TRELLIS QUANTIZATION OF FRAMES

Mohamed F. Mansour
Texas Instruments Inc.
DSP Solutions R&D Center
Dallas, Texas, USA

ABSTRACT

We address the optimization of the quantization of overcomplete frames. The optimization problem is formulated as a quadratic integer programming problem. We develop a suboptimal dynamic programming solution that uses a scalable trellis expansion. The proposed solution offers significant improvement of the quantization error and it can be combined with projection-based quantization, e.g., sigma-delta quantization. The effectiveness of the algorithm is established using harmonic tight frames.

Index Terms— Frames, quantization, dynamic programming, trellis.

1. INTRODUCTION

Quantization of the coefficients of frame expansion is an important design issue in many practical systems. This problem has been studied in some earlier work (e.g., [1], [2], [3]) and many important design criteria and error bounds were described. Oversampled frames are frequently encountered in many common signal processing systems, e.g., sigma-delta converter, discrete wavelet transform, and Gabor transform.

In this work, we study the problem using an optimization theory perspective. In particular, we formulate the optimization problem as an integer programming problem with a quadratic cost function. We study two optimization criteria and show a general optimization model with different weight operators. In the finite dimensional case the problem is reduced to the conventional quadratic integer programming model. This problem is known to be NP-hard. Therefore, we develop a suboptimal dynamic programming solution to the optimization problem. The metric in the trellis expansion is derived from the objective function of the optimization problem. The proposed algorithm can be combined with other quantization procedure, e.g., sigma-delta quantization to further improve the performance.

In our discussion we assume that the domain of the frame operator is $L^2(R)$ or any subspace of it. We will use the following notations: $I$ is the identity operator, $\lfloor \cdot \rfloor$ is the floor integer function. $l^2(Z)$ is the space of square summable sequences. Underlined italic letters denote vectors (i.e., elements of $l^2(Z)$), $\text{Ran}(T)$ is the range of the operator $T$, and $\ker(T)$ is the kernel of the operator $T$.

2. BACKGROUND

The set $\{\phi_j\}_{j \in J}$ is a frame of a Hilbert space $H$ if there exists two positive numbers $A, B$ such that for any $x \in H$ we have

$$A \|x\|^2 \leq \sum_{j \in J} |\langle x, \phi_j \rangle| \leq B \|x\|^2 \quad (1)$$

A frame is tight if $A = B$. Frames [4] can be viewed as generalization of bases of a Hilbert space. If $\{\phi_j\}_{j \in J}$ is an overcomplete frame, then any element $x \in H$ can be expressed as

$$x = \sum c_j \phi_j \quad (2)$$

where $c = \{c_j\}_{j \in J}$ may not be unique. The analysis operator $T : H \rightarrow l^2(Z)$ and the adjoint (synthesis) operator $\tilde{T} : l^2(Z) \rightarrow H$ are defined as:

$$Tf \triangleq \{(f, \phi_j)\}_{j \in J} \quad (3)$$

$$\tilde{T}\{c_j\} \triangleq \sum_{j \in J} c_j \phi_j \quad (4)$$

Note that, if $c$ satisfies (2), then any solution of the form $c + h$ where $h \in \ker(\tilde{T})$ (or equivalently $h \perp \text{Ran}(T)$) also satisfies (2).

The frame operator $F : H \rightarrow H$ is the self-adjoint operator $\tilde{T}^* T$. The dual frame $\{\tilde{\phi}_j\}_{j \in J}$ is defined as:

$$\tilde{\phi}_j = F^{-1} \phi_j$$

for tight frames we have $\tilde{\phi}_j = \phi_j/A$.

One possible solution of particular importance for $\{c_j\}$ in (2) is the minimum-norm solution, which is computed as:

$$\hat{c}_j = \langle x, \tilde{\phi}_j \rangle$$

Any other solution of (2) can be expressed as

$$c = \hat{c} + \sum_n \gamma_n Z^{(n)} \quad (5)$$
3. OPTIMIZATION PROBLEM

The objective of the optimization problem is to minimize the reconstruction error due to the quantization of the frame coefficients. Define the quantization error $Q(\varepsilon)$ as:

$$Q(\varepsilon) = \varepsilon + \sum_{n \in N_K} \gamma_n L^{(n)} - (\|\varepsilon\| + \alpha)$$  \hspace{1cm} (6)

where $\{L^{(n)}\}_{n \in N_K}$ are as defined in (5) and $\alpha$ is an integer vector that represents the quantization approximation. For example, if $Q(c_j) = [c_j]$, then $\alpha_j = 0$, and if $Q(c_j) = [c_j]$, then $\alpha_j = 1$. Let

$$\varepsilon = \varepsilon - \|\varepsilon\|$$  \hspace{1cm} (7)

Note that $\tilde{T}$ $(\sum_{n \in N_K} \gamma_n L^{(n)}) = 0$ because $\{L^{(n)}\} \in \ker(\tilde{T})$. Hence, the reconstruction error is:

$$e^{\text{rec}} = \tilde{T}(Q(\varepsilon)) = \tilde{T}(\varepsilon - \alpha) = \sum_{j \in J} (e_j - \alpha_j) \phi_j$$  \hspace{1cm} (8)

Therefore the optimization problem can be written as minimizing the objective function:

$$\Psi = \|e^{\text{rec}}\|^2 = \langle \tilde{T}(\varepsilon - \alpha), \tilde{T}(\varepsilon - \alpha) \rangle$$

$$\Psi = \langle L(\varepsilon - \alpha), \varepsilon - \alpha \rangle$$  \hspace{1cm} (9)

where

$$L = T\tilde{T}$$  \hspace{1cm} (10)

In some quantization scenarios, e.g., rounding and sigma-delta quantization, the objective of the optimization problem is to minimize the quantization error rather than minimizing the reconstruction error due to quantization. In the absence of oversampling, the two criteria are equivalent. In the following we show that minimizing the quantization error leads to an objective function similar to (9) with a different self-adjoint operator. The objective function in this case is to minimize:

$$\Psi = \|Q(\varepsilon)\|^2$$  \hspace{1cm} (11)

where $Q(\varepsilon)$ is as defined in (6). This objective function can be written as:

$$\Psi = \|\gamma + (\varepsilon - \alpha)\|^2$$  \hspace{1cm} (12)

where $\gamma \in \ker(\tilde{T})$, $\varepsilon$ is as defined in (7). For any value of $\alpha$, the value of $\gamma$ that yields the minimization of (12) is equivalent to the projection of $(\varepsilon - \alpha)$ onto $\ker(\tilde{T})$, i.e.,

$$\gamma_j = -\langle \varepsilon - \alpha, \varepsilon^{(j)} \rangle$$  \hspace{1cm} (13)

Define the projection operator $P : l^2(Z) \rightarrow \ker(\tilde{T})$ as:

$$P(\varepsilon) = \sum_{j \in N_K} \langle \varepsilon, \varepsilon^{(j)} \rangle \varepsilon^{(j)}, \text{where } \varepsilon \in l^2(Z)$$  \hspace{1cm} (14)

Then the objective function in (12) can be rewritten as:

$$\Psi = \| (I - P)(\varepsilon - \alpha) \|^2$$  \hspace{1cm} (15)

Note that $(I - P)$ is the projection operator onto $\text{Ran}(\tilde{T})$, hence it is a self-adjoint operator [5]. Therefore, the optimization problem has the same form as in (9) but with

$$L = (I - P)^2 = (I - P)$$  \hspace{1cm} (16)

Note that by applying Schwartz inequality on (8) we get:

$$\|e^{\text{rec}}\| = \|\tilde{T}(Q(\varepsilon))\| \leq \|\tilde{T}\| \cdot \|Q(\varepsilon)\|$$  \hspace{1cm} (17)

The resulting optimization problem is a quadratic integer programming problem. This problem is known to be NP and several heuristic approaches were developed for suboptimal solutions [6, 7]. Typically, search techniques, e.g., Tabu search or simulated annealing, are used to find a solution. These techniques are not practical in some applications where sequential quantization is need, e.g., in oversampled data converters. In the following section, we describe a suboptimal dynamic programming algorithm that is suited for this class of applications.

4. TRELIS QUANTIZATION

In the following analysis, we focus on minimum reconstruction error optimization in (9), however, exact analysis can be applied to the minimum quantization error problem. Using (3), (4) and after straightforward arithmetic, the objective function in (9) can be written as:

$$\Psi = \sum_{j \neq j} (e_j - \alpha_j)(e_t - \alpha_t)\langle \phi_j, \phi_t \rangle$$  \hspace{1cm} (18)

where $e_j$ and $\alpha_j$ are the $j$-th components of $\varepsilon$ and $\alpha$ respectively. For notational simplicity, denote

$$e_j \triangleq e_j - \alpha_j$$  \hspace{1cm} (19)

Note that,

$$\langle \phi_j, \phi_t \rangle = \langle \phi_t, \phi_j \rangle^*$$

by substituting in (18) we get,

$$\Psi = \sum_{j < t} 2\Re\{e_j e_t^* \langle \phi_j, \phi_t \rangle\} + \sum_t |e_t|^2 \langle \phi_t, \phi_t \rangle$$  \hspace{1cm} (20)
To solve the quantization problem using a dynamic programming technique we need to decompose the objective function to a sequential formula. Define

\[ J_r = \sum_{t \leq \tau} \sum_{j < i} 2Re\{ \varepsilon^* \varepsilon_j \phi_j, \phi_i \} + \sum_{t \leq \tau} |\varepsilon_t|^2 \phi_t, \phi_t \]  

(21)

Note that, \( \Psi = \lim_{\tau \to \infty} J_r \). A recursive formula for \( J_r \) is:

\[ J_{\tau+1} = J_r + \sum_{t < \tau+1} 2Re\{ \varepsilon_t \varepsilon_{\tau+1}^* \phi_t, \phi_{\tau+1} \} + |\varepsilon_{\tau+1}|^2 \phi_{\tau+1}, \phi_{\tau+1} \]  

(22)

To get the optimal solution using the above formula, we need to keep track of all the previous quantization states to evaluate the middle term in (22). This results in an exponential growth of the middle term in (22). In the simplest case, each state at time \( \tau \) represents one possible quantization value of the frame coefficient \( c_\tau \) and there is a possible transition between each pair of states at times \( \tau \) and \( \tau + 1 \). In this case the recursive metric for the \( k \)-th state at time \( \tau + 1 \) is computed as:

\[ J^{(k)}_{\tau+1} = \max_k \left\{ J^{(i)}_r + \sum_{t < \tau+1} 2Re\{ \varepsilon_t \varepsilon^*_{\tau+1} \phi_t, \phi_{\tau+1} \} + |\varepsilon^*_{\tau+1}|^2 \phi_{\tau+1}, \phi_{\tau+1} \right\} \]  

(23)

Note that, \( \{ \varepsilon^{(i)}_t \}_{t \leq \tau} \) specifies the trellis path up to state \( i \) at time \( \tau \). If the best previous state is \( p \), then the updated path for state \( k \) at \( \tau + 1 \) is specified by:

\[ \varepsilon^{(k)}_t = \varepsilon^{(p)}_t, \quad \text{for} \quad t = 1, 2, \ldots, \tau \]  

(24)

and \( \varepsilon^{(k)}_{\tau+1} \) is specified by the quantization value of the \( k \)-th state. The above expansion is suboptimal because the relation between the quantization of the \( (\tau + 1) \)-th coefficient and the quantization of coefficients of index less than \( \tau \) is evaluated only through the trellis path ending at time \( \tau \). Therefore not all possible combinations of these quantizations are tested.

Note that, if the quantization resolution \( \beta \) is more than one bit, then the number of states at each trellis step grows exponentially as \( 2^\beta \). However, as the best quantization value is usually the upper or lower integer values, the number of states can be kept to only two that represent the upper and lower integer quantization values of the \( \tau \)-th coefficient. This simplification gives a huge saving in the algorithm complexity with small performance degradation.

The proposed algorithm works with both finite and infinite dimensional frames. In fact, the infinite dimensional frames give in general better performance because the correlation between successive frame vectors (in time) is in general bigger than the correlation between distant vectors. Therefore, the first term of the objective function in (20) is well approximated by the trellis expansion.

It should be mentioned that the proposed trellis quantization is an alternative to the simple rounding of the frame coefficients at the quantizer. It can be combined with other noise shaping algorithms for frame quantization, e.g., \([1]\) and \([8]\), by replacing the rounding stage in these algorithms with the proposed trellis expansion. Note that, the proposed trellis quantization does not involve any feedback loop, therefore it does not affect the system stability.

5. SIMULATION RESULTS

In our simulation we use geometrically uniform tight frames that are generated as described in \([3]\). The test vectors in \( R^N \) are zero-mean Gaussian iid sequences with unity variance. The frame functions \( \{ \phi_j \} \) in the synthesis process are not quantized.

In the first experiment, we evaluate the proposed trellis quantization scheme. We fix the frame size to \( N = 4 \), and increase the redundancy factor \( r \), i.e., the total number of vectors in the frame is \( M = rN \). In Fig. 1, we compare the reconstruction error due to quantization of the trellis quantization scheme and the rounding procedure with quantization resolutions \( \beta = 1, 8 \). The rounding error behaves as \( O(1/r) \) when \( \beta = 8 \) as the additive uniform error model \([9]\) for the quantization error is valid for this case. When \( \beta = 1 \), this additive model is no longer valid and the performance of the rounding procedure saturates at high redundancy factors. On the other hand, the trellis quantization gives a consistent and significant improvement at higher redundancy factors because it solves the optimization model (20) regardless of the redundancy factor. The trellis expansion behaves almost as \( O(1/r^4) \) (12 dB/octave) at high redundancy factors.

In the second experiment, we verify the versatility of the proposed trellis algorithm by combining it with another frame quantization procedure \([1]\) by replacing the rounding procedure by trellis quantization. The quantization algorithm in \([1]\) projects the quantization error after quantizing the \( \tau \)-th coefficient onto the space spanned by the \( (\tau + 1) \)-th frame vector. It resembles a first-order sigma-delta modulator. Therefore the quantization behavior is \( O(1/r^3) \). When this algorithm is combined with trellis quantization of coefficients (after projection), the performance is consistently better than \( O(1/r^4) \) at all redundancy factors as illustrated in Fig. 2.

6. DISCUSSION

The paper studies the problem of quantizing oversampled frames from an optimization theory perspective. The two main contributions of this work are:

1. Modeling the frame quantization problem as a quadratic
integer programming problem as in (9). We showed that a similar problem model (with different self-adjoint operators) is obtained for both minimum reconstruction error due to quantization and minimum quantization error. The explicit solution to this problem is known to be NP-hard. Therefore all practical frame quantization algorithms are in fact suboptimal.

2. Providing a suboptimal dynamic programming procedure for solving the optimization problem. The metric of the dynamic programming problem is derived from (9). We showed that the proposed algorithm offers significant improvement in the reconstruction error. Moreover, the proposed algorithm can be combined with noise shaping quantization algorithms.

Future work will extend the proposed trellis quantization to sigma-delta modulators which is an example of infinite-dimensional frames.

7. REFERENCES


