43rd IEEE Conference on Decision and Control
December 14-17, 2004
Atlantis, Paradise Island, Bahamas

Multi-step Optimal Quantization in Oversampled Filter Banks

Daniel E. Quevedo*
dquevedo@ieee.org

Graham C. Goodwin*
eegcg@ee.newcastle.edu.au

Helmut Bölcskei†
boelcskei@nari.ee.ethz.ch

***Abstract*— Using concepts from the receding horizon control framework, we propose a novel approach to quantization in oversampled filter banks. The key idea is to pose the quantization problem as a multi-step optimization problem, where the decision variables are restricted to belong to a finite set. It is shown that the resulting architecture yields enhanced performance when compared to the well-known noise shaping coder. In particular, the quantizer proposed can be tuned with stability concepts in mind.**

I. Introduction

Oversampled Filter Banks (FBs) are widespread in many practical applications, mainly in relation to subband signal processing and coding of audio and image signals, see e.g. [1]–[6]. Fig. 1 depicts such a scheme. The main purpose of these schemes is to compress a discrete time signal y into sequences u_i, such that the amount of bits deployed is small, whilst providing a good reconstructed signal, i.e. $\hat{y} \approx y$. For that purpose, a series connection of an *analysis FB*, an *abstract quantizer* $\mathcal{Q}$ and a *synthesis FB* is deployed. As can be seen in Fig. 1, an oversampled FB first decomposes the discrete-time input signal y into subband sequences v_i through analysis filtering and downsampling. The abstract quantizer $\mathcal{Q}$ processes the v_i (in a linear or nonlinear fashion) and quantizes them resulting in the output sequences u_i. The synthesis FB then up-samples and filters the sequences u_i to provide an estimate of y.

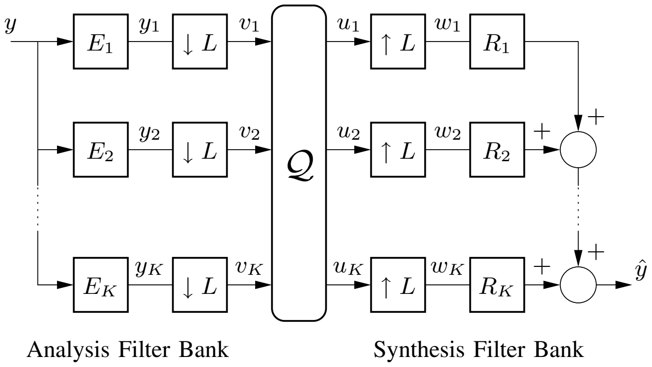


Fig. 1. K-channel FB with sub-sampling by factor L.

Reconstruction errors in an oversampled FB basically arise from three sources: aliasing and imaging (as a consequence of sampling rate conversions), phase and amplitude distortion (of the filters deployed), and quantization and coding effects. In the common case of linear FBs, the effects of aliasing, imaging, phase and amplitude distortion can be dealt with by proper filter design, see e.g. [2]–[4]. On the other hand, even if the subband signals are not processed and lossless coding is employed (which we will assume in this work), quantization introduces loss of information, and commonly precludes $\hat{y}$ from being equal to y.

Quantization effects are nonlinear (and non-smooth) in nature. Nonetheless, at times, linear analysis and design methodologies can be deployed, see e.g. [7], [8]. Of course, these techniques are only approximate and one can expect linear design methods to be, in general, outperformed by nonlinear ones.

Here we present a novel quantization/coding method for oversampled FBs which does not utilize a linear model of quantization. The scheme is aimed at minimizing the reconstruction error. For that purpose, it utilizes concepts from finite-set constrained Predictive Control, see e.g. [9]. The methodology extends our previous work on scalar analog-to-digital conversion [10]–[12] and contains the *Noise Shaping Quantizer* in [8] as a special case.

The remainder of this paper is organized as follows: In Section II, we introduce some basic concepts in relation to (oversampled) FBs. Section III presents the proposed multi-step quantization approach. Its properties are investigated in Section IV. Section V documents a simulation study. Section VI draws conclusions.

II. Background on (Oversampled) Filter Banks

In this section we will give a brief review of some fundamental aspects of (oversampled) FBs. More thorough treatments can be found e.g. in [2]–[4], [13], [14].

A. Basic Aspects

As can be seen in Fig. 1, the discrete-time input signal y is simultaneously filtered by the analysis filters $E_i(z)$, $i \in \{1, 2, \dots K\}$. The K signals y_i are then down-sampled by a factor of L, i.e. only every L-th element is retained. The sampling rate of the resultant subband sequences $v_i(\ell)$ is L times lower than the sampling rate of y, i.e. there is a change in time scale. For example, with $L = 2$, we have:

$$\begin{array}{cccccccc} y_i: & y_i(0) & y_i(1) & y_i(2) & y_i(3) & y_i(4) & y_i(5) & \dots \\ & \downarrow & & \downarrow & & \downarrow & & \\ v_i: & y_i(0) & & y_i(2) & & y_i(4) & & \dots \end{array}$$

*School of Electrical Engineering & Computer Science, The University of Newcastle, Callaghan, NSW 2308, Australia

†Communication Technology Laboratory, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland

0-7803-8682-5/04/$20.00 ©2004 IEEE

The subband sequences are then quantized resulting in the K sequences u_i so that at each instant $\ell \in \mathbb{Z}$:

$$u_i(\ell) \in \mathbb{U}_i, \quad \forall i \in \{1, 2, \dots, K\}, \tag{1}$$

where the $\mathbb{U}_i$ are given finite sets. By allowing for different quantization sets $\mathbb{U}_i$, with subband coding, the available bit budget can be utilized more efficiently than if y was quantized directly, see Section II-D. The performance of the FB depends on how the sequences u_i are chosen inside the abstract quantizer $\mathcal{Q}$. The design of $\mathcal{Q}$ is the main focus of the current paper.

In the synthesis FB, the quantized sequences u_i are up-sampled by a factor of L. This is achieved by padding the sequences with zeroes and gives rise to the signals w_i. For instance, in the case of $L = 3$, we have:

$$\begin{array}{lccccccccccc} u_i\colon & u_i(0) & & & u_i(1) & & & u_i(2) & & \dots \\ & \downarrow & & & \downarrow & & & \downarrow & & \\ w_i\colon & u_i(0) & 0 & 0 & u_i(1) & 0 & 0 & u_i(2) & 0 & \dots \end{array}$$

The K filters R_i then perform an interpolation/smoothing function and provide the estimate $\hat{y}(\ell)$.

Most Subband Coding FBs, correspond to the so-called *maximally decimated* case, where $K = L$. Recently, *oversampled* FBs, where $K > L$ have been studied in the context of frame expansions, see e.g. [8], [13], [14]. In particular, in [8] it is shown how redundancies in the representation of y via v_i can be used in order to mitigate quantization noise effects. In the present work, we include both situations by allowing $K \geq L$.

B. Polyphase Representation

The oversampled FB of Fig. 1 is a multi-rate system, which can conveniently be represented in terms of the polyphase description, see e.g. [2], [7], [14], described in the following. We start by defining the vector sequences:

$$\begin{aligned} v(\ell) &\triangleq \begin{bmatrix} v_1(\ell) & v_2(\ell) & \dots & \dots & v_K(\ell) \end{bmatrix}^T \\ u(\ell) &\triangleq \begin{bmatrix} u_1(\ell) & u_2(\ell) & \dots & \dots & u_K(\ell) \end{bmatrix}^T \\ d(\ell) &\triangleq \begin{bmatrix} d_1(\ell) & d_2(\ell) & \dots & d_L(\ell) \end{bmatrix}^T \\ \hat{d}(\ell) &\triangleq \begin{bmatrix} \hat{d}_1(\ell) & \hat{d}_2(\ell) & \dots & \hat{d}_L(\ell) \end{bmatrix}^T, \quad \text{where:} \end{aligned}$$

$$d_i(\ell) \triangleq y(\ell L + i - 1), \quad \hat{d}_i(\ell) \triangleq \hat{y}(\ell L + i - 1) \tag{2}$$

are the polyphase components of y and $\hat{y}$, respectively. Note that all these vectors are updated only every L sampling instants of y.

Given these definitions and by using the so-called *noble identities*, see e.g. [2], [7], direct algebraic manipulation yields that the FB of Fig. 1 can be characterized via:

$$v(\ell) = E(z)d(\ell), \quad \hat{d}(\ell) = R(z)u(\ell). \tag{3}$$

In this expression, $E(z)$ is the $K \times L$ *analysis polyphase matrix*, while $R(z)$ is the $L \times K$ *synthesis polyphase matrix*. They are defined via $[E(z)]_{i,n} = E_{i,n}(z)$ and $[R(z)]_{n,i} = R_{i,n}(z)$, $i = 1, 2, \dots, K$, $n = 1, 2, \dots, L$, where:

$$\begin{aligned} E_{i,n}(z) &\triangleq \sum_{\ell=-\infty}^{\infty} e_i(\ell L - n + 1) z^{-\ell}, \\ R_{i,n}(z) &\triangleq \sum_{\ell=-\infty}^{\infty} r_i(\ell L + n - 1) z^{-\ell}, \end{aligned}$$

and e_i and r_i are the following impulse responses:

$$E_i(z) = \sum_{\ell=-\infty}^{\infty} e_i(\ell) z^{-\ell}, \quad R_i(z) = \sum_{\ell=-\infty}^{\infty} r_i(\ell) z^{-\ell}.$$

The resultant scheme is depicted in Fig. 2. As can be seen, the polyphase description consists of a MIMO single-rate system, which operates on the down-sampled signals. The multi-rate aspect is implicit, see (2). Since the filter operations are performed only at the reduced sampling rate, the polyphase representation is not only more compact, but also more computationally efficient than the structure of Fig. 1. It avoids unnecessary operations due to up- and down-sampling.

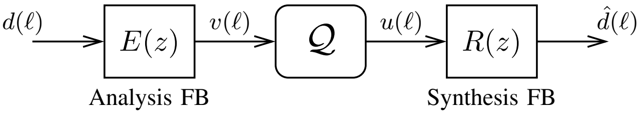


Fig. 2. Oversampled FB in Polyphase Form.

C. Perfect Reconstruction in the Idealized Case

Independent of its description, the oversampled FB of Fig. 1 is a multi-rate system and is thus affected by aliasing and imaging, see e.g. [2], [3]. If the sequences y_i satisfy appropriate bandwidth constraints,[1] then alias-free down-sampling can be ensured. Also, *images* created by the up-sampling process can be filtered out in the synthesis FB.

However, since sharp anti-aliasing filters are of high-order and often exhibit poor characteristics near the cut-off frequencies, it is often convenient to relax requirements on E_i and R_i allowing for some aliasing. In the absence of quantization and coding effects, perfect reconstruction of $y(\ell)$ is possible, as can be appreciated directly from (3) (see also [6]) and aliasing introduced in the analysis FB can be compensated for in the synthesis FB.

Indeed, neglecting quantization, then $u = v$, and hence (3) implies that if[2]

$$R(z)E(z) = I_L, \tag{4}$$

then $d = \hat{d}$ and consequently $y = \hat{y}$. FBs which satisfy (4) are termed *perfect reconstruction* FBs. Note that, choosing $R(z)$ as any left-inverse of $E(z)$, will give a perfect reconstruction FB. In particular, if quantization effects are modeled as additive white Gaussian noise, then choosing $R(z)$ as the pseudo-inverse will minimize the reconstruction mean square error, see also [8].

[1] They need to satisfy a bandpass version of the sampling theorem [15].

[2] I_L denotes the $L \times L$ identity matrix.

D. Quantization Effects

Quantization induces loss of information when transforming the sequence v into u. As a consequence, in general the reconstructed signal $\hat{y}$ will differ from y. As stated in [2], the effect of quantization of the subband signals is very difficult to analyze accurately. The usual paradigm adopted, models the quantizer by a linear gain-plus-noise model, as in the case of standard analog-to-digital conversion, see e.g. [16]. This type of model is simple and may, in some cases, be useful. However, it fails to predict instability related phenomena, such as the appearance of idle tones.

A simple subband coder is described in [2]. It uses one scalar quantizer for each subband, i.e.:[3]

$$u_i(\ell) = q_{\mathbb{U}_i}(v_i(\ell)), \quad \forall \ell,\ \forall i \in \{1, 2, \dots, K\} \qquad (5)$$

Based upon the linear quantization model, the number of bits assigned to each quantizer can be determined according to the energy content of each subband signal in an optimal manner. This type of methodology is widespread, especially in audio applications. To visualize this method, consider a two-channel subband coder, where $E_1(z)$ is a low-pass filter and $E_2(z)$ is high-pass. Given the reduced sensitivity of the human ear to high frequency noise and the fact that typical audio and speech signals have their main energy concentrated at the lower end of the frequency spectrum, it makes sense to assign a higher bit rate to u_1 than to u_2, see also [5].

A more effective approach can be developed by extending the $\Sigma\Delta$-Modulator, see e.g. [18], to the oversampled FB context. In [8], it is shown how the noise shaping coder depicted in Fig. 3 can reduce overall quantization effects in $\hat{y}$. As can be seen in this figure, the scheme consists of a feedback loop which contains a MIMO *noise-shaping filter* G. The quantized set $\mathbb{U}$ has $n_{\mathbb{U}}$ elements. It represents the constraint set for u, i.e. it is given by the Cartesian product:

$$u(\ell) \in \mathbb{U}, \quad \mathbb{U} \triangleq \mathbb{U}_1 \times \mathbb{U}_2 \times \cdots \times \mathbb{U}_K \subset \mathbb{R}^K, \qquad (6)$$

see (1). Thus, $q_{\mathbb{U}}(\cdot)$ is a vector quantizer. The linear quantization model can then be utilized in order to design an *optimal* filter G, which can be either *full MIMO*, or also of restricted complexity, see [8].

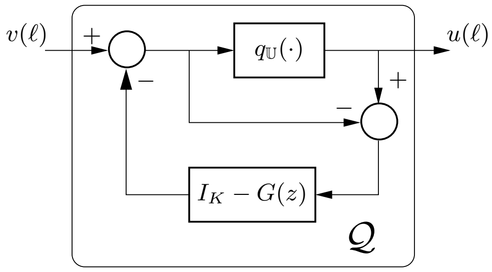


Fig. 3. Noise Shaping Coder.

In the next section we will propose a novel quantization method for oversampled FBs. As will be apparent, the new scheme embeds the noise shaping coder of [8] in a broader framework and allows for enhanced performance. Moreover, stability can be guaranteed for the proposed scheme.

III. Multi-step Optimal Quantization

The main aim of the quantizer is to make the reconstruction error $d - \hat{d}$ small by proper selection of u while aiming to minimize the number of bits per unit time. In what follows, we will show how methods stemming from the Model Predictive Control Framework, see e.g. [9], [19], can be deployed in order to design an abstract quantizer $\mathcal{Q}$, which is aimed at minimizing a weighted measure of the reconstruction error. The method extends our previous work on audio quantization documented in [10]–[12] to oversampled FBs or, equivalently, to non-square MIMO systems.

A. Frequency Selective Coding

The quantization problem can be embedded in the more general problem of minimizing the filtered distortion sequence:

$$e_d(\ell) \triangleq W(z)(H(z)d(\ell) - u(\ell)).$$

In this expression, $H(z)$ is of dimension $K \times L$ and $W(z)$ is of dimension $n_e \times K$. The two filters ($H(z)$, $W(z)$) and n_e are design variables. They allow one to shape the frequency content of the quantization errors. More precisely, if e_d is small, then u will approximate the sequence $H(z)d$ and $W(z)(H(z)d - u)$ will have an approximately flat spectrum.

One possible design choice resides in setting:

$$H(z) = E(z), \quad W(z) = R(z), \qquad (7)$$

in which case $n_e = L$ and $e_d = R(z)v - \hat{d}$, i.e. the difference between the reconstructed sequence $\hat{d}$ and the sequence which would be obtained in the absence of quantization effects. If, furthermore, $E(z)$ and $R(z)$ are chosen to correspond to a perfect reconstruction FB, see (4), then e_d reduces to $e_d = d - \hat{d}$.

Note that, although setting $n_e = L$ appears to be a natural choice, since then the dimension of e_d is equal to that of d and $\hat{d}$, we will show in Section IV-C, that setting $n_e = K$ may be sensible as well.

For our development, it is convenient to describe W via:

$$W(z) = D + C(zI_n - A)^{-1}B,$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times K}$, $C \in \mathbb{R}^{n_e \times n}$, $D \in \mathbb{R}^{n_e \times K}$ and $n \in \mathbb{N}$. With this, e_d is the output of:

$$\begin{aligned} x(\ell + 1) &= Ax(\ell) + B(a(\ell) - u(\ell)) \\ e_d(\ell) &= Cx(\ell) + D(a(\ell) - u(\ell)), \end{aligned} \qquad (8)$$

where $a(\ell) \triangleq H(z)d(\ell) \in \mathbb{R}^{K \times 1}$ and $x \in \mathbb{R}^{n \times 1}$ is the system state.

[3] Here, and in the remainder of this work, $q_X(\cdot)$, where X is any quantized finite set, will denote a standard nearest neighbour quantizer, see e.g. [17].

B. Optimization Criterion

We propose to minimize, at instant $\ell = k$, the following measure of e_d defined over a finite and fixed horizon N:[4]

$$V_N(\vec{u}(k)) \triangleq \|x'(k+N)\|_P^2 + \sum_{\ell=k}^{k+N-1} \|e_d'(\ell)\|^2, \quad (9)$$

where P is a given positive semidefinite matrix and:

$$\vec{u}(k) \triangleq \begin{bmatrix} u'(k)^T & u'(k+1)^T & \dots & u'(k+N-1)^T \end{bmatrix}^T$$

contains the decision variables (candidate quantization levels). The functional (9) examines predictions of the filtered distortion e_d and the *final* state $x(k+N)$ in (8). These predicted trajectories are formed as:

$$\begin{aligned} x'(\ell+1) &= Ax'(\ell) + B(a(\ell) - u'(\ell)), \\ e_d'(\ell) &= Cx'(\ell) + D(a(\ell) - u'(\ell)), \end{aligned}$$

with $\ell = k, k+1, \dots, k+N-1$. The initial condition is $x'(k) = x(k)$, which, given past values of u and d (and hence a), can be computed exactly from (8). The final state weighting term $\|x'(k+N)\|_P^2$ is included in the cost to ensure stability-like properties, see Section IV-A.

C. Moving Horizon Optimization

Minimization of V_N in (9) gives rise to the optimizer

$$\vec{u}^\star(k) \triangleq \arg \min_{\vec{u}(k) \in \mathbb{U}^N} V_N(\vec{u}(k)), \quad (10)$$

where $\mathbb{U}^N \subset \mathbb{R}^{KN}$ is defined via $\mathbb{U}^N \triangleq \mathbb{U} \times \cdots \times \mathbb{U}$.

The vector $\vec{u}^\star(k)$ contains information about decisions to be made at N future time instants. Since its last few components depend only on a small window of the filtered distortion, e_d, we propose to utilize only the *first* K components of $\vec{u}^\star(k)$:[5]

$$u^\star(k) \triangleq \begin{bmatrix} I_K & 0_K & \dots & 0_K \end{bmatrix} \vec{u}^\star(k). \quad (11)$$

It is this quantity, whose effect on e_d is, in principle, best captured within the horizon examined by V_N. At time $\ell = k$ the output of the resultant multi-step optimal quantizer is:

$$u(k) \longleftarrow u^\star(k). \quad (12)$$

This value is also used in (8) in order to deliver $x(k+1)$. At the next sampling instant, this new state value is used to minimize the cost $V_N(\vec{u}(k+1))$, yielding $u(k+1)$. This procedure is repeated *ad-infinitum*. The past is propagated forward in time via the state sequence x, see (8), thus, yielding a recursive scheme.

Fig. 4 depicts the proposed quantizer, which constitutes the main contribution of this work. Of particular interest is the case where H is chosen to be equal to the analysis polyphase matrix E, see also (7). With this choice, the input to the quantizer can be regarded as $a = v$.

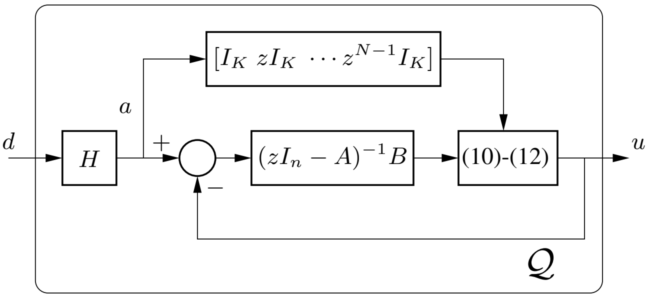


Fig. 4. Implementation of the quantizer as a feedback loop.

It is worth emphasizing that, in general, larger values for the prediction horizon N provide better performance, since more data is taken into account in the quantization process. In fact, one can expect that, if N is chosen large enough relative to the time scale of W, then the effect of $u(k)$ on $e_d(\ell)$ for $\ell \geq k+N$ will be negligible. Since, in general, the computational time needed to obtain (10) is exponential in N, this parameter allows the designer to trade-off performance versus on-line computational effort. As will be shown in Section V, excellent performance can often be achieved with relatively small horizons.

The quantizer proposed can be regarded as a (non-square) MIMO Model Predictive Controller, see e.g. [19], [20], with *plant* W, *controlled input* u and *reference* $W(z)H(z)d$. The special feature of this particular problem, is that u is restricted to satisfy the quantization constraint (6), which puts the problem into a finite set constrained predictive control framework, see e.g. [9], [21], [22].

IV. Properties of the Quantizer

The quantization scheme proposed is not only related to finite-set constrained predictive control schemes, but also generalizes our previous work on audio quantization, see [10], [11], and especially the *Multi-step Optimal Converter* (MSOC) described in [12]. Indeed, as will be apparent from what follows, some features of the quantizer are closely related to those of the MSOC.

A. Stability

A very significant, but largely unsolved, issue which arises when including a quantizer in a feedback loop is that of stability. Poor stability properties manifest themselves in the appearance of idle tones (limit cycles), see also [23]–[25]. By applying ideas stemming from the Model Predictive Control literature, see e.g. [9], [19], [20] we can obtain the following stability-related result:

Theorem 1: Suppose that the sequence a is such that for a finite value k, it holds that $a(\ell) \in \mathbb{U}, \forall \ell \geq k$. Then, provided P is chosen to satisfy the Lyapunov Equation

$$A^T PA + C^T C = P, \quad (13)$$

we have $e_d(\ell) \to 0 \in \mathbb{R}^{n_e \times 1}$, as $\ell \to \infty$.

Proof: The proof is included in Appendix A. ■

Despite the fact that this result is valid only for a restricted class of inputs, namely those that make the sequence

[4] $\|x\|_P^2$ denotes the quadratic form $x^T Px$, where x is any vector and P is a matrix, $\|e_d\|^2 = e_d^T e_d$.

[5] $0_K \triangleq 0 \cdot I_K$.

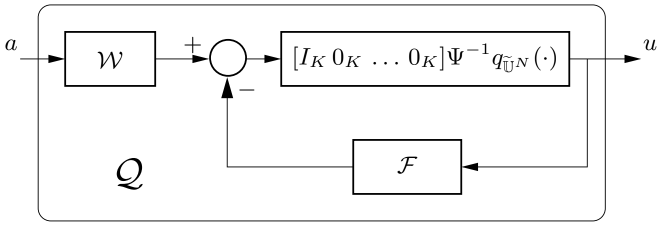
a eventually take on values in $\mathbb{U}$, simulation studies, such as those included in Section V (see also [12]), indicate that setting P as the *stabilizing* value given in (13) is often a good choice to avoid limit cycles and performance degradation.

B. Closed Form Solution

The proposed quantizer requires that one solve the (non-convex) finite-set-constrained optimization problem (10). In the particular case of having a square filter W, i.e. of choosing $n_e = K$, see Section III-A, we can extend the results of [12] to state the solution in terms of a vector quantizer immersed in a closed loop as depicted in Fig. 5. In this figure, $\mathcal{F}(z)$ and $\mathcal{W}(z)$ are the $NK \times K$ filters:

$$\mathcal{F}(z) \triangleq \Psi^{-T}(\Phi^T\Gamma + M^T P A^N)(zI_n - A)^{-1}B,$$
$$\mathcal{W}(z) \triangleq \Psi \begin{bmatrix} I_K & zI_K & \dots & z^{N-1}I_K \end{bmatrix}^T + \mathcal{F}(z)$$

with the matrices Ψ and Φ defined in Appendix B.



a
$\mathcal{W}$
+
−
$[I_K\, 0_K\, \dots\, 0_K]\Psi^{-1}q_{\widetilde{\mathbb{U}}^N}(\cdot)$
u
$\mathcal{F}$
$\mathcal{Q}$

Fig. 5. Implementation of the quantizer as a feedback loop ($n_e = K$).

It should be emphasized that, if n_e is chosen smaller than K (as is the case of $n_e = L$ with an oversampled FB, see (7)), then the matrices $\Phi^T\Phi$ and Ψ in (18) will be rank deficient. Thus, with $n_e < K$, Ψ is not invertible, and the scheme depicted in Fig. 5 does not make sense.

C. Relationship to the Noise-Shaping Quantizer

Examination of Figs. 3 and 5 shows that the multi-step optimal quantizer and the noise-shaping coder proposed in [8] are related. Indeed, consider the special (and simple) case of a horizon $N = 1$, no terminal state weighting, $P = 0$, and filter $W(z)$ with $n_e = K$ and $D = I_K$.

In this case, the definitions included in Appendix B yield that $\Gamma = C$, $\Psi = \Phi = D = I_K$, so that $\widetilde{\mathbb{U}}^N = \mathbb{U}$, $\mathcal{W}(z) = W(z)$ and $\mathcal{F}(z) = W(z) - I_K$. Thus, the dynamics of Fig. 5 are governed by:

$$u(\ell) = q_{\mathbb{U}}(W(z)H(z)d(\ell) - (W(z) - I_K)u(\ell)).$$

On the other hand, the noise-shaping coder of Fig. 3 is characterized via:

$$u(\ell) = q_{\mathbb{U}}(G^{-1}(z)v(\ell) - (G^{-1}(z) - I_K)u(\ell)).$$

As a consequence, both schemes are equivalent if:

$$W(z) = G^{-1}(z), \quad H(z) = E(z). \tag{14}$$

Therefore, the multi-step optimal quantizer can be regarded as a generalization of the noise-shaping coder.

We thus see that the proposed quantizer extends the noise-shaping coder in [8] in two ways: It allows for multi-step optimality (horizon $N > 1$) and stability concepts can be directly included in the design via the terminal state penalization matrix P. As documented in the example included in the following section, these design parameters can be used to give enhanced performance.

V. Simulation Study

As a simple example, adopted from [8], we consider a two-channel FB, with $L = 1$, which gives an oversampling factor of 2. The analysis bank is formed by the Haar filters:

$$E_1(z) = (1 + z^{-1})/\sqrt{2}, \quad E_2(z) = (1 - z^{-1})/\sqrt{2},$$

and the synthesis bank is characterized by:

$$R_1(z) = \frac{1}{2\sqrt{2}}(1 + z), \quad R_2(z) = \frac{1}{2\sqrt{2}}(1 - z).$$

This choice gives a polyphase description which satisfies the *perfect reconstruction* condition (4). We utilize the noise-shaping filter of [8], namely:

$$G(z) = I_2 - z^{-1} \begin{bmatrix} 0.5 & 0.5 \\ -0.5 & -0.5 \end{bmatrix}$$

and tune $W(z)$ and $H(z)$ according to (14) and P as in (13).

To visualize performance of the proposed quantizer, we carried out a simulation, with $\mathbb{U}_1 = \mathbb{U}_2 = \{-1, 0, 1\}$, and where y was chosen as $T_f = 1000$ samples of an i.i.d. Gaussian process having zero-mean and unit variance.

Fig. 6 illustrates the results. It shows the values of the sample variance of the reconstruction error:

$$S \triangleq \frac{1}{T_f}\sum_{\ell=1}^{T_f}(y(\ell) - \hat{y}(\ell))^2, \tag{15}$$

for various horizons N. In this figure, $N = 0$ denotes direct quantization as in (5).

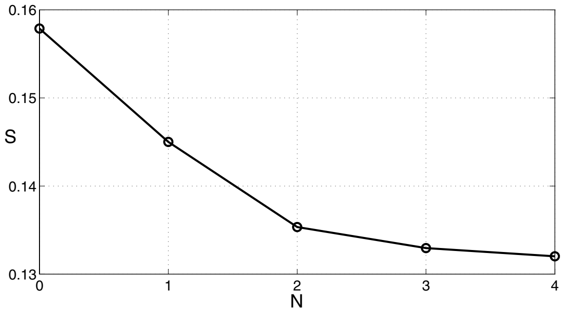


Fig. 6. Performance as a function of N.

As can be seen from this figure, performance improves as the horizon N is increased. Thus, the proposed quantizer outperforms direct quantization and also the noise shaping coder (which correspond to $N = 0$ and $N = 1$, respectively). Moreover, it can be observed that performance is asymptotic in N and that a modest value, say $N = 2$, achieves most of the performance gain. These observations are typical, and also agree with those obtained in the direct signal quantization context, see [10]–[12].

VI. Conclusions

We have proposed a novel quantization method for oversampled filter banks. The scheme utilizes the receding horizon optimization concept and can be implemented as a closed loop. The quantizer presented here includes the noise shaping coder as a special case. More precisely, while the noise-shaping coder is *one-step optimal*, the new quantizer is multi-step optimal. This, and the possibility of a stability-enforced design, gives enhanced performance, as documented by a simulation study.

Appendix

A. Proof of Theorem 1

The proof uses the sequence of optimal costs defined as

$$V_N^\star(\ell) \triangleq V_N(\vec{u}^\star(\ell)), \quad \ell \in \mathbb{N}$$

as a Lyapunov function and follows closely that of Theorem 2 in [12].

Suppose that, at sample $\ell = k$, the optimal sequence is

$$\vec{u}^\star(k) = \begin{bmatrix} u_k^T & u_{k+1}^T & \dots & u_{k+N-1}^T \end{bmatrix}^T.$$

Next, at sample $\ell = k+1$, consider the related sequence

$$u^s = \begin{bmatrix} u_{k+1}^T & u_{k+2}^T & \dots & u_{k+N-1}^T & a(k+N)^T \end{bmatrix}^T \in \mathbb{U}^N. \tag{16}$$

Due to optimality, it follows that

$$V_N^\star(k+1) \leq V_N(u^s).$$

Direct calculation then yields:

$$\begin{aligned} V_N(u^s) &= V_N^\star(k) + \|Ax(k+N)\|_P^2 - \|x(k+N)\|_P^2 \\ &+ \|e_d(k+N)\|^2 - \|e_d(k)\|^2 = V_N^\star(k) \\ &+\|Ax(k+N)\|_P^2 - \|x(k+N)\|_P^2 + \|Cx(k+N)\|^2 - \|e_d(k)\|^2 \end{aligned}$$

since, with u^s given by (16), predictions of e_d used in $V_N^\star(k)$ and in $V_N(u^s)$ coincide. Hence,

$$V_N^\star(k+1) - V_N^\star(k) \leq -\|e_d(k)\|^2 \leq 0, \tag{17}$$

where we have used (13). As a consequence of (17), it follows that $\lim_{\ell \to \infty} V_N^\star(\ell)$ exists and

$$V_N^\star(\ell+1) - V_N^\star(\ell) \to 0.$$

Also from (17) it follows that $e_d(\ell) \to 0$, which completes the proof.

B. Definitions for Fig. 5

In Fig. 5 the matrix Ψ has dimensions $NK \times NK$ and is defined implicitly via:

$$\Psi^T \Psi = \Phi^T \Phi + M^T P M, \tag{18}$$

where:

$$\Phi \triangleq \begin{bmatrix} D & 0 & \dots & 0 \\ CB & D & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ CA^{N-2}B & \dots & CB & D \end{bmatrix}, \; \Gamma \triangleq \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{N-1} \end{bmatrix},$$

$$M \triangleq \begin{bmatrix} A^{N-1}B & A^{N-2}B & \dots & AB & B \end{bmatrix}.$$

The image of the quantizer $q_{\widetilde{\mathbb{U}}^N}(\cdot)$ is the set:

$$\widetilde{\mathbb{U}}^N \triangleq \{\tilde{\eta}_1, \, \tilde{\eta}_2, \dots, \tilde{\eta}_r\} \subset \mathbb{R}^{NK},$$

with

$$\tilde{\eta}_i = \Psi \eta_i, \; \eta_i \in \mathbb{U}^N.$$

References

[1] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[2] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, New Jersey: Prentice-Hall, 1993.

[3] N. Fliege, *Multirate Digital Signal Processing: Multirate Systems, Filter Banks, Wavelets*. New York, NY: John Wiley & Sons, 1994.

[4] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press, 1996.

[5] N. Gilchrist and C. Grewin, eds., *Collected Papers on Digital Audio Bit-Rate Reduction*. New York: Audio Eng. Soc, 1996.

[6] Z. Cvetković and J. D. Johnson, "Nonuniform oversampled filter banks for audio signal processing," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 393–399, Sept. 2003.

[7] A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets*. San Diego, CA: Academic Press, 1992.

[8] H. Bölcskei and F. Hlawatsch, "Noise reduction in oversampled filter banks using predictive quantization," *IEEE Trans. Inform. Theory*, vol. 47, pp. 155–172, Jan. 2001.

[9] D. E. Quevedo, G. C. Goodwin, and J. A. De Doná, "Finite constraint set receding horizon control," *Int. J. Robust Nonlin. Contr.*, vol. 14, pp. 355–377, Mar. 2004.

[10] G. C. Goodwin, D. E. Quevedo, and D. McGrath, "Moving-horizon optimal quantizer for audio signals," *J. Audio Eng. Soc.*, vol. 51, pp. 138–149, Mar. 2003.

[11] D. E. Quevedo and G. C. Goodwin, "Audio quantization from a receding horizon control perspective," in *Proc. Amer. Contr. Conf.*, pp. 4131–4136, 2003.

[12] D. E. Quevedo and G. C. Goodwin, "Multi-step optimal analog-to-digital conversion," *IEEE Trans. Circuits Syst. I*, 2004, to appear.

[13] Z. Cvetković and M. Vetterli, "Oversampled filter banks," *IEEE Trans. Signal Processing*, vol. 46, pp. 1245–1255, May 1998.

[14] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Processing*, vol. 46, pp. 3256–3268, Dec. 1998.

[15] V. Sathe and P. P. Vaidyanathan, "Effects of multirate systems on the statistical properties of random signals," *IEEE Trans. Signal Processing*, vol. 41, pp. 131–146, Jan. 1993.

[16] S. P. Lipschitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, pp. 355–375, May 1992.

[17] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic, 1992.

[18] S. R. Norsworthy, R. Schreier, and G. C. Temes, eds., *Delta–Sigma Data Converters: Theory, Design and Simulation*. Piscataway, N.J.: IEEE Press, 1997.

[19] J. M. Maciejowski, *Predictive Control with Constraints*. Prentice-Hall, 2002.

[20] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Optimality and stability," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[21] G. C. Goodwin and D. E. Quevedo, "Finite alphabet control and estimation," *Int. J. Contr., Automation, and Syst.*, vol. 1, pp. 412–430, Dec. 2003.

[22] A. Bemporad, "Multiparametric nonlinear integer programming and explicit quantized optimal control," in *Proc. IEEE Conf. Decis. Contr.*, pp. 3167–3172, 2003.

[23] O. Feely, "A tutorial introduction to non-linear dynamics and chaos and their application to Sigma-Delta modulators," *Int. J. Circuit Theory Appl.*, vol. 25, pp. 347–367, 1997.

[24] D. F. Delchamps, "Nonlinear dynamics of oversampling A-to-D converters," in *Proc. IEEE Conf. Decis. Contr.*, pp. 480–485, 1993.

[25] P. J. Ramadge, "On the periodicity of symbolic observations of piecewise smooth discrete-time systems," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 807–813, July 1990.