The Choice of MPEG-4 AAC encoding parameters as a direct function of the perceptual entropy of the audio signal

Claus Bauer, Mark Vinton

Abstract— **This paper proposes a new procedure of low-complexity to determine the encoding parameters for the MPEG-4 AAC encoder under real-time constraints. In particular, it addresses the optimization problem of minimizing the distortion subject to a rate constraint for an MPEG-4 AAC encoder. Existing implementations use the heuristic Two Loop search to solve this optimization problem. This paper presents a new solution algorithm which achieves distortion values significantly lower than the Two Loop Search and which, due to its low computational complexity, is a promising technology for future AAC implementations. We show via simulations that the technology presented in this paper significantly outperforms previous technologies.**

I. Introduction

In recent years, the delivery of multimedia content over wireless networks has rapidly gained importance. Multimedia applications are expected to be the driving applications for high bandwidth Third Generation Cellular, WiMAX, and WiFi networks. The success of these technologies also depends on the availability of low bit rate audio codecs such as MPEG4-AAC.

AAC achieves perceptual qualities at low bitrates by exploiting perceptual redundancies in the signal. The AAC encoder partitions each audio frame into a number of bands, and for each band, dynamically allocates bits to encode the transform coefficients. In addition to the encoded transform coefficients, the quantizer step size and the Huffman Code Book of each band are transmitted to the decoder as side information. The total transmission rate from the encoder to the decoder is the sum of the bits needed to encode the transform coefficients and the bits needed to encode the side information. The side information itself is encoded differentially and thus depends on the relation of the parameters at adjacent bands. To ensure a low bit rate encoding, the encoder must choose the quantizer step size and the Huffman Code Books such that the total transmission rate is below a predefined threshold, while ensuring that a predefined objective measure for the perceptual quality of the decoded signal is satisfied. The inter-band relationship of the transform parameters shows the complexity of the optimization problem.

The most common perceptual quality measures is the average weighted ($ANMR$) noise to mask ratio, which are also called the average distortion. We define the $ANMR$ in section II. In this paper, we investigate the problem of determining the set of the *encoding parameters* that optimize the $ANMR/MNMR$ subject to an upper bound on the permissible transmission rate. Obviously, low transmission rates that do not compromise the quality of the received audio are of strong interest to the delivery of audio content over bandwidth constrained networks. The most common procedure to solve the $ANMR$ problem is the Two Loop Search (TLS) [3]. The TLS uses a heuristic approach, which neglects the inter-band dependencies of the side information and thus simplifies the problem significantly by optimizing the two encoding parameters - quantizer step size and Huffman Code Books - *independently* for each band. This leads to an increased total transmission rate and/or an increase of the $ANMR$.

In [1], a *joint* optimization of the encoding parameters of all bands has been proposed. The possible choices of the encoding parameter are modeled as a trellis and the $ANMR$ problem is solved using an iterative Viterbi search through the trellis. The cost function includes a Lagrangian multiplier that penalizes any violation of the target rate. The Trellis Search has been further refined in [13].

In [2], algorithms that find the optimal parameter settings for both the SFs and $HCBs$ are presented. The authors show that both the Trellis Search [1], which achieves distortions on average 10% above the optimum value, and the optimal solution algorithms in [2] are computationally too complex for real time applications.

In this paper, we propose the *Fast Trellis Search* algorithm, which qualitatively performs as well as the Trellis Search [1], but is of significantly lower complexity than the Trellis Search. The complexity of the Trellis Search is caused by the iteration over the Lagrangian multiplier and the solution of an optimization problem for each iteration. Using results from mathematical analysis and digital signal processing theory, we show that it is possible to derive a good estimate for the final Lagrangian multiplier - without having to iterate over all initial Lagrangian multipliers - from the properties of the signal. In particular, we express the final Lagrangian multiplier as a function of the *Perceptual Entropy* of the signal and the target rate of the encoding process. This relationship is the base of the Fast Trellis Search as it avoids the iteration over various Lagrangian multipliers, which reduces the computational complexity. We will show that the Fast Trellis Search is a candidate for future real-time implementations of AAC.

In the next section, we develop an analytic formulation of the problem under consideration. In section III, we develop

The authors are with Dolby Laboratories, San Francisco, CA, 94103, Fax: 415 8631373, C. Bauer: Tel: 415 5580343, cb@dolby.com, Mark Vinton: Tel : 415 5580785, msv@dolby.com.

1-4244-0000-7/05/$20.00 ©2005 IEEE.

the Fast Tellis Search algorithm. We numerically evaluate the Fast Trellis Search in section III-C and conclude in section IV.

II. Problem definition

The AAC encoder converts the time domain signal into the spectral domain using the modified discrete cosine transform (MDCT). The 1024 spectral coefficients obtained via the MDCT are grouped into N scale factor bands ($SFBs$). Within each band, all coefficients are quantized using the same scalar quantizer. As not all 1024 coefficients are relevant for the perception of the decoded signal by the human ear, not all coefficients are quantized. The quantizer step size is controlled by a scale factor (SF) selected from a range of typically 60 SFs. Within a SFB, the quantized coefficients are entropy encoded using a Huffman Code Book (HCB) selected from typically 12 $HCBs$. The SF and HCB parameters are transmitted as side information.

In order to formalize the problem of minimizing the distortion subject to a rate constraint, we introduce the following notation. We assume that a frame consists of N scale factor bands SFB_i, $1 \leq i \leq N$, and that the encoder can choose from a set of M_1 scale factors and M_2 Huffman Code Books. Further, let s_i be the SF value and h_i be the HCB value for the i^{th} scale factor band in the frame. We define vectors $S = \{s_1, .., s_N\}$ and $H = \{h_1, .., h_N\}$. We assume that the s_i and the h_i only take integer values and require $1 \leq s_i \leq M_1$, and $1 \leq h_i \leq M_2$, $\forall i$, $1 \leq i \leq N$. Both the s_i and h_i are indexes into sets of pre-determined Scale Factors and Huffman Code Books, respectively.
The average noise to mask ratio ($ANMR$) [3] is defined as the ratio of the quantization noise to the masking threshold [14]. To express the $ANMR$ analytically, we define $d(s_i)$ as the quantization error of the i-th scale factor band if the i-th scale factor is chosen equal to s_i. w_i denotes the weight of the i-th scale factor band which is defined as the inverse of the masking threshold (see [14]) of the i-th band. The ANMR is expressed as

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

In the following, we derive an analytic expression for the transmission rate. The transmission rate consists of three parts:

- Let $Q_i(s_i, h_i)$ be the bits required to encode the quantized coefficient indices of the ith band with the SF value chosen as s_i and the HCB value chosen as h_i. We note that the function $Q_i(s_i, h_i)$ is also a function of the actual signal X., i.e. $Q_i(s_i, h_i) := Q_{X,i}(s_i, h_i)$. In general, for any two different signals X and Y, $Q_{X,i}(a, b) \neq Q_{Y,i}(a, b)$. As we only consider a fixed signal, we omit the index X.
- The function $F(s_{i-1}, s_i)$ gives the number of bits required to specify the SF for SFB_i. As the SFs are encoded differentially, we note that $F(s_{i-1}, s_i) := F(s_{i-1} - s_i)$.
- Finally, $G(h_{i-1}, h_i) = G(h_i - h_{i-1})$ represents the number of bits needed to encode the HCB value of SFB_i.

The transmission rate $R(S, H)$ is defined as

$$R(S,H) = Q_1(s_1,h_1) + \sum_{i=2}^{N}\Big(Q_i(s_i,h_i) + F(s_{i-1}-s_i) + G(h_{i-1}-h_i)\Big). \tag{2}$$

For a given rate threshold R_t, the $ANMR$ problem is then defined as follows:

$$Minimize \qquad ANMR(S) \tag{3}$$
$$such\,that \qquad R(S,H) \leq R_t. \tag{4}$$

III. Fast Trellis Search

A. Fast Trellis Search: A direct estimate of the final Lambda multiplier in the Trellis Search

We first revisit the Trellis Search in [1]. The author builds a trellis consisting of N stages where each stage corresponds to a SFB. The states of each stage i are the set of all possible choices of the parameters h_i and s_i. Each path through the trellis corresponds to a specific choice of the quantization parameters s_i and h_i. For the solution of the joint optimization problem as defined in (3), the author uses an "unconstrained" cost function that is the sum of the distortion as expressed in (1) and the product of a Lagrangian multiplier Λ and the rate $R(S, H)$. Using the definitions introduced in section II, the cost of choosing a path in the Trellis that includes the subpath from the $i-1$-th band at stage (s_{i-1}, w_{i-1}) to the i-th scale factor band at stage (s_i, w_i) is defined as

$$C(s_{i-1},h_{i-1},s_i,h_i) = w_i d(s_i) + \Lambda\left(Q_i(s_i,h_i) + F(s_{i-1}-s_i) + G(h_{i-1}-h_i)\right).$$

The Trellis Search iterates over the Lagrangian multiplier Λ and uses a Viterbi search to find the cheapest path relative to the norm $C(s_{i-1}, h_{i-1}, s_i, h_i)$ through the trellis for each Λ. The iteration stops when $ANMR(S)$ does not decrease any further and the rate constraint (4) is satisfied. The iteration over Λ leads to the high computational complexity of this approach which makes it not feasible for practical applications. In the next section, we propose the Fast Trellis Search which avoids this computational complexity. It derives a close guess for the value of the final Lagrangian multiplier Λ_{final} in the iteration without having to iterate over all previous values of Λ.

B. Facts from digital signal processing theory

In (1), we have introduced the average noise to mask ratio $ANMR$ as a unit-less number. Alternatively, using a logarithmic measure, we use the definitions in [15] to express the quantities average noise to mask ratio $ANMR_R^{dB}$, the Average Signal to Noise Ratio $ASNR_R^{dB}$ and the Average Signal to Mask Ratio $ASMR^{dB}$ in dB. The Average Noise to Mask Ratio $ANMR_R^{dB}$ can then be expressed as

$$ASNR_R^{dB} - ASMR^{dB} \quad = \quad -ANMR_R^{dB}. \tag{5}$$

We have indexed the $ASNR_R^{dB}$ and the $ANMR_R^{dB}$ with the rate index R, because both quantities depend on the actual rate R. Using the definition (1), we see

$$ANMR_R^{dB} \quad = \quad 10\log_{10}\left(\frac{1}{N}\sum_{i=1}^{N} w_i d(s_i)\right). \qquad (6)$$

We will make use of the notion of Perceptual Entropy PE that was introduced by Johnston [7] and is summarized in [9]. Perceptual Entropy is defined as a measure of perceptually relevant information contained in any audio signal. Expressed in bits per sample, PE represents a theoretical limit on the compressibility of a particular signal. An explicit calculation of PE has been given in [9]:

$$\begin{aligned} PE \quad &= \quad \frac{1}{M}\sum_{i=1}^{N}\sum_{b_i=bl_i}^{bh_i}\left[\log_2\left(2\left[\frac{Re(t_{b_i})}{\sqrt{6/w_i k_i}}\right]+1\right)\right. \\ &+ \quad \left.\log_2\left(2\left[\frac{Im(t_{b_i})}{\sqrt{6/w_i k_i}}\right]+1\right)\right], \qquad (7)\end{aligned}$$

where

i index of critical band,
bl_i upper bound of band i,
bh_i lower bound of band i,
k_i number of transform components in band i,
b_i, index of the transform coefficients in the i^{th} critical band,
t_{b_i}, transform coefficient in the i^{th} critical band,

and w_i is the inverse masking threshold as defined in section II. We note that in the literature the PE is often defined with an additional factor $\frac{1}{M}$ on the right side of (7).

It is known from [10] that a noise, in particular the quantization noise, that falls below the masking threshold is inaudible. Thus, when a signal is (theoretically) encoded with PE bits, at each band the masking threshold is chosen as equal to the inverse of the distortion, i.e.,

$$w_i = \frac{1}{d(s_i)}. \qquad (8)$$

We see from (5), (6), and (8) that

$$ASNR_{PE}^{dB} \quad = \quad ASMR^{dB}. \qquad (9)$$

Assuming a uniform quantization step, we know from [15] that for any rate R, there holds statistically

$$ASNR_R^{dB} \quad = \quad \frac{cR}{M} - k, \qquad (10)$$

where $c = 6.02$, M is the number of quantized coefficients (see section II), and k is a constant that depends on the PDF of the signal. As AAC uses non-uniform quantization steps, it is of interest to understand if a relation similar to (10) holds for non-uniform quantization steps. We answered this question affirmatively by doing the following experiment: For a set of MPEG audio test items (see section **??** for a more detailed description of the test items), we applied an AAC quantizer to compute a $\frac{R}{M}$ - bit quantization of the $MDCT$ coefficients. Then, we computed the $ASNR_R^{dB}$ as defined in [14]. Figure 1 shows the test result for a typical MPEG audio test item: Statistically the $ASNR_R^{dB}$ is a linear function of the ratio $\frac{R}{M}$ with a slope close to 6.02. This result was achieved consistently over the set of considered MPEG audio test items. Thus, we assume in the sequel that also for non-uniform quantization the $ASNR_R^{dB}$ can - statistically - be expressed as a linear function of $\frac{R}{M}$ as in (10).

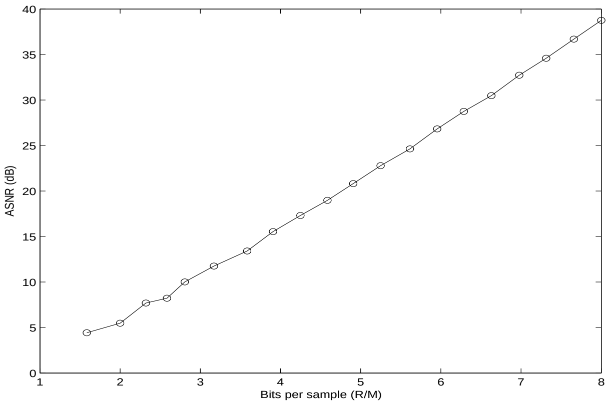

Fig. 1. The $ASNR_R^{dB}$ as a linear function of bits per sample $\frac{R}{M}$

We see from (5), (6), (9), and (10),

$$\begin{aligned}\frac{c(R-PE)}{M} \quad &= \quad ASNR_R^{dB} - ASMR^{dB} \\ &= \quad -10\log_{10}\left(\frac{1}{N}\sum_{i=1}^{N} w_i d(s_i)\right). \qquad (11)\end{aligned}$$

Using (1) and (11), we express for fixed N, M, and PE the average distortion $ANMR(S)$ as a function of R, i.e., we write $ANMR(S) = D_R$, where

$$D_R \quad = \quad 10^{\frac{c(PE-R)}{10M}}. \qquad (12)$$

C. An interpretation of the final Lagrangian multiplier Λ_{final}.

We know from the theory of the Lagrangian multiplier ([4],[5],[11],[12]) that the negative slope of the distortion rate function is equal to the final Lagrangian multiplier Λ_{final}^R, where R denotes the rate, i.e.,

$$-\frac{dD_R}{dR} \quad = \quad \Lambda_{final}^R. \qquad (13)$$

Using (12), we write D_R as follows:

$$\begin{aligned} D_R \quad &= \quad 10^{\frac{cPE}{10M}} 10^{\frac{-cR}{10M}} \\ &=: \quad 10^{\frac{cPE}{10M}} F(R). \qquad (14)\end{aligned}$$

As $F(R) = e^{-R\frac{c\,ln\,10}{10M}}$, there is

$$\frac{dF(R)}{dR} = -F(R)\frac{c\,ln\,10}{10M}. \tag{15}$$

From (13), (14), and (15) we obtain the following estimate for Λ^R_{final} :

$$\Lambda^R_{final} = \frac{c\,ln\,10}{10M} \times 10^{\frac{c(PE-R)}{10M}}. \tag{16}$$

For a fixed value of R, and assuming that c is a positive constant, we see from (16) that Λ^R_{final} increases with increasing PE. This can be intuitively explained as follows: For increasing PE, the minimum number of bits needed by the encoder to encode an audio sample increases, such that the transmission rate tends to increase and potentially violates the rate constraint. However, for an unconstrained cost function defined as in (5), if the transmission rate is high, the value of Λ must be large in order to strongly penalize a violation of the rate constraint.sectionApplications and Simulation results

D. Applications of the Fast Trellis Search

For practical applications, we note that the MPEG specification foresees a *bit reservoir* for the encoding process, i.e., a reservoir of bits is provided to the encoder to encode the content of a given audio file. Consequently, it is not strictly necessary that the encoder meets the target rate for all frames of an audio file, but it must ensure that it meets the target rate on average over the whole file. Indeed, the experiments presented in section III-E show that although the rate constraint is violated by the Fast Trellis Search on individual frames, it is met on average over all samples of a typical audio file. This is due to the fact that the Fast Trellis Search often finds solutions that require less transmission bits than the target rate.

As another possible application, one could use the Λ_{final} determined by formula (16) as an initial value for an iterative Trellis Search that - due to the close guess of Λ_{final} - would require significantly less iterations over Λ than a Trellis Search that randomly picks an initial Λ value.

E. Simulation results

In this section, we apply the formula (16) to determine the value of Λ_{final} and compare it with the actual experimental value for Λ. For the experiments, we used the AAC encoder in *bit reservoir* mode and used common MPEG audio test samples. We compared the different methods using $\sim$ 30000 audio frames, 43 $SFBs$, 60 SFs, and 12 $HCBs$. Our sample rate was 44100 Hz, we used a single channel (mono), and we chose a rate of $\sim$ 32 kbps.

As the formula (10) - and so (16) - is only asymptotically correct, we try to improve the accuracy of (16) by introducing an additional degree of variability into the relation (16), and write

$$\Lambda^R_{final} = \frac{c_1\,ln\,10}{10M} \times 10^{\frac{c_2PE-c_3R}{10M}}. \tag{17}$$

In order to determine the values of the constants c_1, c_2 and c_3, we execute the Trellis Search over a wide range of values for N, M, PE, and R and determine Λ_{final} by iteration over all initial Λ values. Taking the logarithm on both sides of the formula (17), we use the experimental data to determine the constants c_1, c_2 and c_3 as a two-dimensional least square problem. We obtain the values $c_1 = -9248.3$, $c_2 = 11.712$, and $c_3 = 8.897$. We see that the constants c_1 - c_3 differ significantly from the value $c = 6.02$ in (10). This can be expected as the relation (10) is only a very rough approximation for non-uniform quantization. In order to evaluate the formula (17) with the constant values chosen as above, we compared the formula (17) with experimental values for Λ_{final} obtained by iterating over all initial Λ values. Figure 2 shows that the actual values for Λ_{final} are indeed very well approximated by the formula (17). Fig. 3 shows that the distortions that are respectively achieved susing the Fast Trellis Search, by estimating Λ_{final} using formula (17), and the Trellis Search, by iterating over all initial values of are nearly identical. Figure 4 shows that the rate which is achieved by the Trellis Search using the analytic Λ_{final} value violates the rate constraint in $\sim$ 30% of all samples. However, the simulations also show that the achieved transmission rate is always less than 10% above the target rate. Also, the simulation results show that the rate constraint is met on average which allows a deploy-

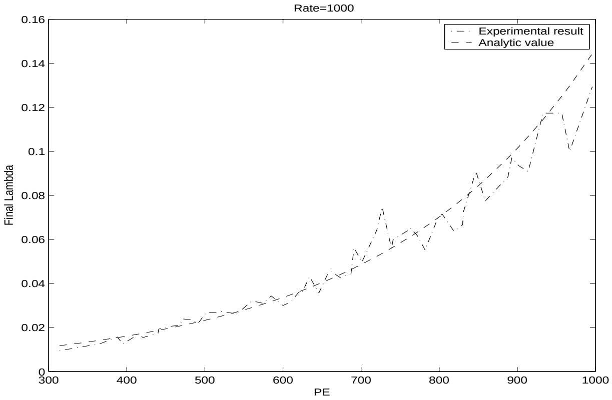

Fig. 2. Comparison of the experimental value and analytic estimate of Λ_{final} for a target rate R=1000, M= 768, N=1024

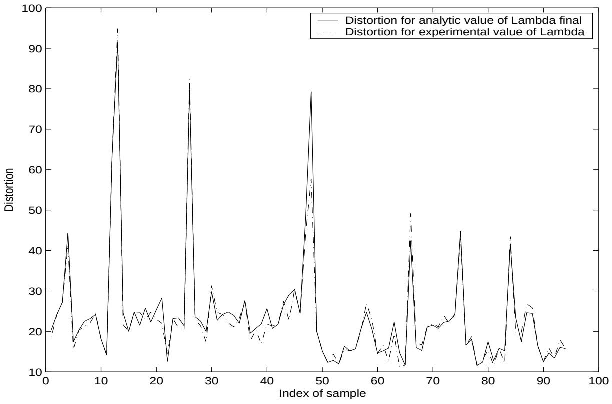

Fig. 3. Comparison of the average distortion for experimental and analytic values of Λ_{final} for a target rate R=800

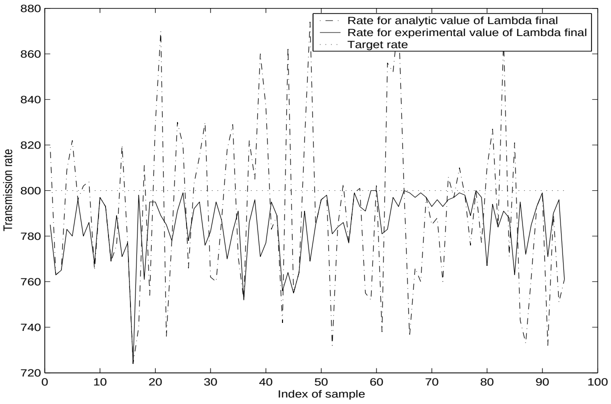

Fig. 4. Comparison of the total transmission rate for experimental and analytic values of Λ_{final} for a target rate R=800

TABLE I

Distortion relative to masking curve

Optimization technique	Probab. that ANMR is below masking curve
Two Loop Search	57%
(Fast) Trellis Search	86%
Optimal solution	93%

ment of the Fast Trellis Search in the bit reservoir mode.

We recall from [2], that the Trellis and thus - based on the results from this section - also the Fast Trellis Search is on average 10% above the optimum value. Whilst the $ANMR$ is an effective measure of ultimate performance of the encoding process, the objective of perceptual compression is to achieve an $ANMR$ that is *satisfactory,* regardless of whether it is the best $ANMR$ possible or not. In the context of audio coding, an $ANMR$ is widely considered as $satisfactory$ if it is below or at least not far above the masking curve. The $ANMR$ is below the masking curve if $ANMR < 1$. Our experiments - see I - show that whereas the TLS achieves an $ANMR$ below the masking curve with a probability of 57%, both the Trellis as well as the Fast Trellis Search obtain $ANMR$ values below the masking curve with a probability of 86%, whereas the optimal solutions, which we calculated using the algorithms in [2], achieve such $ANMR$ values with a probability of 93%.

The experiments show that the Trellis Search needs about 2 seconds, whereas the Fast Trellis Search needs about 80 ms to solve the $ANMR$ problem. Currently, the TLS is the only real-time method as it needs only several milliseconds to solve the ANMR problem. However, in view of the rapidly increasing capacity of commercial processors, the Fast Trellis Search could soon be suitable for real-time implementations. Optimal methods [2] are far too complex for real-time implementations.

IV. Conclusions

This paper proposes a new method to solve the problem of minimizing the average distortion subject to a constraint on the total transmission rate for the MPEG4-AAC encoder. The Fast Trellis Search outperforms the Two Loop Search as it finds $ANMR$ values which on average are only 10% above the optimal values and are below the masking curve with a probability of 86%. The simulations show that when using the MPEG bit reservoir the Fast Trellis Search and the Trellis Search achieve nearly identical distortions for each frame and the Fast Trellis Search meets the rate constraint on average over the samples of an audio file. Due to its low complexity, the Fast Trellis Search is a promising technology for future AAC implementations.

References

[1] Aggarwal. A., *Towards weighted mean-squared optimality of scalable audio coding.* Dissertation, UCSB, Dec. 2002.

[2] Bauer, C., Vinton, M., *Joint Optimization of Scale Factors and Huffman Code Books for MPEG-4 AAC,* IEEE International Workshop on Multimedia Signal Processing 2004, Siena, Italy.

[3] Bosi, M., et al., *ISO/IEC MPEG-2 Advanced Audio Coding.* Journal of the AES, Vol. 45, No. 10, 1997, Oct, pp.791 -811.

[4] Chou, P.A., Lookabaugh, T., Gray, R.M., *Entropy-Constrained Vector Quantization.* IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, no. 1, Jan. 1989, pp. 31 - 42.

[5] Everett, H., *Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources.* Operations Research, Vol. 11, 1963, pp. 399 - 417.

[6] Geoffrion, A.M., *Duality in Nonlinear programming: A simplified applications-oriented development.* SIAM Review, Vol. 13, No.1 , Jan 1971.

[7] Johnston, J.D., *Transform Coding of Audio Signals Using perceptual Noise critera.* IEEE JSAC, Vol. 6, No. 2, Feb. 1998.

[8] Najafzadeh, H., Kabal, P., *Perceptual bit allocation for low rate coding of narrowband audio.* ICASSP 2000, Istanbul, June 2000.

[9] Painter, T., Spanias, A., *Perceptual Coding of Digital Audio.* Proc. of the IEEE, Vol. 88, No. 4, April 2000, pp. 451 - 513.

[10] Schroeder, M.R., Atal, B.S., Hal, J.L., *Optimized Digital Speech Coders by exploiting masking propoerties of the human ear.* J. acoustic. Soc. Am. 66(6), Dec. 1979, pp. 1647 - 1652.

[11] Shoham, Y., Gersho, A., *Efficient Bit Allocation for an Arbitrary set of Quantizers.* IEEE Transactions on Acoustics, Speech and signal processing, Vol. 36, Sept. 1988, pp. 1445 - 1453.

[12] Wiegand, T., Girod, B.; *Lagrange Multiplier Selection in Hybrid Video Coder Control.* Proc. of ICIP'01, Thessaloniki, Greece.

[13] Yang, C.H., Hang, H.M.,*Cascaded Trellis-Based Optimization for MPEG-4 Advanced Audio Coding.* 115th AES conv., 2003, NYC.

[14] Zwicker, E., Fasthl, H., *Psychoacoustics: Facts and Models.* Second Edition, Springer Publishing House, 1990c1999.

[15] Zoelzer, U., *Digital Audio Signal Processing.* John Wiley Sons Ltd. 1997.