QUANTIZATION NOISE SHAPING ON ARBITRARY FRAME EXPANSIONS

Petros T. Boufounos and Alan V. Oppenheim

Massachusetts Institute of Technology - Digital Signal Processing Group
77 Massachusetts Avenue, Rm. 36-615, Cambridge, MA 02139
{petrosb,avo}@mit.edu

ABSTRACT

Quantization noise shaping is commonly used in oversampled A/D and D/A converters. This paper considers quantization noise shaping for arbitrary frame expansions of signals based on generalizing the view of first order noise shaping as a compensation of the quantization error through a projection. Two levels of generalization are developed, one a special case of the other, and two different cost models are proposed to evaluate the quantizer structures. Within our framework, the implementation of the quantizer and reconstruction are computationally straightforward. The computational complexity is in the initial determination of frame vector ordering, which is part of the quantizer design and carried out off-line. We show that in the case of frame representation corresponding to uniform oversampling, the natural ordering implied by sequential time sampling is optimal. Furthermore, for general finite frame expansions, the problem of optimal ordering corresponds to known problems in graph theory.

1. INTRODUCTION

Quantization methods for frame expansions have received considerable attention in the last few years. Simple scalar quantization applied independently on each frame expansion coefficient, followed by linear reconstruction, is well known to be suboptimal [1, 2]. Several algorithms have been proposed that improve performance although with significant complexity either at the quantizer [3] or in the reconstruction method [3, 4]. More recently, frame quantization methods inspired by uniform oversampled noise shaping (referred to generically as Sigma-Delta noise shaping) have been proposed for oversampled filterbanks [5] and for uniform tight frames [6]. Our approach is along the lines of that proposed in [6] but incorporates the use of projections and explores the issue of frame vector ordering. It is applicable to arbitrary frame expansions, both finite and infinite, and provides a design method for optimal first order noise shaping quantizers.

We view noise shaping as compensation of the error from quantizing each frame expansion coefficient through a projection to the space defined by another synthesis frame vector. This requires only knowledge of the synthesis frame set and a pre-specified ordering and pairing for the frame vectors. As we show, this improves the error in reconstruction due to quantization, even for non-redundant frame expansions (i.e. a basis set) when the frame vectors are non-orthogonal.

In section 2 we present a brief summary of frame representations to establish notation and we describe classical first-order Sigma-Delta quantizers in the terminology of frames. In section 3 we propose two generalizations, which we refer to as the sequential quantizer and the tree quantizer, both assuming a known ordering of the frame vectors. Section 4 proposes two different cost models for determining the frame vector ordering, one based on a stochastic representation of the error and the other on deterministic upper bounds. In section 5 we determine the optimal ordering of coefficients assuming the cost measures in section 4 and show that for Sigma-Delta noise shaping the natural (time-sequential) ordering is optimal. We also show that for finite frames the determination of frame vector ordering can be formulated in terms of known problems in graph theory.

2. CONCEPTS AND BACKGROUND

2.1. Frame representation

A vector $\mathbf{x}$ in a space $\mathcal{W}$ of finite dimension N is represented with the finite frame expansion:

$$\mathbf{x} = \sum_{k=1}^{M} a_k \mathbf{f}_k, \; a_k = \langle \mathbf{x}, \bar{\mathbf{f}}_k \rangle \tag{1}$$

The space $\mathcal{W}$ is spanned by both sets: the synthesis frame vectors $\{\mathbf{f}_k, k = 1, \ldots, M\}$, and the analysis frame vectors

This work was supported in part by: participation in the Advanced Sensors Collaborative Technology Alliance (CTA) sponsored by the U.S. Army Research Laboratory under Cooperative agreement DAAD19-01-2-008, the Texas Instruments Leadership University Consortium Program, BAE Systems Inc., and MIT Lincoln Laboratory. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

$\{\bar{\mathbf{f}}_k, k = 1, \ldots, M\}$. This condition ensures that $M \geq N$. Details on the relationships of the analysis and synthesis vectors can be found in a variety of texts such as [1]. The ratio $r = M/N$ is referred to as the *redundancy* of the frame.

2.2. Sigma-Delta noise shaping

Oversampled signals are a well studied example of frame expansions. A signal $x[n]$ or $x(t)$ is upsampled or oversampled to produce a sequence a_k. In the terminology of frames, the upsampling operation is a frame expansion in which $\bar{\mathbf{f}}_k[n] = r\mathbf{f}_k[n] = \mathrm{sinc}(\pi(n-k)/r)$,with $\mathrm{sinc}(x) = \sin(x)/x$. The sequence a_k is the corresponding ordered sequence of frame coefficients:

$$a_k = \langle x[n], \bar{\mathbf{f}}_k[n] \rangle = \sum_n x[n]\mathrm{sinc}(\pi(n-k)/r) \quad (2)$$

$$x[n] = \sum_k a_k \mathbf{f}_k[n] = \sum_k a_k \frac{1}{r}\mathrm{sinc}(\pi(n-k)/r) \quad (3)$$

with similar expressions in continuous time. Sigma-Delta quantizers can be represented in a number of equivalent forms [7]. The representation shown in Figure 1 most directly represents the view that we extend to general frame expansions. Performance of Sigma-Delta quantizers is typically analyzed using an additive white noise model for the quantization error. Based on this model it is straightforward to show that the in-band quantization noise power is minimized when the scaling coefficient c is chosen to be $c = \mathrm{sinc}(\pi/r)$. With typical oversampling ratios, this coefficient is close to unity and is often chosen as unity for computational convenience.

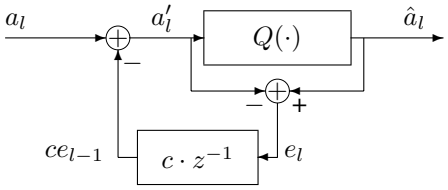

Fig. 1. Traditional first order noise shaping quantizer

The process in figure 1 can be viewed as a sequence of coefficient quantization and error projection. Consider $x_l[n]$, such that the coefficients up to a_{l-1} have been quantized and e_{l-1} has already been scaled by c and subtracted from a_l to produce a'_l:

$$x_l[n] = \sum_{k=-\infty}^{l-1} \hat{a}_k \mathbf{f}_k[n] + a'_l \mathbf{f}_l[n] + \sum_{k=l+1}^{+\infty} a_k \mathbf{f}_k[n] \quad (4)$$

$$= x_{l+1}[n] + e_l(\mathbf{f}_l[n] - c \cdot \mathbf{f}_{l+1}[n]) \quad (5)$$

The incremental error $e_l(\mathbf{f}_l[n] - c \cdot \mathbf{f}_{l+1}[n])$ at the l^{th} iteration of (5) is minimized if we pick c such that $c \cdot \mathbf{f}_{l+1}[n]$ is the projection of $\mathbf{f}_l[n]$ onto $\mathbf{f}_{l+1}[n]$:

$$c = \langle \mathbf{f}_l[n], \mathbf{f}_{l+1}[n] \rangle / ||\mathbf{f}_{l+1}[n]||^2 = \mathrm{sinc}(\pi/r) \quad (6)$$

This choice of c projects the error of quantizing a_l to $\mathbf{f}_{l+1}[n]$ and compensates for this error by modifying a_{l+1}.

3. NOISE SHAPING ON FRAMES

In this section we propose two generalizations of the discussion of section 2.2 to general frame representations.

3.1. Single coefficient quantization

Throughout this discussion we assume the ordering of the analysis frame vectors, denoted by $(\mathbf{f}_1, \ldots, \mathbf{f}_M)$, and correspondingly the ordering of the coefficients $(a_1, \ldots, a_M)$ is predetermined. Issues of determining the optimal ordering are addressed in sections 4 and 5. It should be emphasized that the ordering can be determined off-line using only the synthesis frame vector set, not the specific signal being represented or the expansion frame vectors, i.e. it is part of the off-line design of the quantizer.

To illustrate our approach, we consider quantizing the first coefficient a_1 to $\hat{a}_1 = a_1 + e_1$, with e_1 denoting the additive quantization error. Equation (1) then becomes:

$$\mathbf{x} = \hat{a}_1 \mathbf{f}_1 + \sum_{k=2}^{M} a_k \mathbf{f}_k - e_1 \mathbf{f}_1. \quad (7)$$

As in (5), we then perform the projection using coefficient $c_{1,2}$ to obtain:

$$a'_2 = a_2 - e_1 c_{1,2} \quad (8)$$

where $c_{i,j}$ is, in general, different for each pair of frame vectors ($\mathbf{f}_i$,$\mathbf{f}_j$):

$$c_{i,j} = \langle \mathbf{f}_i, \mathbf{u}_j \rangle / ||\mathbf{f}_j|| \quad (9)$$

where $\mathbf{u}_k = \mathbf{f}_k / ||\mathbf{f}_k||$ are the unit vectors in the direction of the synthesis frame vectors. After the projection, the residual component has direction

$$\mathbf{r}_{1,2} = (\mathbf{f}_1 - c_{1,2}\mathbf{f}_2) / ||\mathbf{f}_1 - c_{1,2}\mathbf{f}_2|| \quad (10)$$

The corresponding error is $e_1 \langle \mathbf{f}_1, \mathbf{r}_{1,2} \rangle \mathbf{r}_{1,2} = e_1 \tilde{c}_{1,2} \mathbf{r}_{1,2}$, where $\tilde{c}_{1,2} = \langle \mathbf{f}_1, \mathbf{r}_{1,2} \rangle$ is the *error coefficient* for this pair of vectors. Substituting the above, equation (7) becomes

$$\mathbf{x} = \hat{a}_1 \mathbf{f}_1 + a'_2 \mathbf{f}_2 + \sum_{k=3}^{M} a_k \mathbf{f}_k - e_1 \tilde{c}_{1,2} \mathbf{r}_{1,2} \quad (11)$$

The component $e_1 \tilde{c}_{1,2} \mathbf{r}_{1,2}$ is the final quantization error after one step is completed. We iterate the process above by quantizing the next (updated) coefficient until the last coefficient has been quantized. We call this procedure the *sequential* first order noise shaping quantizer.

3.2. The tree noise shaping quantizer

The sequential quantizer can be generalized by relaxing the sequence of error assignments: Again, we assume that the coefficients have been pre-ordered and that the ordering defines the sequence in which coefficients are quantized. However, we associate with each frame vector $\mathbf{f}_k$ another possibly not adjacent frame vector $\mathbf{f}_{l_k}$ further in the sequence (and therefore for which the corresponding coefficient has not yet been quantized) to which the error is projected using equation (8). With this more generalized approach some frame vectors can be used to compensate for more than one quantized coefficient. For finite frames, this defines a tree, in which every node is a frame vector or associated coefficient. If a coefficient a_k uses coefficient a_{l_k} to compensate for the error, then a_k is a child of a_{l_k} in that tree. The root of the tree is the last coefficient to be quantized, a_M. We refer to this as the *tree* noise shaping quantizer. The sequential quantizer is, of course, a special case of the tree quantizer where $l_k = k + 1$.

The resulting expression for $\mathbf{x}$ is given by:

$$\mathbf{x} = \sum_{k=1}^{M} \hat{a}_k \mathbf{f}_k - \sum_{k=1}^{M-1} e_k \tilde{c}_{k,l_k} \mathbf{r}_{k,l_k} - e_M \mathbf{f}_M \quad (12)$$

$$= \hat{\mathbf{x}} - \sum_{k=1}^{M-1} e_k \tilde{c}_{k,l_k} \mathbf{r}_{k,l_k} - e_M ||\mathbf{f}_M|| \mathbf{u}_M \quad (13)$$

where $\hat{\mathbf{x}}$ is the quantized version of $\mathbf{x}$, after noise shaping and the e_k are the quantization errors in the coefficients *after* the correction of the $k-1$ iteration has been applied to a_k. Thus, the total additive error of the process is:

$$\mathcal{E} = \sum_{k=1}^{M-1} e_k \tilde{c}_{k,l_k} \mathbf{r}_{k,l_k} + e_M ||\mathbf{f}_M|| \mathbf{u}_M \quad (14)$$

4. ERROR MODELS AND ANALYSIS

To simplify the analysis of eq. (14) we focus on cost measures for which the incremental cost at each step is independent of the whole path and the data. We call these *incremental* cost functions. In this section we examine two such models, one stochastic and one deterministic.

4.1. Additive noise model

The first cost function assumes the additive uniform white noise model for quantization error, and minimizes the expected energy of the error $E\{||\mathcal{E}||^2\}$. All the error coefficients e_k are assumed white and identically distributed,with variance $\Delta^2/12$, where Δ is the interval spacing of the quantizer. They are also assumed to be uncorrelated with the quantized coefficients. Thus, all error components contribute additively to the error power, resulting in:

$$E\{||\mathcal{E}^2||\} = \frac{\Delta^2}{12} \left(\sum_{k=1}^{M-1} |\tilde{c}_{k,l_k}|^2 + ||\mathbf{f}_M||^2 \right) \quad (15)$$

4.2. Error magnitude upper bound

As an alternative cost function, we can also consider an upper bound for vector addition to analyze equation (14). For any set of vectors $\mathbf{u}_i$, $||\sum_k \mathbf{u}_k|| \leq \sum_k ||\mathbf{u}_k||$, with equality only if all vectors are collinear, in the same direction. This leads to the following upper bound on the error:

$$||\mathcal{E}|| \leq \frac{\Delta}{2} \left(\sum_{k=1}^{M-1} |\tilde{c}_{k,l_k}| + ||\mathbf{f}_M|| \right), \quad (16)$$

The vector $\mathbf{r}_{M-1,l_{M-1}}$ is by construction orthogonal to $\mathbf{f}_M$ and the $\mathbf{r}_{k,l_k}$ are never collinear, making the bound very loose. Thus, a shaping quantizer can be expected in general to perform better than what the bound suggests.

4.3. Error analysis

In terms of the above cost functions for any quantization ordering, the sequential or tree quantizers perform better than or equal to direct scalar quantization of the frame expansion coefficients. This is true even if the frame is not redundant. The equality holds if and only if all the pairs $(\mathbf{f}_k, \mathbf{f}_{l_k})$ of the quantizer are orthogonal. Note also that independent of the coefficient ordering the cost function has a multiplicative term of $\Delta^2/12$ or $\Delta/2$. These terms only depend on the design of the quantization intervals, and we ignore them when comparing orderings. For uniform frames the additive contribution of the last coefficient quantized, $||\mathbf{f}_M||^2$ or $||\mathbf{f}_M||$, can also be ignored since it is the same for any ordering. For infinite frames there is no "last" coefficient. Thus, the comparison in both cases only depends on the coefficients $\tilde{c}_{k,l_k}$. We use these properties in considering the optimal ordering.

5. NOISE SHAPING QUANTIZER DESIGN

As indicated earlier, the essential issue in quantizer design based on the strategies outlined in this paper is determining the ordering of the frame vectors. The optimal ordering depends on the specific set of synthesis frame vectors, but not on the specific signal. Consequently, the quantizer design (i.e. the frame vector ordering) is carried out off-line and the quantizer implementation is a sequence of projections based on the ordering chosen for either the sequential or tree quantizer.

In this section we propose simple design strategies, and a lower bound on the cost of the minimum-cost quantizer for a given frame. We show that the Sigma-Delta noise shaping

quantizer meets this lower bound. Furthermore, in the case of finite frames we map the design problem to well-studied graph theory problems.

5.1. Simple design strategies

An obvious design strategy is to pair the coefficients such that the quantization of every coefficient a_k is compensated as much as possible by the coefficient a_{l_k}. This can be achieved if for any $\mathbf{f}_k$ we use the $\mathbf{f}_{l_k}$ that is closest to compensate for the error, i.e. that minimizes $|\tilde{c}_{k,l_k}|$. If this strategy is possible to implement, it results in the optimal ordering under both the cost models we discussed. In fact, this is exactly how a traditional Sigma-Delta quantizer works, and, therefore, it is one of the optimal first order structures given the oversampled frame (another optimal example is the anticausal version). A structure generated using this approach is optimal under both the cost models we discussed.

For certain frames, however, this optimal pairing might not be possible. The resulting ordering might contain cycles whereby the quantization of one coefficient a_k should be compensated using an already quantized coefficient a_{l_k}. However, even this case, this pairing can be useful in providing a loose lower bound for the cost of the optimal quantizer. Furthermore, it suggests a heuristic for a good coefficient pairing: at every step k, the error from quantizing coefficient a_k is compensated using the coefficient a_{l_k} that can compensate for most of the error, picking from all the frame vectors whose corresponding coefficients have not yet been quantized. This is a suboptimal but implementable heuristic. We discuss optimal design in the next section.

5.2. Optimal design for finite uniform frames

From section 3.2 it is clear that a tree quantizer can be represented as a graph—specifically, a tree—in which all the nodes of the graph are coefficients to be quantized. Similarly for a sequential quantizer, which is a special case of the tree quantizer: the graph is a linear path passing through all the nodes a_k in the correct sequence. In both cases, the graphs have edges (k, l_k), pairing coefficient a_k to coefficient a_{l_k} if and only if the quantization of coefficient a_k assigns the error to the coefficient a_{l_k}.

Assuming a uniform synthesis frame, we can measure the total cost of the quantizer—subject to the transformations of section 4.3—using this graph. To do so, we assign to every edge (k, l_k) of the graph a weight $w(k, l_k) = |\tilde{c}_{k,l_k}|^2$ or $w(k, l_k) = |\tilde{c}_{k,l_k}|$ for the additive noise model or the error upper bound cost functions respectively. By summing the total weight of all the edges in the graph, we deduce the total weight of the graph, i.e. the total cost of the corresponding quantizer.

Using this graph, designing the optimal first order quantizer for a given uniform finite frame corresponds to well studied graph-theory problems. The minimum cost sequential quantizer corresponds to the solution of the *traveling salesman problem (TSP)*, i.e. the minimum weight path that goes through all the nodes. The TSP is NP-complete in general, but has been extensively studied in the literature[8]. Similarly, the minimum cost tree quantizer is given by the *minimum spanning tree*, another well studied problem, solvable in polynomial time [8]. Since any path is also a tree, if the minimum spanning tree is an acyclic path through the graph, then that is also the solution to the traveling salesman problem. The solutions to these two problems are the optimal sequential or tree quantizer for a given frame, respectively. Note that in general the optimal ordering and pairing depend on which of the two cost functions we choose.

6. REFERENCES

[1] I. Daubecies, *Ten Lectures on Wavelets*, CBMS-NSF regional conference series in applied mathematics. SIAM, Philadelphia, PA, 1992.

[2] Z. Cvetkovic and M. Vetterli, "Overcomplete expansions and robustness," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, 1996.*, Jun 1996, pp. 325–328.

[3] V.K. Goyal et al., "Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 16–31, Jan 1998.

[4] N.T. Thao and M. Vetterli, "Reduction of the MSE in R-times oversampled A/D conversion $O(1/R)$ to $O(1/R^2)$," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 200–2003, Jan 1994.

[5] H. Bolcskei and F. Hlawatsch, "Noise reduction in oversampled filter banks using predictive quantization," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 155–172, Jan 2001.

[6] J.J. Benedetto et al., "Sigma-delta quantization and finite frames," in *Proceedings of IEEE ICASSP 2004*, Montreal, Canada, May 2004, IEEE.

[7] J.C. Candy and G.C Temes, Eds., *Oversampling Delta-Sigma Converters*, IEEE Press, 1992.

[8] H.T. Cormen et al., *Introduction to algorithms*, MIT Press and McGraw-Hill, 2nd edition, 2001.