBU-SQAM database [23]. The methods proposed are of higher complexity than the reference encoder but such complexity only impacts encoding which is typically an offline operation, while the end-user does not experience any additional decoding delay. Preliminary results of this work have been reported in [24] and [25].

The organization of this paper is as follows. Section II provides a brief background to the problem. The problem within the AAC setting is formulated in Section III. The two-layered trellis solution to the problem is described in Section IV. Section V summarizes the results.

II. Background

A. MPEG Advanced Audio Coding

The implementation of the proposed approach is in the MPEG AAC setting. The high-level description of AAC given in Section I is refined here with more details for the relevant blocks.

1) Window Switching: The audio file is divided into overlapping frames and each frame is multiplied by a window. The frames are 2048 samples each in the LONG configuration [Fig. 2(a)]. If the 1024 samples in the center of the frame (between the dotted lines of Frame k in Fig. 2(a)) are nonstationary, the frame is instead encoded as a series of eight SHORT overlapped windows of 256 samples each [frame k in Fig. 2(b)] to achieve better time resolution. Adjacent LONG and SHORT windows, due to their incompatible shapes, would disrupt the perfect reconstruction properties of the transform discussed further. This is prevented by replacing the LONG window preceding a series of SHORT windows with a START window of suitable shape [Frame $k-1$ in Fig. 2(b)] and the one succeeding a SHORT window with a STOP window [Frame $k+1$ in Fig. 2(b)]. Window switching was first suggested for audio coding by Edler in [26]. Window switching decisions are usually made by the psychoacoustic model, based on heuristic thresholds of perceptual entropy [27] or transient detection [28], [29].

2) Modified Discrete Cosine Transform (MDCT): Each audio frame is transformed to the frequency domain using the forward MDCT [30]–[32]. Despite requiring overlapped frames, the MDCT is critically sampled. MDCT of a LONG (also START and STOP) frame yields 1024 transformed coefficients and 128 coefficients for each SHORT block (or 1024 total for the eight SHORT windows).

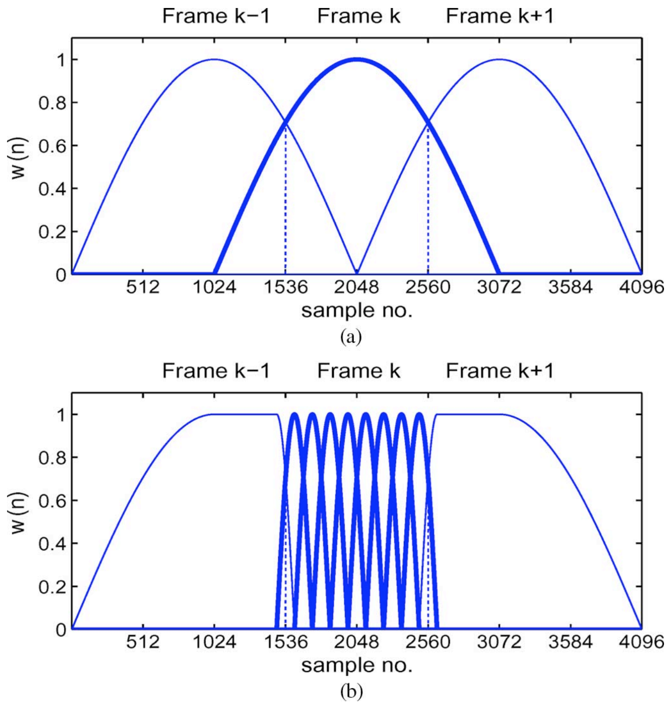

Fig. 2. Frame k in LONG and SHORT configurations and corresponding effect on neighboring LONG frames.

3) Quantization and Coding (QC) Module: The quantization and coding module receives MDCT coefficients grouped into SFBs and corresponding masking thresholds from the psychoacoustic model, selects the SFs and HCBs, and quantizes and encodes the coefficients. The difference in SF values of consecutive SFBs is encoded using a single standard specified Huffman table. The HCB values are run-length coded, i.e., a fixed number of bits is used to convey the HCB value (whenever it changes from an SFB to the next), and the number of consecutive SFBs having the same HCB. The SF and HCB bits thus consume part of the bit-rate and have to be accounted for in the rate calculation. In the MPEG Verification Model (VM) [28] the implicit rate-distortion tradeoff is accomplished using a two loop search (TLS). The TLS inner loop is a distortion loop that searches through the set of SFs for each SFB such that a near-uniform target NMR is maintained across SFBs. Once this is achieved the encoder steps into the outer, rate loop, finds the best HCBs to encode the quantized spectra and calculates the total number of bits consumed by the frame. If the rate constraint for that frame is not met the target NMR is increased (to spend fewer bits), and the inner loop executed again.

4) Bit Reservoir: AAC allows coding different frames with a different number of bits, though achieving a target average bit-rate might still be necessary. The VM implementation employs a bit-reservoir. If the QC module spends less than the available bit quota for the frame (e.g., when the frame corresponds to silence), excess bits may be used by future frames of higher demand.

B. Distortion Measure

A distortion metric for audio coding should be able to properly account for the various perceptual artifacts caused by coding. Simple measures, such as the mean squared quantization error of the spectral coefficients, ignore psychoacoustic effects, while complicated metrics such as the Perceptual Evaluation of Audio Quality (PEAQ) [33], [34], entail intractable optimization complexity. The most widely used metric is NMR [9]–[12] which divides the squared quantization error in a coding band (SFB) by the band's masking threshold.

Consider a frame of AAC whose MDCT coefficients have been grouped into L SFBs. Let e_i be the squared quantization error of the coefficients in SFB i. Let μ_i be the reciprocal of the masking threshold in the band. The NMR d_i in SFB i is given by

$$d_i = \mu_i e_i, \quad 0 \le i \le L-1. \tag{1}$$

Several variants of the frame distortion can be derived from the above definition, for example, the Total NMR (TNMR) denoted by D_T is

$$D_T = \sum_{i=0}^{L-1} d_i. \tag{2}$$

In [11]–[17] the Average NMR (ANMR), i.e., NMR averaged across SFBs has been used (clearly, $ANMR = D_T/L$). Since the number of SFBs varies for LONG and SHORT windows, TNMR is used in this work for a fair comparison between window configurations. Note that L in the SHORT configuration corresponds to the total number of SFBs of the eight SHORT windows together. Alternatively, the distortion of a frame could be defined as the Maximum NMR (MNMR) [12]–[17], D_M, across all SFBs, i.e.,

$$D_M = \max_{i=0}^{L-1} d_i. \tag{3}$$

Using the above as building blocks we can extend to consider distortion evaluation for the entire audio file (say of N frames):

$$\begin{aligned} &\text{Average TNMR (ATNMR)}: \\ &\mathcal{D}_{\mathcal{AT}} = \frac{1}{N}\sum_{k=0}^{N-1} D_T(k) && (4) \\ &\text{Maximum TNMR (MTNMR)}: \\ &\mathcal{D}_{\mathcal{MT}} = \max_{k=0}^{N-1} D_T(k) && (5) \\ &\text{Maximum MNMR (MMNMR)}: \\ &\mathcal{D}_{\mathcal{MM}} = \max_{k=0}^{N-1} D_M(k). && (6) \end{aligned}$$

$D_T(k)$ and $D_M(k)$ denote the distortion of frame k according to TNMR of (2) and MNMR of (3), respectively. It is important to note that there is no single audio distortion measure that is known to capture well, all artifacts produced by restricted bit-rate audio coding and the consideration of all the above candidates will demonstrate the generality of the proposed approach.

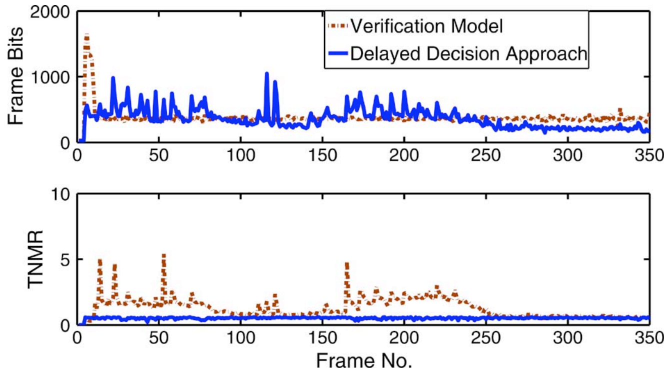

Fig. 3. Distribution of rate and distortion (TNMR) across frames when using the VM and delayed-decision based approach for glockenspiel at 16 kbps.

C. Problem Motivation and Challenges

1) Window Switching: As already mentioned, current encoders rely on heuristics to make decisions about window switching, but such decisions are not optimal in the sense of minimizing a pre-specified distortion measure. One approach (see [17] and [21]) is to design an encoder that compares the frame distortion under different window configurations and makes a window choice for that frame, but different windows encompass a different number of samples, as is evident in Fig. 2, and such comparison would not be fair. In addition, two consecutive frames cannot independently be encoded as a LONG-SHORT pair and thus, independent window choices for each frame may not form an ‘allowable’ window sequence. One could, on the other hand, compare distortion in two sequences of window decisions which start and end in the same audio samples, for instance, the LONG-LONG-LONG sequence of Fig. 2(a) and START-SHORT-STOP sequence of Fig. 2(b). This of course entails delay. This simple example provides motivation for investigating delayed decisions for window switching.

2) Bit Reservoir: The bit-reservoir of VM allows a frame to utilize bits saved (i.e., unused) in the past but cannot “borrow from the future.” Nor can it optimally borrow from the past, as the encoder cannot anticipate future needs. Some encoders, including 3GPP’s Enhanced AACplus [29] encoder, intentionally save some bits for future use by employing perceptual entropy based algorithms that specify the bit requirement for a frame. Such algorithms involve heuristic thresholds. Fig. 3 compares the effect on distortion (TNMR) due to the distribution of bit resource according to VM versus MTNMR minimization by the delayed-decision approach discussed later. The spikes in TNMR values for VM correspond to artifacts caused by a lack of sufficient bits in nonstationary frames of the audio sample (glockenspiel). It is evident that delayed decision redistributes bits to mitigate such coding artifacts.

3) Quantization and Coding Module: TLS, as described previously, separates the calculation of rate and distortion into individual loops and does not simultaneously control them. Moreover, SFs for consecutive SFBs are differentially encoded, and HCBs are run length encoded. Hence, selecting these parameters for each band independently is suboptimal. The trellis-based optimal parameter selection of [13] and [14] is a rate-distortion optimal alternative to TLS, but the procedure there was based on the assumption that the bit-rate for each frame was fixed. Modifications are necessary to incorporate this trellis into a system that relies on delayed decisions for distributing bits to frames. Another limiting assumption was that all windows were encoded in the LONG configuration. Modifications are also necessary to jointly deal with eight SHORT frames.

III. Joint Selection of Encoding Parameters: Problem Formulation

We describe here the problem formulation in the AAC setting.

A. Problem Setting

Consider an audio file of N frames. Frame k $(0 \leq k \leq N-1)$ is associated with a window configuration w_k from the set $\{\mathrm{LONG}, \mathrm{START}, \mathrm{SHORT}, \mathrm{STOP}\}$. The number of SFBs L_k in frame k depends on the window configuration. In the SHORT configuration, L_k corresponds to the number of SFBs of all eight SHORT windows. SFB i of frame k is associated with a scalefactor s_i^k and Huffman code book h_i^k $(0 \leq i \leq L_k - 1)$. Parameters s_i^k and h_i^k take value in finite sets of SF and HCB choices as prescribed in the AAC standard. Thus the intra-frame decisions produce L_k-tuples $S_k = (s_0^k, \ldots, s_{L_k-1}^k)$ and $H_k = (h_0^k, \ldots, h_{L_k-1}^k)$. All the above encoding parameters for a frame are summarized in $P_k = (w_k, S_k, H_k)$. Additionally, we denote by X_k the segment of 2048 audio samples encompassed by frame k in the LONG configuration. Clearly, other window configurations use a subset of X_k.

The number of bits of information representing frame k depends on the actual samples it contains and the choice of encoding parameters and is, hence, denoted by $B(X_k, P_k)$. An average rate constraint $\mathcal{R}$ is imposed on the encoding process, i.e.,

$$\frac{1}{N}\sum_{k=0}^{N-1} B(X_k, P_k) \leq \mathcal{R}. \tag{7}$$

The window decisions sequence is also constrained so that a START window is always used when transitioning from a LONG to a SHORT window, and a STOP window is inserted between SHORT and LONG windows. These conditions will be referred to as the ***Window Switching Constraints***.

B. Rate and Distortion Calculation

The information, in the bitstream, about SFB i of frame k can be summarized as follows.

- We denote by $\mathcal{Q}(X_k, w_k, s_i^k, h_i^k)$ the number of bits needed to encode the spectral coefficients in SFB i, as it naturally depends on the audio samples in the frame X_k in addition to the quantizer (scalefactor s_i^k), the Huffman code book h_i^k, and the window choice w_k (which influences the transform applied on X_k and hence the unquantized spectral coefficient values).
- The scalefactor s_i^k is transmitted as $s_i^k - s_{i-1}^k$. Therefore, the scalefactor bits for SFB i can be written as $\mathcal{E}(s_{i-1}^k, s_i^k)$ (with $s_{-1}^k = 0$).
- The run-length encoding of HCBs produces a fixed number of bits to indicate the run-length whenever $h_i^k \neq h_{i-1}^k$ and 0 bits otherwise. Thus the number of HCB information bits for SFB i is of the form $\mathcal{F}(h_{i-1}^k, h_i^k)$ (with $h_{-1}^k \neq h_0^k$).

Additionally, the encoder conveys the window configuration using $\mathcal{G}(w_k)$ bits. Thus, the total number of bits to encode the frame with parameters P_k can be enumerated as

$$B(X_k, P_k) = \mathcal{G}(w_k) + \sum_{i=0}^{L_k-1} \left\{ \mathcal{Q}\left(X_k, w_k, s_i^k, h_i^k\right) + \mathcal{E}\left(s_{i-1}^k, s_i^k\right) + \mathcal{F}\left(h_{i-1}^k, h_i^k\right) \right\} \quad (8)$$

where the number of SFBs L_k depends on w_k.

The psychoacoustic model produces a masking threshold for each SFB of a frame by analyzing it in the frequency domain. Thus, the weight μ_i in (1) is a function of the audio signal X_k and the transform (and hence w_k) used for time to frequency conversion. Similarly, the squared quantization error e_i depends on the quantizer (i.e., scalefactor s_i^k) and the unquantized transform coefficients. Thus, using (1), the distortion d_i in SFB i of frame k can be represented as

$$d_i\left(X_k, w_k, s_i^k\right) = \mu_i(X_k, w_k) e_i\left(X_k, w_k, s_i^k\right). \quad (9)$$

The above definition of d_i is subsequently used in (2) or (3) to obtain the frame distortion. In either case we employ the generic notation $D(X_k, P_k)$, where it is clear from the context whether $D_T(k)$ or $D_M(k)$ is in use. The distortion of the entire file is then obtained from (4)–(6). Let the encoding parameter set for the entire file be $\mathcal{P} = (P_0, \ldots, P_{N-1})$, while $\mathcal{X}$ represents the entire audio signal itself. The overall distortion, therefore, can be denoted as $\mathcal{D}(\mathcal{X}, \mathcal{P})$, and the overall bit consumption is given by

$$\mathcal{B}(\mathcal{X}, \mathcal{P}) = \sum_{k=0}^{N-1} B(X_k, P_k). \quad (10)$$

Note that H_k is specified in P_k and needed to determine the rate, but it plays no role in determining the value of $D(X_k, P_k)$, as is evident from (9).

C. Problem Definition

Find the parameter set $\mathcal{P}^*$ that minimizes the overall distortion, i.e.,

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \quad \mathcal{D}(\mathcal{X}, \mathcal{P}) \quad (11)$$

subject to the rate constraint $(1/N)\mathcal{B}(\mathcal{X}, \mathcal{P}) \leq \mathcal{R}$ and the window switching constraints of Section III-A.

Depending on the choice of definition of $\mathcal{D}(\mathcal{X}, \mathcal{P})$ from (4)–(6) we have three different problems which will be referred to as the ***ATNMR***, ***MTNMR*** and ***MMNMR problems***, respectively.

IV. Optimization With a Two-Layered Trellis

A. Minimizing Average Overall Distortion

We address here the problem of minimizing the average distortion of the file

$$\mathcal{D}(\mathcal{X}, \mathcal{P}) = \frac{1}{N} \sum_{k=0}^{N-1} D(X_k, P_k) \quad (12)$$

given the rate constraint (7). Note that if $D(X_k, P_k)$ is defined as TNMR (2) then $\mathcal{D}(\mathcal{X}, \mathcal{P})$ would be ATNMR (4). The above problem is similar to the classical problem of minimizing average distortion of quantizers given a rate constraint. The problem was originally addressed for independent quantizers in [35] and later for dependent quantizers in [36] using a Langrangian based iterative procedure. The constrained optimization problem is converted to that of minimizing the Lagrangian cost

$$\mathcal{J}_A(\mathcal{X}, \mathcal{P}) = \mathcal{D}(\mathcal{X}, \mathcal{P}) + \lambda \frac{1}{N} \mathcal{B}(\mathcal{X}, \mathcal{P}) \quad (13)$$

where λ is the Lagrange parameter. Rewriting (13) as a summation over frames we obtain

$$\mathcal{J}_A(\mathcal{X}, \mathcal{P}) = \sum_{k=0}^{N-1} J_A(X_k, P_k) \quad (14)$$

where

$$J_A(X_k, P_k) = \frac{1}{N} \left\{ D(X_k, P_k) + \lambda B(X_k, P_k) \right\} \quad (15)$$

is the contribution of a particular frame to the Lagrangian cost. Minimization of $\mathcal{J}_A(\mathcal{X}, \mathcal{P})$ for a specific value of λ yields an operating point on the rate-distortion curve. One may adjust λ and re-optimize until the rate constraint is satisfied, to obtain the choice of parameters $\mathcal{P}^* = (P_0^*, \ldots, P_{N-1}^*)$ that minimize the distortion in (12) under the constraint (7). Note that $J_A(X_k, P_k)$, the Lagrangian cost for frame k, is independent of encoding decisions P_l, $l \neq k$ and therefore,

$$\min_{\mathcal{P}} \mathcal{J}_A(\mathcal{X}, \mathcal{P}) = \sum_{k=0}^{N-1} \min_{P} J_A(X_k, P) \quad (16)$$

where $P = (w, S, H)$ is a generic point in the encoding parameter space for a single frame. Thus, for a given value of λ, the overall minimization problem seems separable into N intra-frame minimization problems. Note, however, that $P_k = (w_k, S_k, H_k)$ depends on the window choice. Independent minimization of $J_A(X_k, P_k)$ over all window choices may violate the window switching constraints and yield incompatible windows for neighboring frames, as discussed in Section II-C1. To circumvent this difficulty we define the minimum frame Lagrangian for a given window configuration w as

$$J_k^*(w) = \min_{S,H} J_A(X_k, \{w, S, H\}), \quad \forall w \in \{\mathrm{LONG}, \mathrm{START}, \mathrm{SHORT}, \mathrm{STOP}\} \quad (17)$$

The dependence of $J_k^*(\cdot)$ on X_k is implicit in the subscript k. The above minimization which will henceforth be referred to as the **Intra-frame Minimization Problem I** is discussed in Section IV-C. Assume for now that for every frame k the above minimum cost $J_k^*(w)$, the minimizing parameters $S_k^*(w)$

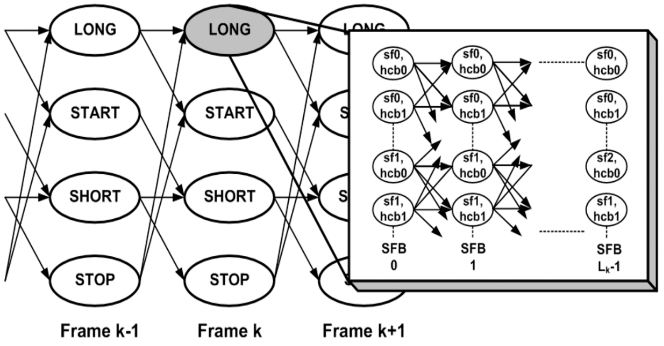

Fig. 4. **Two-Layered Trellis:** The *Window Switching Trellis* (or *Outer Trellis*) runs across frames, with states as window choices. The *Inner Trellis* (in the inset) spans across SFBs and is used in each node of the Outer Trellis to find the best intra-frame parameters.

and $H_k^*(w)$, corresponding distortion $D_k^*(w)$ and frame bit consumption $B_k^*(w)$ have been calculated for every window configuration w. The overall cost $\mathcal{J}_A$ is, therefore, minimized by the window decisions $w_0^*, \ldots, w_{N-1}^*$ given by

$$\left(w_0^*, \ldots, w_{N-1}^*\right) = \arg \min_{(w_0, \ldots, w_{N-1})} \sum_{k=0}^{N-1} J_k^*(w_k) \tag{18}$$

with $(w_0, \ldots, w_{N-1})$ obeying the window switching constraints (Section III-A). The search complexity of the above problem can be reduced drastically while simultaneously imposing these constraints by using a *trellis*-based search, such as the Viterbi algorithm [37], [38]. A trellis (the Outer Trellis in Fig. 4) is constructed with stages corresponding to frames and nodes to window choices per frame. Transitions are allowed only between compatible window choices, e.g., LONG to LONG, LONG to START, etc. Each node is associated with a specific window decision w and is populated with corresponding quantities $J_k^*(w)$, $S_k^*(w)$, $H_k^*(w)$, $D_k^*(w)$, and $B_k^*(w)$. The solution $w_0^*, \ldots, w_{N-1}^*$ to (18) is the path $(w_0, \ldots, w_{N-1})$ through the trellis that minimizes the total cost $\sum_{k=0}^{N-1} J_k^*(w_k)$ along that path.

To formally implement the window switching constraints, associate the window configurations LONG, START, SHORT, and STOP with the numbers 1–4, respectively. We denote by $\mathbf{W}_m$, $1 \leq m \leq 4$, the set of window choices which could precede the window choice m. For example, $\mathbf{W}_1 = \{1, 4\}$—a LONG window can only be preceded by a LONG or STOP window. The path of minimum cost is found as follows.

Outer Trellis Algorithm

```
1) Initialize. For $1 \leq m \leq 4$, set partial sum $\Upsilon(m) = J_0^*(m)$.
   Set counter $k = 1$.
2) Search. For $1 \leq m \leq 4$, in stage $k$, find back pointer
   $\Psi_k(m) = \arg\min_{n \in \mathbf{W}_m} \Upsilon(n)$.
3) Update. For $1 \leq m \leq 4$, set partial sum
   $\Upsilon(m) = \Upsilon(\Psi_k(m)) + J_k^*(m)$.
4) Next Stage. Increment $k$. If $k < N$ go to step 2.
5) Backtrack. Winning path ends in $w_{N-1}^* =$
   $\arg\min_{1 \leq m \leq 4} \Upsilon(m)$. Set $k = N - 1$. While $k \neq 0$, do
   $\{w_{k-1}^* = \Psi_k(w_k^*), k = k - 1\}$.
```

At each stage, only four paths survive and the complexity of this search is linear in N. As is evident, the trellis search naturally incorporates the window switching constraints, hence the name Window Switching Trellis. It is also called the Outer Trellis to differentiate from the Inner Trellis (inset of Fig. 4) that will be used to solve (17). If the rate $(1/N)\sum_{k=0}^{N-1} B_k^*(w_k^*)$ associated with the winning path does not satisfy the rate constraint (7), λ is adjusted, the minimization of (17) redone for each frame and in all window configurations, the outer trellis repopulated, and the above search repeated. When the rate constraint is met the decisions associated with the winning path are the optimal decisions minimizing the overall distortion given by (12).

B. Minimizing Maximum Overall Distortion

Here

$$\mathcal{D}(\mathcal{X}, \mathcal{P}) = \max_{k=0}^{N-1} D(X_k, P_k). \tag{19}$$

Depending on whether $D(X_k, P_k)$ is defined according to TNMR (2) or MNMR (3), the resulting $\mathcal{D}(\mathcal{X}, \mathcal{P})$ will be either MTNMR (5) or MMNMR (6). A Lagrangian solution is not applicable here due to the min-max nature of the problem. Nevertheless, a trellis-based approach offers an effective means to find the solution. Let parameter γ specify the maximum overall distortion

$$\begin{aligned} &\mathcal{D}(\mathcal{X}, \mathcal{P}) \leq \gamma \\ &\Rightarrow D(X_k, P_k) \leq \gamma, \ 0 \leq k \leq N - 1. \end{aligned} \tag{20}$$

We now find the set of encoding parameters $\mathcal{P}^*$ that minimizes the total rate $(1/N)\mathcal{B}(\mathcal{X}, \mathcal{P})$ subject to the above distortion constraint, i.e., the cost function to be minimized is

$$\begin{aligned} \mathcal{J}_M(\mathcal{X}, \mathcal{P}) &= \frac{1}{N}\mathcal{B}(\mathcal{X}, \mathcal{P}) \\ &= \sum_{k=0}^{N-1} J_M(X_k, P_k) \end{aligned} \tag{21}$$

where $J_M(X_k, P_k) = (1/N)B(X_k, P_k)$ is the corresponding cost function for frame k. If the rate thus found exceeds the rate constraint in (7), γ can be increased (allow more distortion in each frame) and the minimization repeated. Thus, we now iterate over γ, similar to the iteration over λ in Section IV-A. We can again split the overall minimization into N separate minimizations as follows:

$$\min_{\substack{\mathcal{P} \ s.t. \\ \mathcal{D}(\mathcal{X}, \mathcal{P}) \leq \gamma}} \mathcal{J}_M(\mathcal{X}, \mathcal{P}) = \sum_{k=0}^{N-1} \min_{\substack{P \ s.t. \\ D(X_k, P) \leq \gamma}} J_M(X_k, P) \tag{22}$$

where we have used (20). The window switching constraints again forbid independent minimization. Thus, the corresponding minimum cost for a frame in window configuration w is defined as

$$J_k^*(w) = \min_{\substack{S,H\ s.t.\\ D(X_k,P)\leq\gamma}} J_M(X_k, \{w, S, H\}) \quad \forall w \in \{\text{LONG}, \text{START}, \text{SHORT}, \text{STOP}\} \tag{23}$$

The above minimization is referred to as **Intra-frame Minimization Problem II** and will be discussed in Section IV-D which derives the optimal cost $J_k^*(w)$ and corresponding $S_k^*(w)$, $H_k^*(w)$, $D_k^*(w)$, and $B_k^*(w)$ for populating the Window Switching Trellis. The Outer Trellis Algorithm of Section IV-A finds the best path (decisions) through the trellis. The rate can be adjusted by varying γ, repeating the minimization of (23), repopulating the trellis, and finding the winning path again.

It should be noted that in Sections IV-A and IV-B the best path is decided at the end of the Window Switching Trellis, thereby clearly implementing delayed decisions. Additional delay is due to iterations over λ or γ values, but such delay can be substantially contained by complexity reduction techniques to be discussed later.

C. *Intra-Frame Minimization Problem I*

In Section IV-A we assumed that the solution to (17) is available. The problem is rewritten here in equivalent form: for frame k, in a specific window configuration w, we need to find

$$\{S_k^*(w), H_k^*(w)\} = \arg\min_{S,H} \{D(X_k, \{w, S, H\}) + \lambda B(X_k, \{w, S, H\})\}. \tag{24}$$

The solution entails a search over all possible combinations of SFs and HCBs, a space whose cardinality is exponential in the number of SFBs. Based on [13] and [14], $S_k^*(w)$ and $H_k^*(w)$ can be obtained in a computationally efficient manner when the frame distortion $D(X_k, P)$ is defined as TNMR or MNMR calculated over the SFBs. In the former case we specifically write

$$D(X_k, P) = \sum_{i=0}^{L_k-1} d_i(X_k, w, s_i). \tag{25}$$

This in conjunction with (8) and (24) and noting that $\mathcal{G}(w_k)$ of (8) is independent of S_k and H_k yields

$$\{S^*(w), H^*(w)\} = \arg\min_{S,H} \sum_{i=0}^{L-1} \{d_i(w, s_i) + \lambda (\mathcal{Q}(w, s_i, h_i) + \mathcal{E}(s_{i-1}, s_i) + \mathcal{F}(h_{i-1}, h_i))\} \tag{26}$$

where the frame index k is implicit and the dependence on the deterministic audio segment X_k has been omitted to simplify notation. The above minimization can be realized using the Inner Trellis of Fig. 4 which has SFBs as stages and states corresponding to combination of SF and HCB values. Thus, each state of stage i (SFB i) can be indexed by an ordered pair (u, v) denoting $s_i = u$ and $h_i = v$, associated with distortion $d_i(w, u)$ and quantization bits $\mathcal{Q}(w, u, v)$. A transition from state (u', v') in stage $i-1$ to state (u, v) in stage i is associated with the rate costs $\mathcal{E}(u', u)$ and $\mathcal{F}(v', v)$ to encode (s_i, h_i). A path through this trellis corresponds to SF and HCB sequences S and H, respectively. We seek the path that minimizes the cost in (26). We define the cost for a node (u, v) in stage i as

$$\Pi_i(u, v) = d_i(w, u) + \lambda\mathcal{Q}(w, u, v) \tag{27}$$

and for transition (u', v') of stage $i-1$ to (u, v) of stage i as

$$\Delta_i((u', v') \rightarrow (u, v)) = \lambda (\mathcal{E}(u', u) + \mathcal{F}(v', v)). \tag{28}$$

The path of minimum cost is found as follows.

Inner Trellis Algorithm

1) *Initialize*. $\forall(u, v)$ partial cost

$$\Gamma(u, v) = \Pi_0(u, v) + \Delta_0((u', v') \rightarrow (u, v))$$

with $u' = 0$ and $v' \neq v$ being forced (Section III-B). Set $i = 1$.

2) *Search*. $\forall(u, v)$ of stage i find back pointers

$$\Theta_i(u, v) = \arg\min_{(u',v')\ \text{in stage}\ i-1} \{\Gamma(u', v') + \Delta_i((u', v') \rightarrow (u, v))\}$$

3) *Update*. $\forall(u, v)$ update partial cost

$$\Gamma(u, v) = \Gamma(\Theta_i(u, v)) + \Delta_i(\Theta_i(u, v) \rightarrow (u, v)) + \Pi_i(u, v)$$

4) *Next Stage*. Increment i. If $i < L$ go to step 2.

5) *Backtrack*. Winning path ends in

$$(s_{L-1}^*, h_{L-1}^*) = \arg\min_{(u,v)\ \text{in stage}\ L-1} \Gamma(u, v)$$

Set $i = L - 1$. While $i \neq 0$, do

$$\{(s_{i-1}^*, h_{i-1}^*) = \Theta_i(s_i^*, h_i^*), i = i - 1\}$$

In step 2 of the above algorithm, only one path into any state survives and thus after each stage there are as many paths as states. Hence, the complexity of the above algorithm is linear in the number of SFBs. The algorithm when performed for frame k in window configuration w gives the best SF and HCB sequence $S_k^*(w)$, $H_k^*(w)$ in (24), and corresponding distortion $D_k*(w)$. The cost and rate associated with the winning path in the above algorithm, in conjunction with the contribution from $\mathcal{G}(w)$ of (8) give $B_k^*(w)$ and $J_k^*(w)$ of (17) used in the outer trellis of Section IV-A.

ATNMR solution: Using the above algorithm in tandem with Section IV-A we can now enumerate a *Two-Layered Trellis*-based solution to the ATNMR problem (Section III-C):

1) *Initialize*. Select a value of Lagrangian parameter λ.
2) *Inner Trellis*. For each frame k and in each window configuration w, using the Inner Trellis Algorithm and node and transition costs as defined in (27) and (28),

respectively, find $S_k^*(w)$, $H_k^*(w)$, $D_k^*(w)$, $J_k^*(w)$, and $B_k^*(w)$ and populate the outer trellis.

3) *Outer Trellis*. Using the Outer Trellis Algorithm find the best window decisions $w_0^*, \ldots, w_{N-1}^*$ and consequently $P_k^* = (w_k^*, S_k^*(w_k^*), H_k^*(w_k^*)) \forall k$, overall rate $\mathcal{B}(\mathcal{X}, \mathcal{P}^*)$, and distortion $\mathcal{D}(\mathcal{X}, \mathcal{P}^*)$.

4) *Iterate*. Check rate $\mathcal{B}(\mathcal{X}, \mathcal{P}^*)$ against rate constraint. If satisfied go to step 5 else change λ and go to step 2.

5) *Encode*. Use the optimal parameter set $\mathcal{P}^*$ to encode the audio file.

D. Intra-Frame Minimization Problem II

We address here the minimization problem in (23), i.e., for frame k, in window configuration w_k

$$\{S_k^*(w), H_k^*(w)\} = \arg \min_{\substack{S,H \\ D(X_k,P) \leq \gamma}} B\left(X_k, \{w, S, H\}\right). \quad (29)$$

As in Section IV-C, a computationally efficient minimization is possible if the frame distortion $D(X_k, P)$ is in the form of sum or maximum of SFB distortions. We describe the solution here for the maximum case, i.e.,

$$D(X_k, P) = \max_{i=0}^{L_k-1} d_i(X_k, w, s_i). \quad (30)$$

Combined with the distortion constraint in (29) it implies that

$$d_i(X_k, w, s_i) \leq \gamma, \qquad \forall i. \quad (31)$$

Using (8) and (31), we can now rewrite (29) as

$$\{S^*(w), H^*(w)\} = \arg \min_{\substack{S,H \\ d_i(w,s_i) \leq \gamma \ \forall i}} \sum_{i=0}^{L-1} \left\{ \begin{array}{l} \mathcal{Q}(w, s_i, h_i) \\ +\mathcal{E}(s_{i-1}, s_i) \\ +\mathcal{F}(h_{i-1}, h_i) \end{array} \right. \quad (32)$$

where, as usual, we omit index k, the dependence on X_k, and the term $\mathcal{G}(w)$. We use the same inner trellis as in Section IV-C to perform the minimization of (32) but the node and transition costs (27), (28) are redefined as

$$\Pi_i(u, v) = \begin{cases} \mathcal{Q}(w, u, v), & \text{if } d_i(w, u) \leq \gamma \\ \infty, & \text{otherwise} \end{cases} \quad (33)$$

$$\Delta_i\left((u', v') \rightarrow (u, v)\right) = \mathcal{E}(u', u) + \mathcal{F}(v', v). \quad (34)$$

The Inner Trellis Algorithm described in Section IV-C can be subsequently used to find $S_k^*(w), H_k^*(w)$ of (29), the corresponding distortion $D_k^*(w)$ as well as the rate cost of the winning path. This, along with $\mathcal{G}(w)$ of (8) gives the minimum cost $J_k^*(w)$ of (23) and can be used in the outer trellis method of Section IV-B.

MMNMR solution: We can now solve the MMNMR problem using the above algorithm and the method described in Section IV-B, in a *Two-Layered Trellis* framework.

1) *Initialize*. Select a value of the maximum distortion parameter γ.

2) *Inner Trellis*. For each frame k and in each window configuration w, using the Inner Trellis Algorithm with node and transition costs of (33) and (34), find $S_k^*(w)$, $H_k^*(w)$, $D_k^*(w)$, $J_k^*(w)$, and $B_k^*(w)$ and populate the outer trellis.

3) *Outer Trellis*. Using the Outer Trellis Algorithm find the optimal window decisions $w_0^*, \ldots, w_{N-1}^*$ and consequently $P_k^* = (w_k^*, S_k^*(w_k^*), H_k^*(w_k^*)) \forall k$, overall rate $\mathcal{B}(\mathcal{X}, \mathcal{P}^*)$, and distortion $\mathcal{D}(\mathcal{X}, \mathcal{P}^*)$.

4) *Iterate*. Check rate $\mathcal{B}(\mathcal{X}, \mathcal{P}^*)$ against the rate constraint. If satisfied go to step 5 else change γ suitably and go to step 2.

5) *Encode*. Use decisions $\mathcal{P}^*$ to encode the audio file.

The MTNMR problem, a hybrid of maximum and cumulative distortions, requires the solution of (23) but with the frame distortion $D(X_k, P)$ being the sum (TNMR) of SFB distortions. Therefore, (23) can be seen as equivalent to finding parameters that minimize the rate $B(X_k, P)$ given a constraint on a cumulative distortion criterion. This is a dual of the problem where the rate for a frame is fixed and parameters that minimize average (or total) distortion have to be found [13]–[17] and can still be solved using the Lagrangian approach described in Section IV-C.

MTNMR solution:

1) *Initialize*. Select a value of the maximum distortion parameter γ.

2) *Inner Trellis*. For each frame k and in each window configuration w do the following.

   a) Select a value of intra-frame Lagrangian parameter λ_{inner}.

   b) Using the Inner Trellis Algorithm with cost definitions (27) and (28) and setting $\lambda = \lambda_{\text{inner}}$ find $S_k^*(w)$, $H_k^*(w)$, $D_k^*(w)$, $J_k^*(w)$, and $B_k^*(w)$.

   c) Check $D_k^*(w)$ against γ. If satisfied go to step (d) else change λ_{inner} and go to step (a).

   d) Populate the corresponding outer trellis node with $S_k^*(w)$, $H_k^*(w)$, $D_k^*(w)$, $J_k^*(w)$, and $B_k^*(w)$.

3) *Outer Trellis*. Using the Outer Trellis Algorithm find the best window decisions $w_0^*, \ldots, w_{N-1}^*$, $P_k^* = (w_k^*, S_k^*(w_k^*), H_k^*(w_k^*)) \forall k$, overall rate $\mathcal{B}(\mathcal{X}, \mathcal{P}^*)$, and distortion $\mathcal{D}(\mathcal{X}, \mathcal{P}^*)$.

4) *Iterate*. Check rate $\mathcal{B}(\mathcal{X}, \mathcal{P}^*)$ against the rate constraint. If satisfied go to step 5 else change γ suitably and go to step 2.

5) *Encode*. Use decisions $\mathcal{P}^*$ to encode the audio file.

Note: If γ, the allowed distortion in each frame, is too small, it is possible that no choice of parameter sets S and H achieves it, i.e., the parameter space for the minimization in (23) could be a null set for certain frames in particular window configurations w. In such a case, $D_k^*(w)$ in step 2(c) of above algorithm will not be less than γ for any value of λ_{inner} and, unless fixed, results in an infinite loop. This pathology can be avoided by including an appropriate exit condition in the program. For example, it is easily seen that a low value of λ_{inner} favors decreasing distortion $D_k^*(w)$ at the cost of increasing rate $B_k^*(w)$. So λ_{inner} could be bound to be greater than a minimum value ζ. If the distortion $D_k^*(w) > \gamma$ in step 2(c) even if $\lambda_{\text{inner}} = \zeta$, then a forced exit is made from step 2(c) with the cost $J_k^*(w)$ being explicitly set to ∞.

E. Modifications for SHORT Configuration

The SHORT window configuration requires some modifications to the inner trellis design of [13] or [14]. The eight SHORT windows in the frame must be encoded jointly, i.e., the QC module (the inner trellis) analyzes the SFBs of all eight windows and jointly determines their SFs and HCBs. Let L_s denote the number of SFBs per SHORT window. The AAC bitstream format dictates that the information regarding the L_s SFBs of the first SHORT window appear first, followed by that of the second and so on. Note that both differential encoding of SFs and run length encoding of HCBs requires the imposition of ordering on the SFBs. The AAC standard allows differential encoding of SFs across SHORT window boundaries within a frame (e.g., the SF of the first SFB in the second SHORT window may be encoded as a difference from that of the last SFB in the first SHORT window), but it restricts run length coding of HCBs from extending beyond the SHORT window boundary. Therefore, the inner trellis has $8L_s$ stages, corresponding to the SFBs of all eight SHORT windows. Transition costs [(28), (34)] which straddle across SFBs of two adjacent SHORT windows are allowed the usual SF contribution of $\mathcal{E}(s_{i-1}, s_i)$ but artificially forced to have a nonzero $\mathcal{F}(h_{i-1}, h_i)$ contribution even if $h_{i-1} = h_i$ (see Section III-B).

Additionally, the AAC standard allows "grouping of SHORT windows" where the encoder can identify consecutive SHORT windows within a frame with similar characteristics and interleave their spectra into a shared set of SFBs [1], [2]. For example, a frame of eight SHORT windows could be partitioned into three groups of two, three, and three windows. Windows in the same group share SFs and HCBs for the same SFB. This is accommodated in the inner trellis by using stages as grouped SFBs rather than individual window SFBs.

Since there are eight windows, 127 groupings are possible and the grouping choice is an additional encoding parameter in the SHORT configuration, but all of these groupings span the same number of audio samples and hence the minimizations in (17) and (23) can be performed in each grouping configuration to select the optimal grouping, and appropriately populate the SHORT node of the outer trellis.

F. Complexity Reduction

The complexity (or encoding time) can be considerably reduced via memory tradeoff. All the above methods require multiple traversals of the audio file, iterating over λ or γ, but the distortion and number of bits associated with a given state of the inner trellis do not depend on the values of these iteration parameters. Thus, concurrent computation of costs for multiple values of λ or γ can eliminate redundant effort. This is akin to maintaining parallel outer and inner trellises each running at a different value of λ or γ while sharing per state results. If a wide and finely divided range of these iteration parameters is used, the best decisions can be obtained in a single traversal of the audio file. Additionally one could also find the best decisions for a range of encoding rates, if desired. The hybrid nature of the MTNMR problem necessitates additional iterations over the inner parameter λ_{inner} to satisfy a specific distortion constraint γ. The maintenance of parallel trellises as described above helps to reuse such iterations for different values of γ.

G. Generalization to Other Codecs

The delayed decisions (beyond the frame) are implemented by the outer Window Switching Trellis. The computational efficiency of the trellis is due to the fact that, in AAC, distortion $D(X_k, P_k)$ and bit usage $B(X_k, P_k)$ for frame k are independent of encoding decisions in other frames. This characteristic is shared by many other audio codecs, including Lucent's PAC [4], Dolby's AC-3 [5], and Sony's ATRAC [3]. These codecs analyze audio samples (in the case of ATRAC, subband outputs of a very low resolution QMF) in frames and switch between different frame resolutions. As in AAC, the frames are encoded separately and share the available bit resource through heuristic allocation.

Moreover, all the above codecs employ a critical band based analysis within each frame, find quantizers (SF equivalents) for the frequency domain signal using the masking thresholds and, with the exception of AC-3, noiselessly encode the quantized spectra. Therefore, an inner trellis scheme with modified node and transition costs can be devised for these codecs.

V. Results

We describe here the experimental setup, including implementation details, and present simulation results. We first list the codecs under comparison.

1) *Reference Model (RM)*: The MPEG-4 Verification Model [28] using only the psychoacoustic model, TLS, bit-reservoir and transient detection based window switching with a restricted set of eight window grouping choices.
2) *Inner-Trellis-only models RM-TB(T) and RM-TB(M)*: use the same blocks as the RM except that greedy TLS is replaced by the trellis-based parameter selection of [13] and [14]. Modifications for SHORT windows as described in Section IV-E are used. RM-TB(T) minimizes TNMR and RM-TB(M) minimizes MNMR within a frame, given a rate constraint. They do not optimize windows and rate distribution across frames.
3) *Outer-Trellis-only models L1-AT, L1-MT, and L1-MM*: use the outer trellis to find the window decisions and bit distributions so as to minimize ATNMR, MTNMR, and MMNMR, respectively. The minimum costs in (17) and (23) have to be obtained to populate the outer trellis. Since the aim of these models is to isolate the effect of the outer trellis, a complete minimization over all possible SF and HCB sets (S, H in (17) and (23)), using the inner trellis, is not effected. Instead a modified TLS is used, in each frame and in every window configuration, as follows. TLS starts off at a low value of distortion (NMR) and corresponding high bit-rate. In subsequent iterations the target NMR is increased in fixed steps till the specified bit-rate for the frame is achieved. Thus, if the bit-rate constraint in the outer loop is set to 0, TLS passes through all of its operational rate-distortion points, each corresponding to one (S, H) pair. The minimization in (17) and (23) is effected only over this restricted set of (S, H) pairs. Thus, the

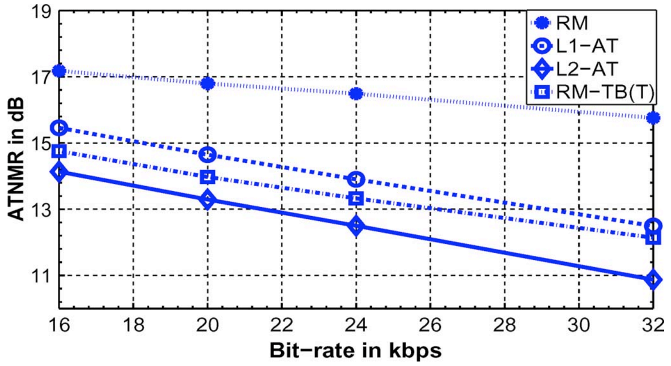

Fig. 5. Comparison of ATNMR produced by RM, RM-TB(T), L1-AT, and L2-AT at different bit-rates.

models L1-AT, L1-MT and L1-MM, by not incorporating the inner trellis, optimize pan-frame decisions but not the choice of parameters within a frame.

4) *Two-Layered Trellis-based models L2-AT, L2-MT, and L2-MM*: use the two-layered trellis-based algorithms (i.e., both inner and outer trellis) to minimize ATNMR, MTNMR, and MMNMR distortion measures, respectively, for the entire file.

At this juncture, we note that although RM, RM-TB(T), and RM-TB(M) can code different frames with a different number of bits, they are still referred to, in general parlance, as constant bit-rate (CBR) codecs. Since these codecs employ a bit-reservoir they ensure that the bitstream can be decoded in real time with constant delay when transmitted over a constant bit-rate channel. The L1- and L2-approaches (in which cases too the instantaneous bit-rate fluctuates) would on the other hand be referred to as average bit-rate (ABR) codecs as they do not employ a bit-reservoir but are still coded to achieve a target mean bit-rate. In case of these codecs, it might be necessary to buffer a larger chunk of the bitstream at the decoder before playback starts.

All the trellis-based approaches used the parallelization methods described in Section IV-F for computational efficiency. A set of ten mono, 16-bit PCM audio files sampled at 44.1 kHz, from the EBU-SQAM [23] database were used for the tests. These samples included tonal signals such as the accordion, signals with attacks such as harpsichord and glockenspiel, speech, and general pop music.

A. Objective Results

Fig. 5 compares the gains (reduction in ATNMR) over RM achieved by: optimizing decisions only across frames (L1-AT), only within frames (RM-TB(T)), and optimizing both intra- and inter-frame decisions (L2-AT). The distortion has been averaged over the ten audio samples. Overall optimization yields the best gains (3–5 dB over RM). Fig. 6 compares the performance of the corresponding encoders when the MTNMR measure is optimized. RM shows hardly any decrease in distortion as the bit-rate is increased. This is due to its suboptimal bit distribution. Most audio samples contain critical frames that require a large number of bits for transparent coding. As the bit-reservoir of RM is inefficient, the maximum distortion (MTNMR) exhibits negligible improvement with increase in average bit-rate. Note that RM-TB(T) also uses the bit-reservoir and hence L1-MT outperforms it by achieving better bit-distribution. This trend in gains is in contrast to the previous case of minimizing average overall distortion (ATNMR). Fig. 7 shows the gains when the MMNMR measure is minimized. The two-layered trellis approach (L2-MM) achieves gains of 10–12 dB over RM and about 8 dB over the single-layered trellis approaches, RM-TB(M) and L1-MM, at various bit-rates. As in the MTNMR case, the outer-trellis-only method L1-MM beats RM-TB(M) at low bit-rates thanks to efficient bit distribution across frames, but at higher bit-rates the inner-trellis-only method RM-TB(M) performs better due to its improved MNMR minimization in each frame, over the suboptimal TLS of L1-MM. Fig. 8 compares window decisions based on transient detection (RM) to that of the Window Switching Trellis (L2-MT), in case of the glockenspiel sample. Rate-distortion optimization leads to different window decisions from that of the RM.

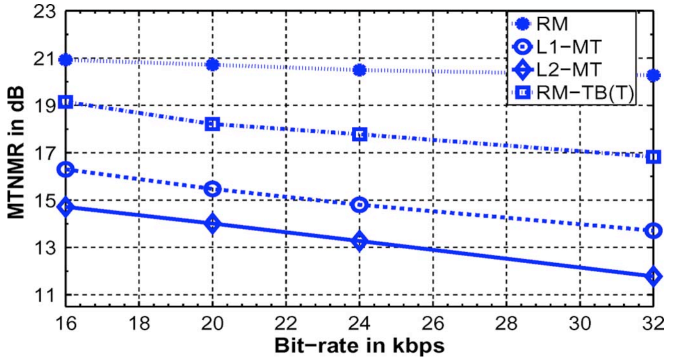

Fig. 6. Comparison of MTNMR produced by RM, RM-TB(T), L1-MT, and L2-MT at different bit-rates.

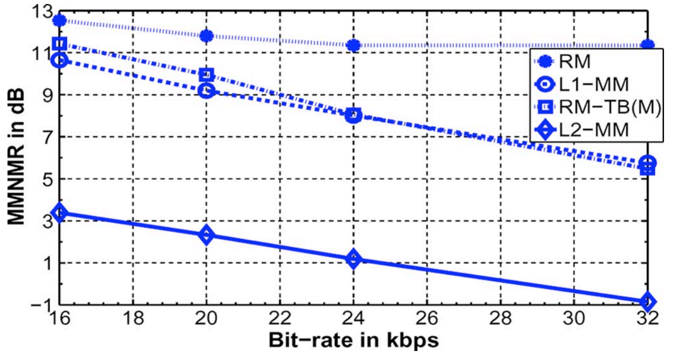

Fig. 7. Comparison of MMNMR produced by RM, RM-TB(M), L1-MM, and L2-MM at different bit-rates.

B. Subjective Evaluation

The effect of optimizing encoding decisions on subjective quality depends critically on the ability of the distortion measure to reflect psychoacoustic effects. Subjective tests indicated that minimizing the MTNMR measure improves audio quality. MUSHRA tests [22] were conducted with 20 listeners and six audio samples (tenor, harpsichord, accordion, side-drums, male German speech, and female English speech) encoded at 16 kbps. Fig. 9 shows the results of these tests. The MUSHRA

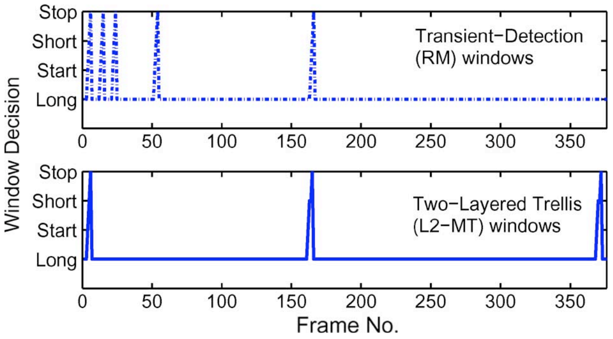

Fig. 8. Comparison of window decisions made by RM and L2-MT for the glockenspiel sample. Peaks indicate transitions to SHORT configuration.

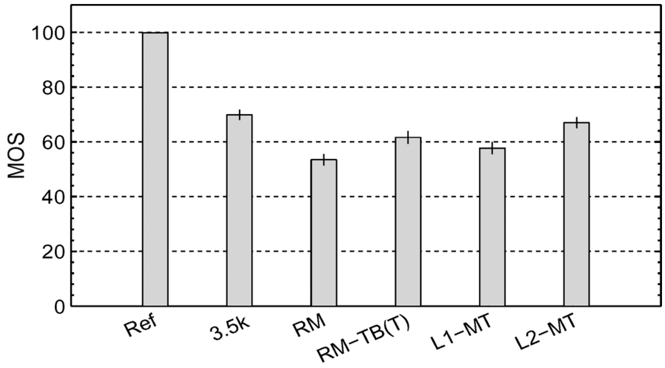

Fig. 9. Comparison of MUSHRA scores of RM, RM-TB(T), L1-MT, and L2-MT for audio encoded at 16 kbps. "Ref" represents the original audio and "3.5 k" is the low pass anchor.

scores have been averaged across samples. The two-layered trellis approach (L2-MT) has the best performance followed by RM-TB(T) and L1-MT. The reference model RM produces the worst quality of audio. Minimizing the MTNMR measure is roughly equivalent to maintaining a constant distortion (TNMR) across frames. The argument for this is as follows. If all the frames do not have the same distortion, then bits used in frames with lesser distortion can be reallocated, thus incrementally increasing distortion in these frames while reducing that in the frame with maximum distortion. This would in effect minimize the overall maximum distortion (MTNMR), but naturally tends to spread the distortion equally over the frames. This uniformity in distortion, which is evident in Fig. 3, may explain why MTNMR minimization yields improved subjective quality, as well as why ATNMR minimization was observed to compromise subjective quality. The MMNMR approach fares comparatively better in this aspect but tends to accentuate some high frequency artifacts. Note that the MMNMR approach also uses maximum overall distortion and hence maintains almost uniform distortion across frames. Additionally, it considers the maximum distortion amongst SFBs of a frame but is not guaranteed to maintain the same NMR in each SFB. This is because the different SFBs (stages of the inner trellis) are connected by nonzero transition costs, i.e., the rate for an SFB depends on the choice of parameters in the previous SFB (8). There are no such transition costs in the outer trellis. This might be a reason why this approach induces some artifacts in the high frequency regions.

TABLE I
RELATIVE FIGURES OF COMPLEXITY OF THE VARIOUS ENCODING METHODS

Encoder	Relative Complexity
RM	1
RM-TB(T,M)	30
L1-(AT,MT,MM)	15
L2-(AT,MM)	450
L2-MT	4500

It should be noted that despite the poorer quality of the ATNMR and MMNMR minimization approaches, these methods should not be dismissed. Since there is no universally precise audio distortion measure, perceptually certain types of audio may benefit from optimization in the ATNMR or MMNMR fashion.

C. Complexity

The encoding complexity of all the methods is linear in the number of frames. Therefore, we simply compare the average time to encode a frame, normalized by that of RM, to get the relative figures of complexity shown in Table I. Note that the delayed decision part of the proposed approach actually comes from the outer trellis but as the table indicates, using the outer trellis to implement better window switching and bit-distribution (i.e., the L1-approaches) is only about 15 times more complex than RM. A major contribution to the complexity of the L2-approaches is actually the inner trellis. This suggests that suboptimal intra-frame parameter selection alternatives to the inner trellis could be used to obtain low complexity delayed-decision based algorithms. One could, for example, prune the number of transitions possible from one stage of the inner trellis to the next, as suggested in [14], and thus reduce the number of paths to be compared and hence the complexity.

Another possibility, in the case of the L2-MT approach, is to linearly interpolate between rate-distortion points for a frame with distortion on the logarithmic scale to get an approximate λ_{inner} that satisfies the bit-rate constraint γ, instead of iterating over multiple values of λ_{inner} as demanded by the MTNMR solution. Such linear interpolation was observed to reduce the complexity figure of the L2-MT approach by a factor of 4 but is suboptimal (reduction in gains by 0.2 dB).

VI. CONCLUSION

In this paper, we derived a two-layered trellis-based optimization scheme for audio coding while minimizing three different overall distortion measures—ATNMR, MTNMR, and MMNMR. The trellis effectively optimizes all the encoding decisions of the reference encoder by making delayed decisions regarding each frame. The delay and one time encoding complexity do not impact the decoder, and the bitstream generated is standard compatible. Scenarios which involve offline encoding of audio may substantially benefit from this overall optimization process. Objective and subjective results in the AAC setting support such a delayed-decision-based optimization procedure.

REFERENCES

[1] *Information Technology—Generic Coding of Moving Pictures and Associated Audio*, ISO/IEC std. ISO/IEC JTC1/SC29 13818-7:1997, 1997.

[2] *Information Technology—Generic Coding of Moving Pictures and Associated Audio*, ISO/IEC std. ISO/IEC JTC1/SC29 14496-3:2005, 2005.

[3] K. Akagiri, M. Katakura, H. Yamauchi, E. Saito, M. Kohut, M. Nishiguchi, and K. Tsutsui, "Sony systems," in *Digital Signal Processing Handbook*, V. Madisetti and D. B. Williams, Eds. New York: IEEE Press, 1998.

[4] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, "The perceptual audio coder (PAC)," in *Digital Signal Processing Handbook*, V. Madisetti and D. B. Williams, Eds. New York: IEEE Press, 1998.

[5] L. D. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, and S. Vernon, "AC-2 and AC-3: Low-complexity transform-based audio coding," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Gerwin, Eds. New York: Audio Eng. Soc., 1996, pp. 54–72.

[6] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–513, Apr. 2000.

[7] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 2nd ed. New York: Springer-Verlag, 1999.

[8] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

[9] K. Brandenburg, "Evaluation of quality for audio coding at low bitrates," in *Proc. 82nd AES Conv.*, 1987, preprint 2433.

[10] K. Brandenburg and T. Sporer, "NMR and masking flag: Evaluation of quality using perceptual criteria," in *Proc. AES 11th Int. Conf.*, May 1992.

[11] H. Najafzadeh-Alaghandi and P. Kabal, "Improving perceptual encoding of narrow-band audio signals at low rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1999, vol. 2, pp. 913–916.

[12] H. Najafzadeh-Alaghandi and P. Kabal, "Perceptual bit allocation for low-rate coding of narrow-band audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jun. 2000, vol. 2, pp. 893–896.

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

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

[15] C.-H. Yang and H.-M. Hang, "Cascaded trellis-based rate-distortion control algorithm for MPEG-4 advanced audio coding," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 998–1007, May 2006.

[16] C. Bauer and M. Vinton, "Joint optimization of scale factors and huffman codebooks for MPEG-4 AAC," in *Proc. 6th IEEE Workshop. Multimedia Signal Process.*, Sep. 2004, pp. 117–189.

[17] C. Bauer, "The optimal choice of encoding parameters for MPEG-4 AAC streamed over wireless networks," in *Proc. 1st ACM Workshop. Wireless Multimedia Netw. Perf. Modeling*, Oct. 2005, pp. 93–100.

[18] E. Camberlein and P. Philippe, "Optimal bit-reservoir control for audio coding," in *Proc. IEEE Workshop. Applicat. Signal Process. Audio Acoust.*, Oct. 2005, pp. 251–254.

[19] O. A. Niamut and R. Heudsens, "R-D optimal time segmentations for the time varying MDCT," in *Proc. Eur. Signal Process. Conf. 2004*, Sep. 2004, pp. 1649–1652.

[20] O. A. Niamut and R. Heudsens, "Optimal time segmentation for overlap-add systems with variable amount of window overlap," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 665–668, Oct. 2005.

[21] J. Boehm, S. Kordon, and P. Jax, "An experimental audio coder using rate-distortion controlled temporal block switching," in *Proc. 120th AES Conv.*, May 2006, preprint 6810.

[22] *Method of Subjective Assessment of Intermediate Quality Level of Coding Systems, ITU-R Recommendation*, BS 1534-1, 2001.

[23] "EBU-SQAM database," [Online]. Available: http://www.ebu.ch/en/technical/publications/tech3000_series/tech3253/index.php

[24] V. Melkote and K. Rose, "Trellis based approach for joint optimization of window switching decisions and bit resource allocation," in *Proc. 123rd AES Conv.*, Oct. 2007, preprint 7216.

[25] V. Melkote and K. Rose, "A two-layered trellis approach to audio encoding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2008, pp. 201–204.

[26] B. Edler, "Codierung von audiosignalen mit überlappender transformation und adaptiven fensterfunktionen," *Frequenz*, vol. 43, no. 9, pp. 252–256, Sep. 1989.

[27] J. D. Johnston, "Estimation of perceptual entropy using noise masking criterion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1984, vol. 5, pp. 2524–2527.

[28] "MPEG Verification Model," [Online]. Available: http://www.standards.iso.org/ittf/PubliclyAvailableStandards/ISO_IEC_14496-5_2001_Software_Reference

[29] "3gpp HE-AAC Reference Software," [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/26410.htm

[30] H. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 6, pp. 969–978, Jun. 1990.

[31] J. P. Princen, A. W Johnson, and A. B. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1987, vol. 12, pp. 2161–2164.

[32] S. Shlien, "The modulated lapped transform, its time-varying forms and its applications to audio coding standards," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 4, pp. 359–366, Jul. 1997.

[33] W. C. Treurniet and G. A. Soulodre, "Evaluation of the ITU-R objective audio quality measurement method," *J. Audio Eng. Soc.*, vol. 48, no. 3, pp. 164–173, Mar. 2000.

[34] *Method for Objective Measurements of Perceived Audio Quality*, ITU-R Std. BS. 1387-1, Nov. 2001.

[35] Y. Shoham and A. Gersho, "Efficient bit-allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 9, pp. 1445–1453, Sep. 1988.

[36] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE. Trans. Image Process.*, vol. 3, no. 5, pp. 533–545, Sep. 1994.

[37] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 4, pp. 260–269, Apr. 1967.

[38] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.

Vinay Melkote (S'08) received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 2005 and the M.S. degree in electrical and computer engineering from the University of California, Santa Barbara (UCSB), in 2006. He is currently pursuing the Ph.D. degree in electrical and computer engineering at UCSB.

He interned in the Multimedia Codecs division of Texas Instruments (TI), India, in the summer of 2004 and was involved in the development of a JPEG decoder compatible with various TI platforms. He interned in the Audio Systems Group of Qualcomm, Inc., San Diego, CA, from June to September, 2006 and was involved in the development of MIDI hardware and audio postprocessing. His research interests include audio and speech processing/coding.

Mr. Melkote is a student member of the Audio Engineering Society. He won the Best Student Paper Award at ICASSP 2009.

Kenneth Rose (S'85–M'91–SM'01–F'03) received the Ph.D. degree from the California Institute of Technology, Pasadena, in 1991.

He then joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, where he is currently a Professor. His main research activities are in the areas of information theory and signal processing, and include rate-distortion theory, source and source-channel coding, audio and video coding and networking, pattern recognition, and non-convex optimization. He is interested in the relations between information theory, estimation theory, and statistical physics, and their potential impact on fundamental and practical problems in diverse disciplines.

Prof. Rose was corecipient of the 1990 William R. Bennett Prize Paper Award of the IEEE Communications Society, as well as the 2004 and 2007 IEEE Signal Processing Society Best Paper Awards.