Multidimensional Optimization of MPEG-4 AAC Encoding
Claus Bauer, Matt Fellers, Grant Davidson
Dolby Laboratories, San Francisco, CA, USA, {cb,mcf,gad}@dolby.com

Abstract—The subjective quality achieved by most audio codes, including MPEG-4 AAC, depends strongly on the algorithms used for encoder parameter selection. As a practical measure in conventional encoders, the overall encoding procedure is usually divided into a sequence of smaller problems that are solved heuristically. In this paper, we formulate the MPEG-4 AAC encoding problem as a multidimensional optimization procedure and present simulation results indicating performance gains relative to conventional approaches.

I. INTRODUCTION

The MPEG-4 AAC audio codec exploits perceptual redundancies to achieve transparent perceptual qualities at low bit rates. The AAC encoder produces audio frames containing either a single long block containing 1024 transform coefficients or a sequence of eight short blocks, each containing 128 transform coefficients. In general, long blocks provide higher frequency domain resolution of the signal to be coded, whereas short block sequences provide higher time domain resolution. If a short block sequence is selected for a particular audio frame, adjacent blocks can be grouped into regions of length one to eight. Within each long block, and within each short block region, the transform coefficients are combined into scale factor bands (SFBs). For each SFB, the AAC encoder allocates bits to encode the transform coefficients. The transform coefficients are quantized using a dynamically chosen scale factor (SF) to determine the quantization step size, and the quantized coefficients are entropy encoded using Huffman codebooks (HCBs). In summary, we see that the encoder must select the following sets of encoding parameters: transform block length, group configuration (for short blocks only), quantization step size, and Huffman codebooks.

These parameters are transmitted to the decoder as side information. The total transmission bit rate is a function of the number of bits needed to encode the transform coefficients and the encoding parameters. The AAC encoder selects the encoding parameters such that the total transmission rate is at or below a given threshold, while ensuring that a predefined measure of decoded signal distortion is minimized. A common perceptual quality measure is the average noise to mask ratio (ANMR) [1].

In a conventional AAC implementation, the encoder first decides if a long or a short transform block length is chosen. In [2], Johnston proposes a technology based on the concept of Perceptual Entropy [3] to decide if a specific audio frame is encoded using a long or a short transform block length. For a long block configuration, the codec must decide the quantization step sizes and the Huffman codebooks for each SFB. The optimal choice of these two parameter sets requires the solution of a complex optimization problem caused by the fact that the information for the quantization step sizes and the Huffman codebooks is encoded differentially between adjacent bands. The commonly used Two Loop Search [4] uses a heuristic approach that neglects the inter-band dependencies and thus simplifies the problem significantly by optimizing the SFs and HCBs independently for each band. A near-optimal joint optimization of the SFs and HCBs for all SFs was first proposed in [5]. Finally, a joint optimization algorithm that always finds the theoretically best solution was presented in [6].

For a short block frame, the optimization problem is even more complex because of the interdependencies of three sets of parameters: group configuration, quantization step size, and Huffman codebooks. For a given group configuration, the encoder calculates multiple masking curves, which in turn influence the selection of quantization step sizes [4]. Step size decisions then affect the Huffman codebook used for entropy coding the quantized transform coefficients. Consequently, the group configuration influences both the choice of the quantization step size and the Huffman codebook. As with long blocks, the values of the quantization step sizes and the Huffman codebooks are encoded differentially in each region.

In previous approaches, the three sets of encoding parameters are chosen sequentially. This simplifies the problem as interdependencies between the parameter sets are neglected. Moreover, in each sequential step, sub-optimal heuristics - with regard to a given metric - are typically used to decide the respective parameter setting. Conventional encoder implementations first choose the group configuration and then determine the SFs and the HCBs. Because the encoder does not know the actual SF and HCB values when choosing a configuration, at best it can only estimate both the distortion and side information cost expected to be induced by the SFs and the HCBs for a specific grouping configuration. In [7], algorithms to determine the grouping configuration of the short blocks into regions are described. The idea of these algorithms is to combine adjacent short blocks into regions with similar spectral envelopes. The SF and HCB are then chosen by the same methods mentioned above for the long block configuration.

In this paper, we develop a novel mathematical model of the complete AAC encoding problem for short blocks and present the first optimal solution algorithm to this problem. In contrast to previous approaches, we do not solve the AAC encoding problem for short blocks by sequentially choosing the three encoding parameters, but we solve the encoding problem as a three-dimensional optimization problem over all possible choices of group configurations, SFs, and HCBs. We describe our algorithm in Section III and show its performance improvements compared to previous methods in Section IV. We conclude in Section V. Throughout this paper, we refer to the AAC encoding problem for short blocks as the AAC encoding problem.

II. PROBLEM DEFINITION

For each audio frame, the spectral MDCT coefficients are divided into blocks, regions, and scale factor bands (SFBs). For a short transform block length, the coefficients are initially divided into 8 blocks which have the approximate same number of coefficients. Adjacent blocks are then grouped into regions where a region consists of between 1 and 8 adjacent blocks. Thus, there are $2^7 = 128$ possible choices of regions. Each region contains $R = 15$ SFBs. We define the variable $N$ as the number of SFBs, i.e., $N = Rq$, $1 \leq q \leq 8$, where $q$ is the number of regions.

In this document, we call a specific blocking and region structure chosen by the encoder a configuration. We number the configurations $q$ from $q = 0$ to $q = G = 127$.

Within the $i$-th band $SFB_i$, all coefficients are quantized using the same scalar quantization step size, which is controlled
by the $i$-th scale factor $s_i$ selected from a range of $M_1$ possible values, i.e., $1 \leq s_i \leq M_1$. Also within $SFB_i$, the quantized coefficients are entropy encoded using a Huffman codebook $h_i$ selected from a set of $M_2$ possible values, i.e., $1 \leq h_i \leq M_2$. We set $S = \{s_1, \ldots, s_N\}$ and $H = \{h_1, \ldots, h_N\}$. Both the $s_i$ and $h_i$ are indexes into fixed sets of Scale Factors and Huffman codebooks, respectively. We define by $d(s_i)$ the quantization error of the $i$-th scale factor band if the $i$-th scale factor is chosen equal to $s_i$. $w_{g,i}$ denotes the weight of the $i$-th scale factor band in the $g$-th configuration which is derived from psychoacoustic properties of the signal. The ANMR is expressed as

$$ANMR(S, g) := \frac{1}{N} \sum_{i=1}^{N} w_{g,i} d(s_i). \quad (1)$$

The transmission rate consists of four parts:

- The number of bits $B$ required to inform the decoder about the chosen regions is $B = (q - 1)$, where $q$ is the number of chosen regions and $c$ is a constant.
- For a given signal, $Q(s_i, h_i, g)$ denotes the bits required for the $g$-th configuration to encode the coefficients of $SFB_i$ with the SF value chosen as $s_i$ and the HCB value chosen as $h_i$.
- $F(s_{i-1}, s_i)$ expresses the number of bits required to specify the SF for $SFB_i$. As for $i \neq (1 \mod R)$, the SF are encoded differently, we set $F(s_{i-1}, s_i) := F(s_{i-1}, s_i)$. For $i = 1 \mod R$, we denote the number of required bits as $F(s_i)$.
- $G_i(H, g)$ represents the number of bits needed to encode the HCB value of $SFB_i$ in the $g$-th configuration. In order to specify $G_i(H, g)$ further, we introduce the notion of a HCB section. A HCB section is defined as a sequence of $SFB_i, a \leq i \leq b$, such that $h_{a-1} \neq h_a$ and $h_a \neq h_{a+1}$ and $h_b \neq h_{b+1}$ where possibly $a = b$. The notion of a HCB section is only defined per region, i.e., HCB sections cannot overlap regions. If the last $SFB$ of a given section and the first $SFB$ of the next section have the same HCB value, then this is not considered as a part of one HCB section, but one says that an HCB of the current section terminates at the last $SFB$ of this section and that a new HCB section starts at the first $SFB$ of the next region. The conditions $h_{a-1} \neq h_a$ or $h_a \neq h_{a+1}$ do not apply if $a$ is the first $SFB$ of a region or $b + 1$ is the last $SFB$ of a region respectively. We say that the HCB section defined in this way starts at $SFB_{a-1}$ terminates at $SFB_b$, and has length $l(a) = b - a + 1$. Based on the MPEG specification [9], there is:

$$G_i(H, g) = \begin{cases} 0, & \text{if } i \neq 1 \mod R, h_i = h_{i+1}, \\
4 + 3k, & \text{if } i \equiv 1 \mod R, l(h_i) \in I_k, \\
4 + 3k, & \text{if } i \neq 1 \mod R, h_i \neq h_{i+1}, l(h_i) \in I_k, \end{cases}$$

where $k \in \{1, 2, 3\}, I_1 = [1, 6], I_2 = [7, 13], I_3 = [14, 15]$.

- The definitions of the functions $F(\cdot)$, $F(\cdot)$, and $G(\cdot)$ show that for $i \equiv 1 \mod R$ the contribution of $SFB_i$ to the rate is calculated differently from the contributions of the $SFB_i$ with $i \neq 1 \mod R$. For $i \equiv 1 \mod R$, we set $Q(s_i, h_i, g) = Q_i(s_i, h_i, g) + F(s_i) + G_i(H, g)$.

We now define the transmission rate $R(S, H, g, B)$:

$$R(S, H, g, B) = \sum_{i=1}^{N} Q_i(s_i, h_i, g) + \sum_{i=1, i \neq 1 \mod R}^{N} \left( F(s_i, s_{i-1}) + G_i(H, g) \right) + B. \quad (2)$$

For a given rate threshold $R_\alpha$, the AAC encoding problem is then defined as follows:

**Minimize** $ANMR(S, g)$ \quad (3)

**such that** $R(S, H, g, B) \leq R_\alpha$. \quad (4)

### III. An optimal solution based on mixed integer linear programming

**A. Preliminary concepts and variables**

We recall that the maximum number of bands for any configuration $g$ is $8R$. In the following, we assume the existence of $8R$ SFBs for any configuration $g$, but we divide the $8R$ SFBs in valid SFBs which are part of the configuration $g$ and invalid SFBs, which are not a part of the configuration $g$, and which we only introduce for modeling purposes. For our model, the invalid SFBs will make a zero contribution to $ANMR(S, g)$ and $R(S, H, g, B)$ in (3) and (4). We denote by $S_{B_j}, 1 \leq j \leq 8$, the 8 short blocks and assign $R SFBs$ to each $S_{B_j}$. We introduce the binary variables $u_j, 0 \leq j \leq 7$, where always $u_0 = 1$, and say that the short blocks $S_{B_j}$ and $S_{B_{j+1}}$ belong to the same region if $u_j = 0, 1 \leq j \leq 7$. Thus, there is a one-to-one relation between all short block configurations $g$ and all possible choices of the set $u_j, 1 \leq j \leq 7$. We call the band $S_{B_j}$ valid if $u_j = 1$. Thus, for a short block configuration with $s_i, 1 \leq s_i \leq 8$ regions, we have $s$ short blocks $S_{B_{R_j+1}}, 1 \leq j \leq 8, 1 \leq a_0 < \ldots < a_j < \ldots < a_8 \leq 8$ with valid bands and the SFBs of $S_{B_{a_m}}$ correspond to the SFB of the $j$-th of the $s$ regions. The other $8 - s$ SFBs are not valid. For any configuration $g$, we denote the set of valid bands as $N_g$.

We consider the difference $a$ of the HCB values of two adjacent SFBs $i, g = a - s_{a-1}$. We define the following function:

$$G(a) = \begin{cases} 0, & a = 0, \\
7, & a \neq 0. \end{cases}$$

$$x_{m,j} = \begin{cases} 3, & u_j = 1, \text{ and } v_m \text{ HCB sections of length } l \in J_m \text{ exists,} \\
between SFB_{R_{j+1}} \text{ and } SFB_{R_{j+1}}, \\
0, & \text{else.} \end{cases}$$

for $m \in \{2, 3, 4\}, v_2 = v_4 = 1, v_3 = 2, J_2 = J_3 = [7, 13], J_4 = [14, 15]$. We note that in the definition of $x_{m,j}$, our usage of the term between includes the bands $SFB_{R_{j+1}}$ and $SFB_{R_{j+1}}$. We see that if $u_j = 0$, then the variables $x_{m,j}, 2 \leq m \leq 4, 0 \leq j \leq 7$ are equal to zero which reflects the fact the short block $S_{B_j}$ only contains invalid SFBs. For $u_j = 1$, if no HCB section of length $\geq 7$ exists, then $G_i(H, g) = G_i(h_i - h_{i+1})$ for all $R_j + 2 \leq l \leq R(j + 1)$. In the other case, we note that there exist either one or two HCB sections of length between 7 and 13 or there is a section of length $\geq 14$. A case by case analysis shows that for the SFBs, where this(these) section(s) start(s), there is $G_i(H, g) = G_i(h_i - h_{i+1}) + r$, where $r = 3 = 3(x_{2j}, x_{4j},)$ if a (first) HCB section of length between 7 and 13 starts at $SFB_i$, or $r = 6 = 3(x_{2j}, x_{4j},) + 3(x_{4j},)$ if a HCB section of length $\geq 14$ starts at $SFB_i$. Combining these definitions and comparing them with (2) we see

$$R(S, H, g, B) = \sum_{i=1, i \neq 1 \mod R}^{N} G_i(H, g) \quad (5)$$

$$= \sum_{i \in N_g}^{8R} G_i(h_i - h_{i+1}) + \sum_{j=0}^{7} 3(x_{2j} + x_{3j} + x_{4j}) \quad (6)$$

We introduce the variable $A_n$ that is set equal to 1 if $SFB_i$ is a valid band and $s_{i-1} = a - M_1$, and is set equal to zero otherwise. Similarly, the variables $B_{i,j}$ describe the difference...
of the values $h_i$ and $h_{i-1}$. Thus, we set

$$A_{i,a} = \begin{cases} 
1, & \text{if } SFR_i \text{ is valid,} \\
0, & \text{else,}
\end{cases}$$

$$B_{i,b} = \begin{cases} 
1, & \text{if } SFR_i \text{ is valid,} \\
0, & \text{else,}
\end{cases}$$

$$2 \leq i \leq 8R, \forall a, 1 \leq a \leq M_1, 1 \leq b \leq 2M_2 - 1,$$

For the same range of parameters, we define the variables $Z_{g,i,a,b}$ that describe which values are taken by the variables $s_i$ and $h_i$ at the $i$-th stage of the $g$-th configuration:

$$Z_{g,i,a,b} = \begin{cases} 
1, & \text{if } s_i = a, h_i = b, \text{ config. } g \text{ is chosen,} \\
0, & \text{else,}
\end{cases}$$

Next, we define the variable $e_{i,m}$ as the weighted distortion that occurs at the $i$-th band when the $g$-th configuration is chosen and if $s_i = m$, i.e., $e_{i,m} = w_{j,d}d(s_i = m), \forall 1 \leq i \leq 8R, 1 \leq m \leq M_1$. Further, in order to avoid negative function arguments we set $F^*(a_2 - a_1) = F(M_1 + a_2 - a_1)$ and $G^*(b_2 - b_1) = G(M_2 + b_2 - b_1)$.

**B. MILP formulation for the ANMR problem**

We formulate the MPEG4-AAC encoding problem as follows:

$$\text{Minimize}$$

$$ANMR(Z) := \frac{1}{N} \sum_{g=0}^{G_1} \sum_{a=1}^{8R} \sum_{i=1}^{M_1} \sum_{b=1}^{M_2} e_{g,i,a}Z_{g,i,a,b},$$

such that (6) - (30) are satisfied, where

$$R(Z, A, B, g, U, X) \leq R_{1},$$

$$R(Z, A, B, g, U, X) = \sum_{g=0}^{G_1} \sum_{a=1}^{8R} \sum_{i=1}^{M_1} \sum_{b=1}^{M_2} Z_{g,i,a,b}Q_{g,i}(a, b)$$

$$+ \sum_{j=0}^{8R} \sum_{i \in \{1 (\text{mod } R)\}}^{M_1} A_{i,a}F^*(a) + \sum_{j=0}^{8R} \sum_{i \in \{1 (\text{mod } R)\}}^{M_1} B_{i,b}G^*(b)$$

$$+ 3 \sum_{j=0}^{8R} \sum_{i \in \{1 (\text{mod } R)\}}^{M_1} (x_{2,j} + x_{3,j} + x_{4,j}) + c \sum_{j=1}^{7} u_{j},$$

$$\sum_{g=0}^{G_1} v_g = 1,$$

$$\sum_{g=0}^{G_1} u_{j}2^{j-1} = \sum_{g=0}^{G_1} g v_g,$$

$$\sum_{a=1}^{M_1} \sum_{b=1}^{M_2} Z_{g,i,a,b} = v_g, \quad 0 \leq g \leq G_1, i \in N_g,$$

$$\sum_{a=1}^{M_1} \sum_{b=1}^{M_2} Z_{g,i,a,b} = 0, \quad 0 \leq g \leq G_1, i \not\in N_g,$$

$$\sum_{a=1}^{M_1} A_{i,a} \leq 1, \quad 1 \leq i \leq 8R, i \not\equiv 1 (\text{mod } R),$$

$$\sum_{b=1}^{M_2} B_{i,b} \leq 1, \quad 1 \leq i \leq 8R, i \not\equiv 1 (\text{mod } R),$$

for $0 \leq j \leq 7$. The variables $Z_{g,i,a,b}$, $x_{m,j}, v_g, u_j, e_{i,j}, f_j, d_{i,j}$, $g_k$, $w_k$ are required to be integers in $\{0, 1\}$. The variables $A_{i,a}$ and $B_{i,b}$ are required to be non-negative real numbers which however will always be chosen $c \in [0, 1]$ by the MILP as we will explain in the following.

We know explain how the MILP formulation correctly models the ANMR encoding problem as defined in (3) and (4). The equation (8) shows that exactly one configuration $g$ is chosen by the encoder. The equation (9) establishes the one-to-one relation between $g$ and a possible choice of the $u_j$ discussed in Section III-A. The equation (10) ensures that for a chosen configuration $g$ and each valid band $SFR$, exactly one variable $Z_{g,i,a,b}$ is equal to 1, whereas all the other variables are equal to zero, i.e., exactly one SF and one HCB are chosen per band. For an invalid band $SFB$, all $Z_{g,i,a,b} = 0$ by equation (11). For fixed values of the variables $Z_{g,i,a,b}$, the equations (12), (14), and
(16) ensure that for any $u_j = 1$, and $R_j + 2 \leq i \leq R(j + 1)$, if there is $s_i - s_{i-1} = a - M_1$, then $A_{i,a} = 1$ and all other $A_{i,a} = 0$. The equations (13), (15), and (17) imply the same relations for the variables $B_{i,s}$. For all invalid bands, the variables $A_{i,a}$, and $B_{i,s}$ are equal to zero.

Finally, we analyze the equations (18) - (30). The equations (18) and (19) force the variables $x_{2,j}$ to be equal to 1 if at least one HCB section of length $\geq 7$ exists between the $SFB_{R_{j+1}}$ and $SFB_{R_{j+1}+1}$. If such an HCB section does not exist, then by (18) and (19) $x_{2,j}$ can take any value $\in \{0, 1\}$ which contradicts the definition of $x_{2,j} = 0$ in Section III-A. By its definition in Section III-A, the variable $x_{2,j}$ does not represent any choice of the encoding variables itself, but merely expresses the impact of specific choices of the encoding variables on the rate. If $x_{2,j}$ is chosen equal to 1, then $x_{2,j}$ makes a higher contribution to the rate $R(Z, A, B, X, Q, U)$ than defined in Section III-A and thus eventually reduce the solution space of the variables $Z_{i,a,b}$, $A_{i,a}$, and $B_{i,s}$. If such a reduction would indeed take place for $x_{2,j} = 1$, the linear program will always set $x_{2,j} = 0$ to avoid the reduction. Thus, we see that with regard to $x_{2,j}$, the MILP formulation allows to find the optimal solution of the MPEG4-AAC encoding problem under the exact constraints formulated in Section III-A. For the discussion of the variables $x_{m,j}$, $3 \leq m \leq 4$, the same arguments apply. Thus, we will only show that if their definition in Section III-A requires them to be equal to 1, then the MILP will force them also to be equal to one, whereas in the case that their definition in Section III-A requires them to be equal to 0, the MILP allows them to take any value $\in \{0, 1\}$.

The relations (20) - (28) imply that $x_{1,j} = 1$ if two HCB sections of length $\geq 7$ exist between $SFB_{R_{j+1}}$ and $SFB_{R_{j+1}+1}$. This can be seen by a case-by-case analysis of all the possible choices of HCBs that contain two HCB sections of length $\geq 7$. The equations (20) - (25) model the case when two HCB sections of length 7 exist, whereas the equations (26) - (28) model the case when one HCB of length 7 and one HCB section of length 8 exist. Finally, the equations (29) and (30) imply that $x_{4,j} = 1$ if one HCB section of length $\geq 14$ exists.

The objective function is defined identically in (1) and (6). The definition of transmission rate in (2) and (5) is identical to the definition in (7). We note that there is $q = 1 - \sum_{2 \leq j \leq u_j} u_j$, where $q$ is the number of regions. Thus the value of $B$ as defined in Section II and (2) is correctly expressed in (7).

### IV. Experimental results

In this section, we compare the performance of the MILP algorithm proposed in this paper with a Reference AAC encoder. The Reference AAC encoder implements a spectral energy difference technique [7] to decide the grouping configuration. and applies the Two-Loop Search routine [4] for computing the SFs and HCBs. Further, we designed an Exhaustive Search AAC encoder as follows: The Exhaustive Search AAC encoder executes the Two-Loop Search for all possible 128 groupings and chooses the grouping with the lowest distortion.

The MILP encoder can be implemented in two mathematically equivalent ways. First, one can directly implement and solve the linear programming formulation in Section III. One uses the simplex method to solve the underlying linear program and then deploys a branch-and-bound method to identify the integer solution. CPLEX provides an implementation of both algorithms. Alternatively, for each of the 128 possible grouping configurations one determines the optimal settings of the SFs and HCBs by using the optimal methods developed in [6], and then chooses the grouping with the lowest ANMR. The first implementation is computationally less complex. However, as for the sake of this evaluation we were not concerned about the execution time of the algorithm and we already had an implementation of [6], we implemented MILP in the second way. We see that this way of implementing MILP differs from the Exhaustive Search AAC in the fact that the optimal methods in [6] are used instead of the heuristic Two Loop Search.

The motivation for comparing these three algorithms was as follows: The comparison of Reference AAC and Exhaustive Search AAC gives us an insight into the quality of the grouping algorithms in [7]. The comparison of both AAC modes with MILP shows us how far the solutions found by both AAC algorithms differ from the theoretical optimum calculated by MILP. We note that both the Exhaustive Search AAC and MILP can currently not be executed in real-time.

The results in Table I show that the Exhaustive search AAC only gives a 0.531 dB improvement over the Reference AAC which shows the strength of the grouping algorithms in [7]. The comparison of both AAC modes with MILP shows that the achieved ANMR values are far from the theoretical optimum.

For the simulations, data was taken from a monophonic audio file, sampled at 44.1 kHz and 100b in length, with 755 short block frames analyzed. The audio file consisted of seven audio sequences containing dynamic program content, such as triangle, glockenspiel, harpsichord, and castanets.

### V. Conclusions

This paper develops the first complete model of the encoding process for the MPEG-4 AAC audio codec that chooses a short block configuration. Previous techniques to solve the AAC encoding problem divide the encoding problems into sub-problems which are then solved heuristically. In contrast, we model the encoding process as a four dimensional optimization problem. The numerical experiments show the large performance gains that can be achieved by our approach compared to previous methods. The calculation of the optimally achievable ANMR values allows us to compare the performance of existing technologies with the theoretical optimum.

### References


