A FRAME-BASED RATE-DISTORTION OPTIMAL CODING SYSTEM USING A LOWER BOUND DEPTH-FIRST-SEARCH STRATEGY

Tom Ryen[1]*, Guido M. Schuster*[2]*, and Aggelos K. Katsaggelos*[3]

[1] Stavanger University College, Department of Electrical and Computer Engineering, P.O.Box 8002, 4068 Stavanger, Norway. E-mail: tom.ryen@tn.his.no
[2] HSR Hochschule für Technik Rapperswil, Abteilung Elektrotechnik, Oberseestrasse 10, 8640 Rapperswil, Switzerland. E-mail: guido.schuster@hsr.ch
[3] Northwestern University, Department of Electrical and Computer Engineering, Evanston, Illinois 60208-3118, USA. E-mail: aggk@ece.nwu.edu

ABSTRACT

The problem of finding the optimal set of quantized coefficients for a frame-based encoded signal is known to be of very high complexity. This paper presents an efficient method of finding the operational Rate-Distortion (RD) optimal set of coefficients. The major complexity reduction lies in the reformulation of the original RD-tradeoff problem, where a new set of coefficients is used as decision variables. These coefficients are connected to the *orthogonalization* of the set of selected frame vectors and not to the frame vectors themselves. Contrary to the original problem, the new problem is practicable to solve optimal in a reasonable amount of time. By organizing all possible solutions as nodes in a solution tree, we use complexity saving techniques to find the optimal solution in an even more efficient way.

1. INTRODUCTION

The use of frame-based coding has been given attention in recent years in topics like lossy compression, sparse signal representation and classification. A frame, or an *overcomplete dictionary*, is a redundant set of column vectors. Consider a one-dimensional signal, $\mathbf{x}$, consisting of NL samples, divided in L blocks of length N. Block no. l, $\mathbf{x}_l \in \mathbb{R}^N$, can be seen as a vector of length N. In frame-based coding, the purpose is to find a best possible approximation to $\mathbf{x}_l$, $\tilde{\mathbf{x}}_l$, as a linear combination of a *small number* of frame vectors. Let $\mathbf{F}$ denote an $N \times K$ matrix whose columns, $\{\mathbf{f}_k\}$, $k = 1, \ldots, K$, constitute a frame. Let $\tilde{\mathbf{w}}_l$ be a vector of length K, where element no. k, $\tilde{w}_{l,k}$, is the quantized coefficient for frame vector $\mathbf{f}_k$. The approximated signal vector, $\tilde{\mathbf{x}}_l$, is

$$\tilde{\mathbf{x}}_l = \mathbf{F}\,\tilde{\mathbf{w}}_l = \sum_{k=1}^{K} \tilde{w}_{l,k}\mathbf{f}_k. \qquad (1)$$

The major part of the coefficients is put to zero, in order to get a sparse representation. The motivation of using an $N \times K$ frame, where $K > N$, instead of an $N \times N$ transform is to have more column vectors to choose from, and thus have a better chance to find a sparse representation that fits the input signal well. Another advantage of frames compared to orthogonal transforms, is that a frame can be trained for a particular class of signals [1]. One disadvantage is that finding the optimal set of selected vectors from the frame and their corresponding coefficients is very hard. An unacceptable computational effort is required. This is why fast, but suboptimal algorithms like Basic Matching Pursuit (BMP) [2], Orthogonal Matching Pursuit (OMP) [3], and Fast Orthogonal Matching Pursuit (FOMP) [4] has been developed and become popular in frame-based coding.

When using frame based coding in a lossy compression scheme, it is natural to use one of the following problem formulations: Minimizing the distortion for a given number of bits, or the dual problem, minimizing the bit rate for a given distortion. This paper is based on the first formulation. The complexity makes this problem unpracticable to solve. In Section 2 we restate the problem formulation and get an efficient algorithm that finds the operational rate-distortion optimal solution, given the frame and the knowledge of the code word lengths. The complexity reduction is connected to the decoupling of the signal blocks and the introduction of a new set of independent coefficients for each signal block. In Section 3 we design a solution tree helping us to get further complexity reduction for the problem defined in Section 2. Experimental results are presented in Section 4, where we compare the new approach performance to Rate-Distortion optimal (RDO) BMP and RDO FOMP [5]. Section 5 concludes this paper.

2. PROBLEM FORMULATION

Consider a frame-based coder where the bit rate of coding coefficient vector no. l, $\tilde{\mathbf{w}}_l$, is $R_l(\tilde{\mathbf{w}}_l)$, and the distortion is $D_l(\tilde{\mathbf{w}}_l)$. We want to find the minimal total distortion subject to a given bit budget, R_{budget}. Our initial optimization problem is

$$\begin{aligned} \min_{\tilde{\mathbf{w}}_l} \quad & \sum_{l=1}^{L} D_l(\tilde{\mathbf{w}}_l) \\ \text{s.t.} \quad & \sum_{l=1}^{L} R_l(\tilde{\mathbf{w}}_l) = R_{budget}. \end{aligned} \tag{2}$$

This problem is extremely hard to solve. In this section we will reason out a new problem without lack of validity, that is solvable in a reasonable amount of time.

First, we apply the Lagrangian Multiplier method [6, 7], to transform the hard constrained problem of (2) into a family of easier, unconstrained problems. As can be seen from the equation below, this results in a decomposing of the original problem into series of independent block optimizations.

$$\begin{aligned} &\min_{\tilde{\mathbf{w}}_l} \Big(\sum_{l=1}^{L} D_l(\tilde{\mathbf{w}}_l) + \lambda \sum_{l=1}^{L} R_l(\tilde{\mathbf{w}}_l) \Big) \\ &= \sum_{l=1}^{L} \Big[\min_{\tilde{\mathbf{w}}_l} \Big(D_l(\tilde{\mathbf{w}}_l) + \lambda R_l(\tilde{\mathbf{w}}_l) \Big) \Big], \end{aligned} \tag{3}$$

for $\lambda \in \mathbb{R}^+$. The optimal solution of this problem is the sum of the best RD-tradeoff for each signal block. In order to find the rate that is equal or close up to the given bit budget in (2) the problem in (3) is solved iteratively to find the appropriate λ-value.

Let the distortion for signal block no. l, $D_l(\tilde{\mathbf{w}}_l)$, be defined as the inner product of the error between the original and the reconstructed signal vector:

$$D_l(\tilde{\mathbf{w}}_l) = \left\| \mathbf{x}_l - \tilde{\mathbf{x}}_l \right\|^2 = \left\| \mathbf{x}_l - \mathbf{F}\, \tilde{\mathbf{w}}_l \right\|^2. \tag{4}$$

When using frames, we expect sparse coefficient vectors, i.e., a large number of the K elements in $\tilde{\mathbf{w}}_l$ is zero. Therefore it is convenient to use a Run-Length Coder (RLC) as a part of the entropy coder. The RLC counts the number of zeros between each nonzero coefficient. After the last nonzero coefficient in each block, we use an End Of Block (EOB) symbol to indicate the start of the next block. Each nonzero coefficient has a value taken from a given set of *values*. We use two different Variable Length Code (VLC) tables, one for the coefficient *values* and one for the *runs* between the nonzero coefficients. We can now define the rate for signal block l, R_l, as

$$R_l(\tilde{\mathbf{w}}_l) = \sum_{k \in nz} (R_{l,k}^{val} + R_{l,k}^{run}) + R^{EOB}, \tag{5}$$

where nz is the set of indices for the nonzero coefficients, $R_{l,k}^{val}$ and $R_{l,k}^{run}$ the number of bits used to code the *value* and the *run* for the k-th coefficient, respectively, and R^{EOB} the number of bits needed to code the EOB symbol.

Even though the signal blocks are independent, there are still dependencies between the coefficients within the same block, both in rate and distortion. The complexity of the minimization in (3) is high, since for every block we need to search between all possible ways to place M nonzero coefficient in a coefficient vector of length K, where $M = 1$, $M = 2$, and so on. The total number of combinations with M nonzero coefficients is $\binom{K}{M}$. In addition, each nonzero coefficient can take on a given number of different *value* symbols, J. For each combination, the number of solutions is J^M. Thus, the total number of different solutions for one single signal block is $\sum_{M=0}^{M_{max}} \binom{K}{M} J^M$, where M_{max} is the largest number of nonzero coefficients we choose to use per block. In all practical cases, $M_{max} << K$, due to the sparse representation idea. A representative number of elements in the coefficient vector is $K = 32$. Let $J = 32$ and $M = \{1, 2, 3, 4\}$. This leads to $\binom{K}{M} = \{32, 496, 4960, 35960\}$ and $J^M = \{32, 1024, 32768, 1048576\}$. A way to get rid of the latter combinatorial explosion is presented in [8]: For each combination of M out of K nonzero coefficients, $M = \{1, 2, \ldots, M_{max}\}$, the *QR-decomposition* [9] is found for the corresponding M frame vectors. Let $\mathbf{\Phi}_l = \mathbf{Q}_l \mathbf{R}_l$ (of size $N \times M$) be the set of selected frame vectors, where the order in $\mathbf{\Phi}_l$ is the same as the order the selected vectors have in $\mathbf{F}$. $\mathbf{Q}_l$ is an $N \times M$ matrix with orthonormal vectors and $\mathbf{R}_l$ is an $M \times M$ upper triangular matrix. For all cases, $M_{max} < N$, and $span(\mathbf{\Phi}_l) \subset \mathbb{R}^N$. The best signal reconstruction we can get when using *continuous valued* coefficients, $\hat{\mathbf{x}}_l$, is found by using the Best Approximation Theorem [9]:

$$\hat{\mathbf{x}}_l = \mathbf{\Phi}_l (\mathbf{\Phi}_l^T \mathbf{\Phi}_l)^{-1} \mathbf{\Phi}_l^T \mathbf{x}_l = \mathbf{Q}_l \mathbf{Q}_l^T \mathbf{x}_l. \tag{6}$$

The error is orthogonal to any vector spanned by the column vectors in $\mathbf{\Phi}_l$. Due to Pythagoras, the distortion in block l, D_l, can be written as the sum of two errors,

$$D_l = \|\mathbf{x}_l - \tilde{\mathbf{x}}_l\|^2 = \|\mathbf{x}_l - \hat{\mathbf{x}}_l\|^2 + \|\hat{\mathbf{x}}_l - \tilde{\mathbf{x}}_l\|^2. \tag{7}$$

The first term in (7) is fixed when the set of selected frame vectors, $\mathbf{\Phi}_l$, is known. For the second term, let a new coefficient vector with continuous valued elements and length M, $\mathbf{v}_l^o$, be such that $\hat{\mathbf{x}}_l = \mathbf{Q}_l \mathbf{v}_l^o$. These coefficients are easily found by

$$\mathbf{v}_l^o = \mathbf{Q}_l^T \hat{\mathbf{x}}_l = \mathbf{Q}_l^T \mathbf{x}_l. \quad (8)$$

$\mathbf{v}_l^o$ is a constant vector as soon as $\mathbf{\Phi}_l$ are known. A new vector of *quantized* coefficients, $\tilde{\mathbf{v}}_l^o$, is connected to the reconstructed signal such that $\tilde{\mathbf{x}}_l = \mathbf{Q}_l \tilde{\mathbf{v}}_l^o$. The elements of $\tilde{\mathbf{v}}_l^o$ can take on values listed in the *value* codeword table. *These coefficients are our new decision variables.* The second term in (7) can be written as

$$\|\hat{\mathbf{x}}_l - \tilde{\mathbf{x}}_l\|^2 = \|\mathbf{v}_l^o - \tilde{\mathbf{v}}_l^o\|^2 = \sum_{m=1}^{M} (v_{l,m}^o - \tilde{v}_{l,m}^o)^2. \quad (9)$$

The distortion, D_l, can now be written as a constant added to the sum of *independent* coefficient distortions. The number of bits used to code the *run* codewords and the EOB symbol are constants. The rate for the *value* codewords is now a sum of *independent* coefficient value rates. For a given combination of M nonzero coefficients, we need to find the combination's $\mathbf{\Phi}_l$, $\mathbf{Q}_l$ and $\mathbf{R}_l$, and then minimize with respect to $\tilde{\mathbf{v}}_l^o$

$$\begin{aligned} & \|\mathbf{x}_l - \mathbf{Q}_l \mathbf{Q}_l^T \mathbf{x}_l\|^2 \\ + \quad & \lambda\Big(\sum_{m=1}^{M} (R_{l,m}^{run}) + R^{EOB} \Big) \\ + \quad & \sum_{m=1}^{M} \min_{\tilde{v}_{l,m}^o} \big((v_{l,m}^o - \tilde{v}_{l,m}^o)^2 + \lambda R_{l,m}^{val} \big). \end{aligned} \quad (10)$$

The problem is now much faster to solve, since only MJ comparisons are necessary, compared to the J^M comparisons needed in our original problem. To find the optimal rate-distortion tradeoff, we must solve (10) for all $\binom{K}{M}$ combinations for $M = \{1, \ldots, M_{max}\}$, and store the minimum of all solutions. When working with low bit rate compression, the maximum number of nonzero coefficients per block is low. Thus, M_{max} could be a small number, and the time used to find the optimal solution could be substantially lower than with a higher M_{max}. It should be mentioned that the decoder needs a QR-decomposition depending on each coefficient vector's nonzero coefficient indices, in addition to the knowledge of the frame and the VLC tables used in the encoder.

3. THE SOLUTION TREE

In this section we organize all the $\sum_{M=0}^{M_{max}} \binom{K}{M}$ possible ways of combining up to M_{max} vectors from the frame. The purpose of building a tree, is to be able to use techniques that gives us additional complexity reduction of coding single signal vectors optimally.

Consider a tree where each node represents a unique set of selected vectors from the frame $\mathbf{F}$, i.e., a unique $\mathbf{\Phi}_l$. The root node represents the selection of *zero* vectors ($M = 0$). It has K children nodes. Each of these nodes represents the selection of exactly *one* frame vector ($M = 1$). The edges that connect the root node to its children nodes is named "1", "2", ... ,"K", to indicate the index of the frame vector that is selected. The node that represents the selection of frame vector no. 1, $\mathbf{f}_1$, has $K-1$ children nodes. These children represent the selection of *two* vectors ($M = 2$), where the first vector of $\mathbf{\Phi}_l$ is $\mathbf{f}_1$ and the second is $\mathbf{f}_i$, where $i = \{2, 3, \ldots, K\}$, respectively. The node representing $\mathbf{\Phi}_l = \mathbf{f}_2$ has $K-2$ children, the node with $\mathbf{\Phi}_l = \mathbf{f}_3$ has $K-3$ children, and so on. Thus, the node with $\mathbf{\Phi}_l = \mathbf{f}_{K-1}$ has one child and the node representing $\mathbf{\Phi}_l = \mathbf{f}_K$ has none. In Figure 1, an illustration of this tree is shown. It will have M_{max} generations, or levels. Level M will have $\binom{K}{M}$ nodes. The entire tree will represent all possible combinations of $0, 1, \ldots, M_{max}$ vectors selected. In each node, we find the minimum rate-distortion solution by solving (10) for the given combination. To find the global optimum solution for the signal block, we need to search through all nodes in the tree.

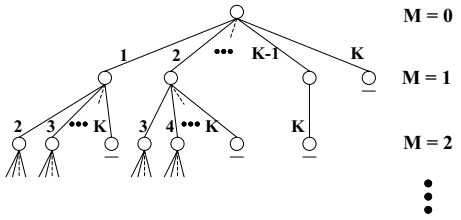

Figure 1: The solution tree. Each node represents the minimum rate-distortion solution given a unique set of selected vectors. The root node has *zero* vectors selected, each node in the 1. generation has *one* selected vector, each node in the 2. generation has *two* selected vectors, and so on. The entire tree represent all possible combinations of selecting up to M_{max} of K frame vectors.

3.1. Time reduction by depth-first-search

By choosing *depth-first-search* (DFS) as the search strategy in this tree, we can build the QR-decomposition for each node recursively. There are two benefits by using DFS: There is less computation in order to find the QR-decomposition, and only one coefficient has to be found in order to find the best set of coefficients for the respective node.

DFS starts in the root node, next it goes to the first child, then to the first child of the first child, and so on, until a node in level M_{max} is reached. Then, it backtracks to the node in level $M_{max} - 1$, where the next child to this node is visited, and so on. In any node in the tree, the next step in the DFS strategy is as follows: If there are unvisited children, go to one of them. If not, backtrack to parent node. DFS stops when all nodes in the tree are visited.

Suppose that we know the minimum cost, $\min C_l$, for a parent node in level $M-1$,

$$\min C_{l,M-1} = \|\mathbf{e}_{l,M-1}\|^2 + \alpha_{l,M-1} + \beta_{l,M-1}, \quad (11)$$

where

$$\begin{aligned}
\mathbf{e}_{l,M-1} &= \mathbf{x}_l - \mathbf{Q}_{l,M-1}\mathbf{Q}_{l,M-1}^T\mathbf{x}_l, \\
\alpha_{l,M-1} &= \lambda\big(\sum_{m=1}^{M-1}(R_{l,m}^{run}) + R^{EOB}\big), \\
\beta_{l,M-1} &= \sum_{m=1}^{M-1}\min_{\tilde{v}_{l,m}^o}\big((v_{l,m}^o - \tilde{v}_{l,m}^o)^2 + \lambda R_{l,m}^{val}\big).
\end{aligned}$$

$\mathbf{Q}_{l,M-1}$ $(N \times M-1)$ is the orthonormal basis in the QR-decomposition of $\mathbf{\Phi}_{l,M-1}$, the set of selected frame vectors. When going from a parent node to a child node, we use the same set of selected frame vectors, only adding a new frame vector, $\phi_{l,M}$, as the last column in the set of selected frame vectors, $\mathbf{\Phi}_{l,M}$ $(N \times M)$, that is

$$\mathbf{\Phi}_{l,M} = \left[\begin{array}{cc} \mathbf{\Phi}_{l,M-1} & \phi_{l,M} \end{array}\right]. \quad (12)$$

The new frame vector will always be added to the right end of the matrix, since it has a higher frame index value than all the previously selected vectors. When using the *Gram-Schmidt process* [9] to find the QR-decomposition, we know that column vector no. i in the orthonormal basis is only a function of column vector 1 to i in the original matrix. Thus, the first $M-1$ columns in the orthonormal basis of the QR-decomposition of $\mathbf{\Phi}_{l,M}$, $\mathbf{Q}_{l,M}$, will be equal to the parent's orthonormal basis, $\mathbf{Q}_{l,M-1}$. Since we know $\mathbf{Q}_{l,M-1}$, only the last vector, $\mathbf{q}_{l,M}$, has to be found by Gram-Schmidt to get the entire $\mathbf{Q}_{l,M}$:

$$\mathbf{Q}_{l,M} = \left[\begin{array}{cc} \mathbf{Q}_{l,M-1} & \mathbf{q}_{l,M} \end{array}\right], \quad (13)$$

where

$$\mathbf{q}_{l,M} = \frac{\phi_{l,M} - \sum_{m=1}^{M-1}\langle\phi_{l,M}, \mathbf{q}_{l,m}\rangle\mathbf{q}_{l,m}}{\left\|\phi_{l,M} - \sum_{m=1}^{M-1}\langle\phi_{l,M}, \mathbf{q}_{l,m}\rangle\mathbf{q}_{l,m}\right\|}. \quad (14)$$

The minimum solution for a child node in level M is

$$\min C_{l,M} = \|\mathbf{e}_{l,M}\|^2 + \alpha_{l,M} + \beta_{l,M}, \quad (15)$$

By (13), the difference between $\mathbf{x}_l$ and its projection on $span(\mathbf{\Phi}_l)$ can be written as

$$\begin{aligned}
\mathbf{e}_{l,M} &= \mathbf{x}_l - \mathbf{Q}_{l,M}\mathbf{Q}_{l,M}^T\mathbf{x}_l \\
&= \mathbf{x}_l - \mathbf{Q}_{l,M-1}\mathbf{Q}_{l,M-1}^T\mathbf{x}_l - \mathbf{q}_{l,M}\mathbf{q}_{l,M}^T\mathbf{x}_l \\
&= \mathbf{e}_{l,M-1} - \mathbf{q}_{l,M}\mathbf{q}_{l,M}^T\mathbf{x}_l. \quad (16)
\end{aligned}$$

The first $M-1$ vectors are not changed, thus the first $M-1$ *run* codewords are the same. A new *run* codeword is added, and $\alpha_{l,M}$ is

$$\begin{aligned}
\alpha_{l,M} &= \lambda\big(\sum_{m=1}^{M}(R_{l,m}^{run}) + R^{EOB}\big) \\
&= \lambda\big(\sum_{m=1}^{M-1}(R_{l,m}^{run}) + R^{EOB}\big) + \lambda R_{l,M}^{run} \\
&= \alpha_{l,M-1} + \lambda R_{l,M}^{run}. \quad (17)
\end{aligned}$$

Each element in coefficient vector $\tilde{\mathbf{v}}_l^o$ is found separately according to (10). Due to (8) the first $M-1$ continuous valued coefficients, $[v_{l,1}^o, \ldots, v_{l,M-1}^o]^T$, will be the same when the first $M-1$ vectors in $\mathbf{Q}_{l,M}$ is not changed. The quantized coefficients, $[\tilde{v}_{l,1}^o, \ldots, \tilde{v}_{l,M-1}^o]^T$, will be the same, and only the last added coefficient, $\tilde{v}_{l,M}^o$, needs to be found. $\beta_{l,M}$ is found by

$$\begin{aligned}
\beta_{l,M} &= \sum_{m=1}^{M}\min_{\tilde{v}_{l,m}^o}\big((v_{l,m}^o - \tilde{v}_{l,m}^o)^2 + \lambda R_{l,m}^{val}\big) \\
&= \sum_{m=1}^{M-1}\min_{\tilde{v}_{l,m}^o}\big((v_{l,m}^o - \tilde{v}_{l,m}^o)^2 + \lambda R_{l,m}^{val}\big) \\
&\quad + \min_{\tilde{v}_{l,M}^o}\big((v_{l,M}^o - \tilde{v}_{l,M}^o)^2 + \lambda R_{l,M}^{val}\big) \\
&= \beta_{l,M-1} + \min_{\tilde{v}_{l,M}^o}\big((v_{l,M}^o - \tilde{v}_{l,M}^o)^2 + \lambda R_{l,M}^{val}\big). \quad (18)
\end{aligned}$$

To find $\mathbf{q}_{l,M}$ in (14) we need $2MN+1$ multiplications and $MN-1$ additions when using Gram-Schmidt. If we would need to use Gram-Schmidt to find the entire $\mathbf{Q}_{l,M}$, the number of multiplications and additions would be $\sum_{m=1}^{M} 2mN+1$ and $\sum_{m=1}^{M} mN-1$, respectively. In all practical cases, $N \gg 1$, thus the number of multiplications and additions is approximately $2NM(M+1)/2$ and $NM(M+1)/2$, respectively. Computing only $\mathbf{q}_{l,M}$ instead of $\mathbf{Q}_{l,M}$ will result in a reduction in the number of multiplications and additions equal to $(M+1)/2$. The complexity reduction factor is larger when M is getting larger, i.e., using DFS strategy is more important for higher M_{max}.

Due to the knowledge of the parent node's local minimum, the number of comparisons needed to find the local minimum of the child node in level M is reduced. The first $M-1$ elements in the coefficient vector is the same as the parent's coefficients. Only the last coefficient must be found. In (18) the number of comparisons is reduced from MJ to J, a reduction factor equal to M.

3.2. Time reduction by introduction of lower bounds

In order to find the global optimum solution, all the nodes in the solution tree have to be visited. But,

M_{max}	2	3	4
Original problem	1	320	74411
Orthogonalized problem	0.064	1	10
Ort. problem using DFS	0.033	0.35	2.61
Orthogonalized problem using DFS and LB	0.014	0.21	1.59

Table 1: Time units relative to coding one signal block using the original formulation in (3) when $M_{max} = 2$. $K = 32$ and $J = 32$ for all cases.

finding the optimal solution in each node is not necessarily a requirement. Let C^* be the *so far* minimum cost. As an initial value, C^* is set to the root node cost, $C_{l,0} = \|\mathbf{x}_l\|^2 + R^{EOB}$, i.e., the cost of selecting none frame vectors. C^* is updated every time one node's solution is better than C^*. For a node in level M, the *lower bound* for the minimum node cost, LB, is given by

$$LB = \|\mathbf{e}_{l,M}\|^2 + \alpha_{l,M} + \lambda M R^{val}_{min}, \qquad (19)$$

where R^{val}_{min} is the minimum number of bits for a *value* codeword. If $C^* < LB$ for a particular node, we know that the global optimum solution is not represented by the particular set of frame vectors, and further computation to find $\beta_{l,M}$ is not required. The DFS algorithm proceeds to next node. In the worst case, $C^* \geq LB$ in all nodes, and a full computation is necessary. But, in an average case many nodes don't need full computation, and the time saved by the introduction of lower bounds is significant. Let us show an example that verify this statement. We use the proposed algorithm to code a Gaussian AR(1) process with $\rho = 0.95$. We use a 16×32 trained frame. The number of *value* symbols is $J = 32$. In Table 1 we show the time consumption when using 1) the original problem in (3), 2) the orthogonalized version in (10), 3) the orthogonalized version by DFS strategy, and 4) the orthogonalized version by DFS strategy and lower bounds (LB) technique. The experiment is done for $M_{max} = \{2, 3, 4\}$. The time units in the table are relative to coding one signal block using the original formulation when $M_{max} = 2$. The largest time reduction is achieved by the orthogonalization of the problem, and this is most important when M_{max} is large. When using the DFS strategy, the time is reduced with an additional factor of M_{max}. What is interesting to see is that the time is further reduced by a factor of 2 caused by using lower bounds in each node in the tree. This is a significant contribution for speeding up the algorithm without loosing optimality.

4. EXPERIMENTS

The optimality of the rate-distortion solution described in previous sections depends on the given frame and the VLC tables. The design of frames [10] is not considered in this work, but we will show the performance of using a trained frame compared to using a Discrete Cosine Transform (DCT) and a random generated frame. The VLC tables are optimized on a training signal, according to the approach in [8].

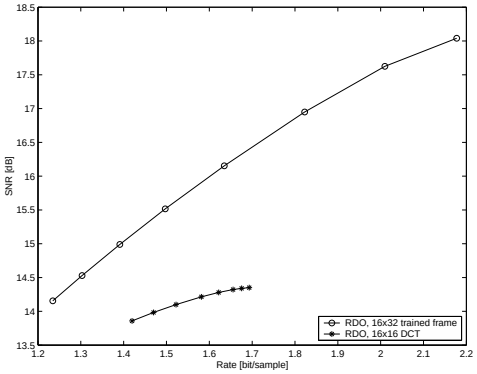

Figure 2: Rate-distortion function for a set of 8 λ-values. The new approach used on a 16×32 trained frame and a 16×16 DCT.

In the following experiments, all signals are Gaussian AR(1) processes with $\rho = 0.95$. The number of signal blocks, $L = 512$, and the number of samples per block, $N = 16$. We use a 16×32 frame trained for selecting 4 out of 32 vectors using OMP [10]. We call this frame *fr4*. The second frame, *frRand*, is a 16×32 random generated frame, where all columns are normalized. The transform used is a 16×16 DCT. In all experiments, $M_{max} = 4$, and $\lambda = 0.0002$.

First, we will motivate for using a designed frame instead of a transform in a RD optimization scheme. Of course, the time complexity is much less when using an orthogonal transform, but the rate-distortion gain is much higher when using the trained frame. This is shown in Figure 2. The signal-to noise ratio (SNR) is given by

$$SNR = 10 \log \frac{\|\mathbf{x}\|^2}{\sum_l D_l}. \qquad (20)$$

We can see from the diagram that using the 16×32 trained frame gives a much better result than using the DCT.

In [5] the RDO Basic Matching Pursuit (BMP) and RDO Fast Orthogonal Matching Pursuit (FOMP) were presented. Both algorithms has the minimization of the rate-distortion tradeoff as their object function, in contradiction to standard BMP and FOMP, where the minimization of the distortion is the object. We will compare these algorithms to our proposed approach. The RDO Matching Pursuit techniques encode the coefficient *indices* and not the *runs* between them. Therefore, we introduce an operational RD optimal encoder (ORDE) where the nonzero coefficient *indices* are encoded in addition to the ORDE where the *runs* are encoded. In Figure 3 we show

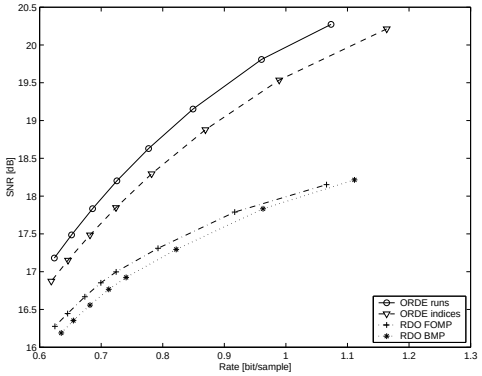

Figure 3: Rate-distortion function for a set of 8 λ-values. ORDE using run-length coding (ORDE runs) and index coding (ORDE indices) compared to RDO BMP and RDO FOMP. Frame used is *fr4*.

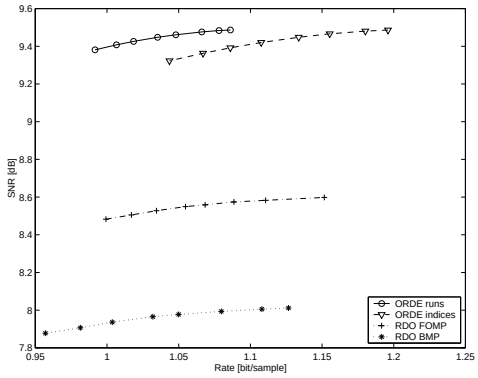

Figure 4: Rate-distortion function for a set of 8 λ-values. ORDE using run-length coding (ORDE runs) and index coding (ORDE indices) compared to RDO BMP and RDO FOMP. Frame used is *frRand*.

the results of using the new approach for coding the AR(1) process compared to the RDO Matching Pursuit techniques. The *fr4* frame is used. The rate-distortion tradeoffs are found for 8 λ-values in the ratio $0.0001 - 0.0008$. The results show that the new approach is better than the Matching Pursuit techniques, due to the claim of optimality. RDO BMP and RDO FOMP are much faster algorithms, but suboptimal. The new approach using *index* coding is between 0.7 dB and 2 dB better than Matching Pursuit at the same bit rates. In this experiment, the use of run-length coding results in an SNR that is ca. 0.5 dB better than by coding the indices directly.

In the last experiment, the frame is replaced to be *frRand*. Keeping all other parameters unchanged, the diagram in Figure 4 shows us the same relation as in the previous experiment. The new perform is better than RDO BMP and RDO FOMP, especially when using run-length coding. Notice the much lower SNR when using a random frame instead of a well designed frame.

5. CONCLUSION

An efficient method for rate-distortion optimal frame-based coding is presented in this paper. The efficiency is achieved by using a QR-decomposition which results in a new set of independent decision variables. This involved a considerably reduction in complexity, and made it possible to find the optimal solution in a reasonable amount of time. Further complexity reduction is gained by representing the problem as a tree, using depth-first-search strategy and introducing lower bounds in each node. Experiments show that this method outperforms RDO Matching Pursuit. We also see the benefit of using a well designed frame through the experiments.

6. REFERENCES

[1] K. Engan, S. O. Aase, and J. H. Husøy, "Multi-frame compression: Theory and design," *Signal Processing*, vol. 80, pp. 2121–2140, Oct. 2000.

[2] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[3] G. Davis, *Adaptive Nonlinear Approximations*, Ph.D. thesis, New York University, Sept. 1994.

[4] M. Gharavi-Alkhansari and T. S. Huang, "A fast orthogonal matching pursuit algorithm," in *IEEE Proc. ICASSP '98*, Seattle, USA, May 1998, pp. 1389–1392.

[5] M. Gharavi-Alkhansari, "A model for entropy coding in matching pursuit," in *IEEE Proc. ICIP '98*, Chicago, USA, Nov. 1998, pp. 778–782.

[6] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion Based Video Compression*, Kluwer Academic Publishers, Boston, 1997.

[7] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, pp. 23–50, Nov 1998.

[8] T. Ryen, G. M. Schuster, and A. K. Katsaggelos, "A rate-distortion optimal coding alternative to matching pursuit," in *IEEE Proc. ICASSP '02*, Orlando, USA, May 2002, pp. 2177–2180.

[9] H. Anton, *Elementary Linear Algebra*, John Wiley and Sons, Inc., New York, 7th edition, 1994.

[10] K. Engan, *Frame Based Signal Representation and Compression*, Ph.D. thesis, Norwegian University of Science and Technology/ Stavanger University College, Sept. 2000.