OPTIMIZATION-QUANTIZATION FOR LEAST SQUARES ESTIMATES AND ITS APPLICATION FOR LOSSLESS AUDIO COMPRESSION

Florin Ghido and Ioan Tăbuş

Institute of Signal Processing, Tampere University of Technology, Finland

ABSTRACT

In this paper we study the problem of optimally quantizing the least square estimates and we introduce a method where the quantized estimate vector is obtained by a sequence of interleaved optimization-quantization scalar stages. We show how the general approach can be reduced to a simple and efficient algorithm when connecting it to the LDL^T solver for general LS problems. The application to quantization of the linear prediction coefficients for audio lossless coding reveals the high performance of the approach, leading to topmost performance in the class of frame based audio coders, surpassing significantly the performance of the current MPEG4-ALS standard.

Index Terms— optimal quantization, least squares, linear prediction, lossless audio compression, MPEG4-ALS standard.

1. INTRODUCTION

The quantization of LS estimates and tracking the loss of performance due to quantization is a generic problem which can be encountered in all areas of engineering and science. We consider the application of least squares linear prediction for asymmetrical lossless audio compression, where the prediction coefficients are transmitted as side information, making decoding very fast. The study of quantization of linear prediction coefficients (LPC) has a long history and we can distinguish two distinct areas of applications: the first is lossy compression (including the important application to speech coding) and second is frame-wise (or forward) lossless compression. Here we use the later technique and we note that there are many equivalent representations of LP coefficients, such as reflection coefficients, log-area ratios of reflection coefficients, and arcsine of reflection coefficients. Scalar quantization can be applied to any of these representations, obtaining different results in the final compression application. We note that the conclusions of most previous studies were favoring always the alternative representations and consequently the quantization of the direct representation of LPC was traditionally considered only a bad choice. However, we are going to show that working with the quantized direct form of LPC we get remarkably good tradeoff predictor complexity-prediction accuracy, and having this technique implemented in an audio codec provides the best performance available in terms of compression ratios and encoding/decoding times, surpassing all existing methods and standards in the field.

1.1. Setting the quantized LS problem

We are given an integer audio signal, mono, stereo, or multichannel, and we need to make prediction frame-wise, by splitting the signal into nonequal frames of sizes in the range of hundreds to thousands samples. In the stereo case, for each frame from a given channel, we select a number of n_i regressor samples from the current channel and a number of n_r regressor samples from a reference channel to create a stereo predictor.

Within a frame of length N, the predictor of $x(t)$ will operate with regressor vectors $u(t)$ of length equal to the maximal predictor order M (it does not matter if it is a mono, stereo, or multichannel predictor). We compute the covariance matrix $R = \sum_{t=1}^{N} u(t)u^T(t)$, the cross-correlation vector $r = \sum_{t=1}^{N} u(t)x(t)$, and the variance $\sigma^2 = \sum_{t=1}^{N} x^2(t)$ as usual in the LS method [1], making use in the vectors $u(t)$ of the values from previous frames when needed. We denote w the linear prediction coefficients for computing the prediction $\hat{x}(t) = w^T u(t)$ and we use as the optimality criterion the sum of squared prediction errors, which can be written in the form $J(w) = \sigma^2 - 2r^T w + w^T R w$. The optimal linear prediction coefficients $w_o = R^{-1} r$ can be found by minimizing $J(w)$ and the corresponding optimal criterion is $J(w_o) = \sigma^2 - w_o^T r$, with which we can rewrite the value of the criterion for an arbitrary w as

$$J(w) = J(w_o) + (w - w_o)^T R (w - w_o). \quad (1)$$

1.2. Scalar quantization of prediction coefficients

For a positive integer Q, the uniform scalar quantization of the real number x will be denoted $x_Q = \text{round}(xQ)/Q$ and we will use throughout the paper the convention that a variable with subscript Q denotes the quantized variable. We can extend this operation to vectors where quantization is applied element-wise, so that the quantized LS solution is $w_{o_Q} = \text{round}(w_o Q)/Q$. Any quantization error $\epsilon = x - x_Q$ can be seen to be bounded in magnitude by $0.5/Q$ (we reserve throughout the paper the symbol ϵ for quantization errors or for vectors of quantization errors). For practical reasons, Q is generally chosen to be an integer power of 2, so that the quantization with $Q = 2^b$ will truncate the fractional part of real numbers to b bits.

In a previous paper [2], an iterative method was proposed for obtaining near-optimal quantized linear prediction coefficients, by observing that the optimum quantized vector w_{opt_Q} is very close to w_{o_Q}, with only some small added differences of $\pm 1/Q$, $\pm 2/Q, \ldots$ for each coefficient. However, the search algorithm in the space of all possible candidates was very laborious, resulting in limited performance improvements when the encoding time was restricted so that real-time encoding is still possible.

We present a simpler and more efficient solution where a very good quantized solution is obtained by optimization arguments rather than brute force search, with a minimal change of the code which computes the unconstrained LS solution, and practically with no added time complexity.

This work was supported by the Academy of Finland (application number 213462, Finnish Programme for Centres of Excellence in Research 2006-2011).

2. DESCRIPTION OF THE INTERLEAVED OPTIMIZATION-QUANTIZATION METHOD

2.1. Optimization-quantization for partitioned Least Squares estimates

The goal is to alleviate the non-optimality of directly quantizing w_o, by splitting w_o into two parts, quantizing one of them, and reoptimizing the second part to account for the quantization already decided.

We consider a split of the vector $w = [w_1^T w_2^T]^T$ into two components w_1 and w_2 of length n_1 and n_2 respectively, and we also define the corresponding partitions of the correlation vector $r = [r_1^T \ r_2^T]^T$ and of the covariance matrix,

$$R = \left[\begin{array}{cc} R_1 & R_3^T \\ R_3 & R_2 \end{array} \right]. \quad (2)$$

The criterion can be rewritten as

$$\begin{aligned} J(w) &= \sigma^2 - 2r_1^T w_1 - 2r_2^T w_2 + w_1^T R_1 w_1 \\ &\quad + w_2^T R_2 w_2 + 2w_1^T R_3^T w_2. \end{aligned} \quad (3)$$

If one fixes the solution $\overline{w}_2$ for the second part of the parameter vector (e.g., taking the last n_2 entries from the quantized w_{o_Q}), the best solution for the free remaining part w_1 can be obtained as

$$R_1 w_1^* = r_1 - R_3^T \overline{w}_2. \quad (4)$$

The new vector formed as $w^* = [w_1^* \ \ \overline{w}_2]$ will correspond to a criterion

$$J(w^*) = \sigma^2 - {w_1^*}^T R_1 w_1^* - 2r_2^T \overline{w}_2 + \overline{w}_2^T R_2 \overline{w}_2 \quad (5)$$

and the criterion for any parameter vector of the form $[w_1^T \ \ \overline{w}_2^T]^T$ can be written

$$J([w_1^T \ \ \overline{w}_2^T]^T) = J(w^*) + (w_1 - w_1^*)^T R_1 (w_1 - w_1^*). \quad (6)$$

We compare now the criteria for the two interesting quantized candidates w_{o_Q} and w_Q^*. Denoting the partition $w_{o_Q} = [\overline{w}_1^T \ \ \overline{w}_2^T]^T$ we get

$$J(w_{o_Q}) = J(w^*) + (\overline{w}_1 - w_1^*)^T R_1 (\overline{w}_1 - w_1^*) \quad (7)$$

$$J(w_Q^*) = J(w^*) + (w_{1_Q}^* - w_1^*)^T R_1 (w_{1_Q}^* - w_1^*). \quad (8)$$

With an argument similar to the one in Section 2.3, one can show that generically $J(w_{o_Q}) > J(w_Q^*)$. To give an intuitive explanation of this we note that the optimization for w_1, when the last part $\overline{w}_2$ is fixed, will always adjust the LS solution to the fact that the last n_2 entries are quantized and fixed.

Subsequently, it becomes natural to perform repeatedly the above process, which involved only one possible split of w. Iteratively we get a longer and longer vector $\overline{w}_2$ and we compute w_1^* by (4) each time, while at next iteration we make the length of $\overline{w}_2$ longer by one, by appending to its top position the last entry from $w_{1_Q}^*$ and continue so until $\overline{w}_2$ gets to the full length of w. Such a process as presented will require solving $n-1$ LS problems, where the size of the unknown vector is successively $n-1, n-2, \ldots, 1$. However, it turns out that the entire process can be implemented by solving a single LS problem of size n, if one resorts to the LDL^T decomposition of the matrix R and at solving the LS problem by back-substitution.

First we show how the partitioned LS problem (4) can be solved using the LDL^T decomposition.

Consider the $n \times n$ covariance matrix decomposed as $R = LDL^T$ and also consider the partitions listed below:

$$L = \left[\begin{array}{cc} L_1 & 0 \\ L_3 & L_2 \end{array} \right]; \ D = \left[\begin{array}{cc} D_1 & 0 \\ 0 & D_2 \end{array} \right] \quad (9)$$

where the blocks R_1, L_1, D_1 have dimensions $n_1 \times n_1$; clearly $R_1 = L_1 D_1 L_1^T$, $R_3 = L_3 D_1 L_1^T$, and $R_2 = L_3 D_1 L_3^T + L_2 D_2 L_2^T$.

We can express the solution (4) in terms of the triangular and diagonal matrices as

$$\begin{aligned} L_1 D_1 L_1^T w_1^* &= r_1 - L_1 D_1 L_3^T \overline{w}_2 \\ L_1^T w_1^* &= D_1^{-1} L_1^{-1} r_1 - L_3^T \overline{w}_2. \end{aligned} \quad (10)$$

Similarly to the unpartitioned optimal solution which can be obtained by back-substitution from $L^T w_o = D^{-1} L^{-1} r$, we can write the block partitions and use (10) to get

$$\left[\begin{array}{cc} L_1^T & L_3^T \\ 0 & L_2^T \end{array} \right] \left[\begin{array}{c} w_1^* \\ \overline{w}_2 \end{array} \right] = \left[\begin{array}{c} D_1^{-1} L_1^{-1} r_1 \\ L_2^T \overline{w}_2 \end{array} \right], \quad (11)$$

which tells that the best solution, which can be obtained by fixing to $\overline{w}_2$ the last n_2 components of the unknown parameters, can be easily solved by back-substitutions in the first n_1 equations of the system (11).

But since our goal is to find only the last entry in w_1^* (not necessarily all entries of w_1^*), then quantize it and append it to the top of the vector $\overline{w}_2$ for preparing the next optimization stage, it becomes clear that our iterative procedure described initially as a sequence of $n-1$ LS problems can be performed "in place" just as a single run of the back-substitution solving, where after solving for an entry of w, the entry is quantized and is considered as part of $\overline{w}_2$ in (11) for the next back-substitution.

2.2. The Optimization-Quantization Least Squares method

We will present now in a detailed way the Optimization-Quantization Least Squares method (OQ-LS) which resulted from the considerations of the previous section. We first decompose the covariance matrix as $R = LDL^T$, with L a lower triangular matrix having all ones on its main diagonal, and D a diagonal matrix with positive elements.

We rewrite the equation for the optimal solution $Rw_o = r$ as $LDL^T w_o = r$, which is equivalent with

$$L^T w_o = D^{-1} L^{-1} r \stackrel{def}{=} g. \quad (12)$$

To get w_o one uses in this last form the back-substitution process, because L^T is an upper triangular matrix.

The back-substitution phase and the fact that L has ones on its main diagonal is the key to our OQ-LS method. We modify the back-substitution algorithm, resulting now in a vector $\tilde{w}$, different of w_o. We will show that the quantized $\tilde{w}_Q$ provides a better solution to the LS criterion than w_{o_Q}. In the original computation of the current component k of w_o by back-substitution we use the line k from L^T,

$$w_{o,k} = g_k - w_{o,k+1} L_{k+1,k} - \ldots - w_{o,M} L_{M,k}. \quad (13)$$

In the alternative computation, the k'th component of $\tilde{w}$ is computed by using the quantized values $\tilde{w}_{k+1_Q}, \ldots, \tilde{w}_{M_Q}$, as follows:

$$\tilde{w}_k = g_k - \tilde{w}_{k+1_Q} L_{k+1,k} - \ldots - \tilde{w}_{M_Q} L_{M,k}. \quad (14)$$

This is used for all $k \in \{M, \cdots, 1\}$ until we determine the full $\tilde{w}$ and the desired quantized version $\tilde{w}_Q$.

The expression for each scalar quantization error can be written

$$\epsilon_k = \tilde{w}_k - \tilde{w}_{k_Q} = (g - L^T \tilde{w}_Q)_k, \quad (15)$$

which is bounded in magnitude, $|\epsilon_k| \leq 0.5 Q^{-1}$, and vector-wise

$$\epsilon = D^{-1} L^{-1} r - L^T \tilde{w}_Q. \quad (16)$$

By substituting $D^{-1} L^{-1} r = L^T w_o$, it results $\epsilon = L^T w_o - L^T \tilde{w}_Q$. If we multiply both terms with L^{-T}, we get $w_o - \tilde{w}_Q = L^{-T} \epsilon$,

which by (1), and using the decomposition $R = LDL^T$ gives

$$J(\tilde{w}_Q) = J(w_o) + (L^{-T}\epsilon)^T (LDL^T)(L^{-T}\epsilon) \quad (17)$$

$$= J(w_o) + \epsilon^T D \epsilon = J(w_o) + \sum_{k=1}^{M} \epsilon_k^2 D_{kk}, \quad (18)$$

which shows how the squared elements of ϵ convert directly to excess minimum squared error in the optimization criterion, simply weighted by D_{kk}. Additionally, we get for a given Q, a strict upper bound for the worst case criterion as

$$J(\tilde{w}_Q) \leq J(w_o) + 0.25 Q^{-2} \sum_{k=1}^{M} D_{kk}. \quad (19)$$

2.3. An approximate analysis

Here we attempt to illustrate the differences between the excess mean square for the two quantized solutions. The relevant terms to compare are $J_1 = (w_{oQ} - w_o)^T R (w_{oQ} - w_o) = \sum_{ij} \epsilon_i' \epsilon_j' r_{ij}$ (from (1)) and $J_2 = (\tilde{w}_Q - w_o)^T R (\tilde{w}_Q - w_o) = \sum_{k=1}^{M} \epsilon_k^2 D_{kk}$ (from (18)).

Since the effect of truncation errors is highly nonlinear, we will attempt just an approximate evaluation of the two terms J_1, J_2 which are deterministic and uniquely defined values, but just for the sake of illustration we make some assumptions about the distribution of the vectors ϵ and ϵ' containing the rounding errors and we will compare the expected values of the expressions J_1, J_2, for a fixed matrix R.

We will assume that all the rounding errors are independent and identically distributed with zero mean and variance σ_ϵ^2, so that the covariance matrices are $E\epsilon\epsilon^T = E\epsilon'\epsilon'^T = \sigma_\epsilon^2 I$.

Furthermore we will assume the matrix R to be symmetric Toeplitz (corresponding to a least square problem where the regressor vectors have the shifting property) for which it is well known that the $(k+1)$'th element, $D_{k+1,k+1} = r_0 - w_o^{[k]T} R^{[k]} w_o^{[k]}$, of the diagonal matrix D in the LDL^T decomposition is the energy of the optimal prediction errors obtained with the optimal predictor, $w_o^{[k]} = R^{[k]-1} r^{[k]}$, of order k (see e.g. [1]). The expected value of the excess mean square for the newly proposed quantization is $EJ_2 = \sigma_\epsilon^2 \sum_{k=0}^{M-1} (r_0 - w_o^{[k]T} R^{[k]} w_o^{[k]})$. The expected value of the excess mean square for the direct quantization of w_o is $EJ_1 = E\epsilon'^T R \epsilon' = trace(RE\epsilon'\epsilon'^T) = M r_0 \sigma_\epsilon^2$. Now it is obvious that $EJ_1 > EJ_2$ due to the fact that each term $r_0 - w_o^{[k]T} R^{[k]} w_o^{[k]}$, except the first one, is much smaller than r_0 for well predictable signals, which is the case with audio signals.

3. IMPLEMENTATION

The lossless audio compressor used in the tests, dubbed here OptimFROG-AS [3], is loosely based on OptimFROG [4], but in an asymmetrical setting. It compresses mono, stereo, and multichannel audio files at any bit depth and sampling rate. It employs stereo prediction [5], adaptive segmentation, and adaptive prediction orders, with prediction coefficients saved in direct form, and a variant of arithmetic coding.

The proposed OQ-LS method can be implemented as a slight modification of the LS solving based on back-substitution. We assume we have the decomposition $R = LDL^T$ and we already obtained $L^T w_o = D^{-1} L^{-1} r = g$. The OQ-LS method differs only in the back-substitution phase, for the computation of $\tilde{w}_Q$.

Computational complexity of the back-substitution phase is the same as the non-quantized LS method using LDL^T decomposition, being $O(M^2)$. In OQ-LS, for each quantization parameter Q we run the back-substitution phase and compute the criterion in $O(M^2)$. In the non-quantized LS method, the exact coefficients are computed only once, requiring $O(M^2)$, then are quantized in $O(M)$, but we also have to separately compute the criterion (1), requiring $O(M^2)$ for each quantization parameter Q. Thus, the overall computational complexities of the two methods are the same.

4. EXPERIMENTAL RESULTS

We tested the proposed OQ-LS method integrated in the lossless audio compressor on a corpus consisting of 80 one minute CD Audio files (44.1 kHz, 16 bit, stereo), produced by extracting the middle minute of the third track from a 80 CD large corpus. For the mono tests, we split each stereo file in two mono files corresponding to left and right channels, producing 160 files.

We compared the following compressors: a) OptimFROG-AS using traditional scalar quantization (OFA-OLD), b) OptimFROG-AS using quantization by the OQ-LS method (OFA-NEW), and c) MPEG-4 ALS standard [6] (Audio Lossless Coding) version RM18 (ALS-V18). We ran the tests on a Intel P4 at 2.8 GHz machine and measured execution times accurately using the total process time.

In order to provide a fair comparison of the efficiency of the coefficient representations, we matched OFA and ALS-V18 so that they will have the same fixed frame size, fixed prediction orders, independent block sizes, and both use arithmetic coding.

For the mono case, we set for OFA the fixed predictor order $n_i = 16$, fixed frame size 1764, independent blocks of 10 seconds, and for equivalence, for ALS-V18 we use: '-b' (BGMC codes), '-g0' (block switching off), '-n1764' (frame size 1764), '-o16' (prediction order 16), and '-r100' (random access frame each 10 seconds).

Compressor	*quantization bits/coeff.*	*Compressed size (%)*	*Encoding time (s)*	*Decoding time (s)*
OFA-NEW	0-8 (LP)	62.2817	175.4	87.4
OFA-NEW	0-15 (LP)	62.2853	184.8	86.3
OFA-NEW	6 (LP)	62.3506	165.8	87.4
OFA-OLD	0-15 (LP)	62.3783	189.5	88.4
OFA-NEW	8 (LP)	62.4000	166.5	88.6
OFA-OLD	0-8 (LP)	62.4137	176.3	87.3
OFA-OLD	8 (LP)	62.5060	165.8	87.8
ALS-V18	6 (RC)	62.5981	185.4	141.0
OFA-OLD	6 (LP)	63.2350	166.2	88.1

(a)

Compressor	*quantization bits/coeff.*	*Compressed size (%)*	*Encoding time (s)*	*Decoding time (s)*
OFA-NEW	0-15 (LP)	59.9096	230.6	86.1
OFA-NEW	0-8 (LP)	59.9379	220.6	85.6
OFA-NEW	6 (LP)	60.0471	208.8	85.6
OFA-OLD	0-15 (LP)	60.0482	231.5	85.5
OFA-NEW	8 (LP)	60.0806	211.9	85.9
OFA-OLD	0-8 (LP)	60.2566	219.6	86.2
OFA-OLD	8 (LP)	60.4659	209.3	85.5
ALS-V18	6 (RC)	60.5786	832.6	144.4
OFA-OLD	6 (LP)	61.8549	207.9	85.7

(b)

Table 1. Overall compressed size (in percents, lower is better) sorted by compressed size; best quantization precision for each frame was searched within the specified bits; (a) mono, (b) stereo.

In Table 1 we can see that the new method achives the best results. The new quantization with 6 bits provides near optimal com-

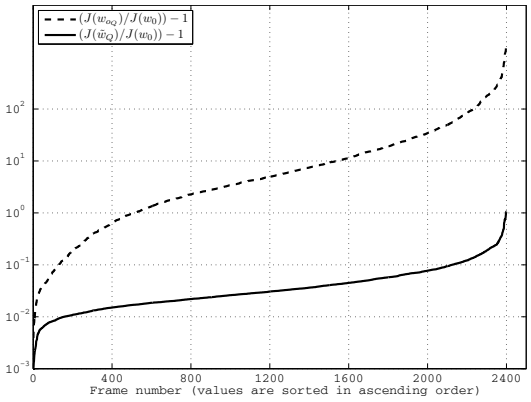

Fig. 1. Misadjustment of the squared error for $Q = 64$.

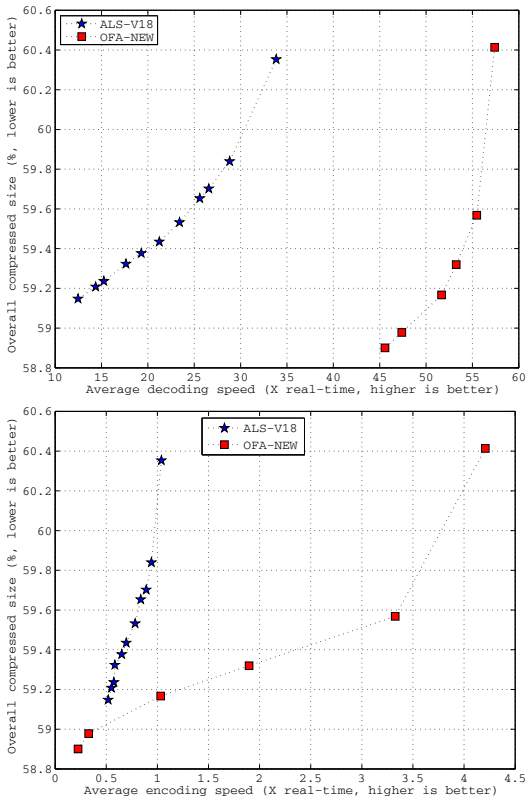

Fig. 2. Overall compression vs. average decoding speed (top) and overall compression vs. average encoding speed (bottom) for ALS-V18 with maximum order 12, 24, 32, 36, 48, 64, 80, 96, 128, 144, and 192 and OFA-NEW with maximum stereo orders 8/4, 16/8, 24/12, 32/16, 48/32, and 64/32 (data points are from right to left).

pression, and ALS-V18 (using quantization with 6 bits for reflection coefficients) provides the second worst compression.

For the stereo case, we set additionally for OFA the fixed side predictor order $n_r = 6$ and for ALS-V18 we added '-t2' which chooses for each frame the best between joint stereo and multichannel correlation (using 3+3 coefficients). The ordering of the overall performance is the same as in the mono case.

To illustrate the performance of the quantization using the OQ-LS method, we used a constant value for Q and investigated the distribution of the misadjustment $J_{ex}(w_{test}) \stackrel{def}{=} J(w_{test})/J(w_o) - 1$, where w_{test} is either w_{o_Q} or $\tilde{w}_Q$. We used one of the stereo files (the one extracted from the 'Enya - Greatest Hits' CD), frame size 2205, and stereo prediction with fixed $n_i = 16$ and $n_r = 8$, all set to match the test made in [2].

For 6 bit quantization, $Q = 64$, the empirical distribution of J_{ex} is shown in Figure 1, from which we also mention the average and maximum values: the direct quantization method obtains $\text{avg}(J_{ex}(w_{o_Q})) = 32.25$, $\max(J_{ex}(w_{o_Q})) = 1826.34$, the OQ-LS algorithm obtains $\text{avg}(J_{ex}(\tilde{w}_Q)) = 0.05$, $\max(J_{ex}(\tilde{w}_Q)) = 1.05$, while [2] obtained an average of 0.17 and a maximum of 5.47, but only when using a large number of iterations, showing that the new quantization method is very effective.

We have compared OFA-NEW and ALS-V18 on the same stereo corpus, for several maximum prediction orders. We used for OFA-NEW adaptive segmentation with maximum 32 segments (ALS also uses up to 32 segments), with a unit size of 588 samples (ALS equivalent is 640 samples). For a fair comparison we used for ALS-V18 optimal compression settings for a given maximum order as "-7 -oMAX_ORDER -t2", but without using long time prediction (LTP).

In Figure 2 we can see that for the same compressed size, the decompression for OFA-NEW is significantly faster than for ALS-V18 (1.7 times faster for low complexity and about 4 times faster for high complexity). The highest compression achieved by ALS-V18 with maximum order 192 is achieved by OFA-NEW with much shorter stereo predictors, bounded above by $n_i = 32$ and $n_r = 16$. For encoding, we can see that for the same compressed size, OFA-NEW is significantly faster than ALS-V18 (4 times faster for low complexity and up to 2 times faster for high complexity).

5. CONCLUSIONS

We have presented a novel method for finding optimally quantized solutions for least squares estimates (OQ-LS) and we applied it to lossless audio compression. When integrated into an asymmetrical lossless audio compressor, the overall achieved performance is superior to the MPEG-4 ALS standard.

6. REFERENCES

[1] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, 4th edition, September 2001.

[2] F. Ghido, "Optimal quantized linear prediction coefficients for lossless audio compression - scalar quantization revisited," in *Proceedings of the 120th Audio Engineering Society Convention (AES 120)*, Paris, France, May 2005.

[3] F. Ghido and I. Tabus, "Adaptive design of the preprocessing stage for stereo lossless audio compression," in *Proceedings of the 122th Audio Engineering Society Convention (AES 122)*, Vienna, Austria, May 2007.

[4] F. Ghido, "OptimFROG lossless audio compressor," on Internet, at *http://www.LosslessAudio.org/*, July 2006, version 4.600ex.

[5] F. Ghido, "An asymptotically optimal predictor for stereo lossless audio compression," in *DCC 2003 Proceedings*. IEEE, March 2003, p. 429.

[6] ISO/IEC, "ISO/IEC 14496-3:2005/Amd 2:2006, Audio Lossless Coding (ALS), new audio profiles and BSAC extensions," on Internet, at *http://www.nue.tu-berlin.de/mp4als*, April 2007, reference software version RM18.