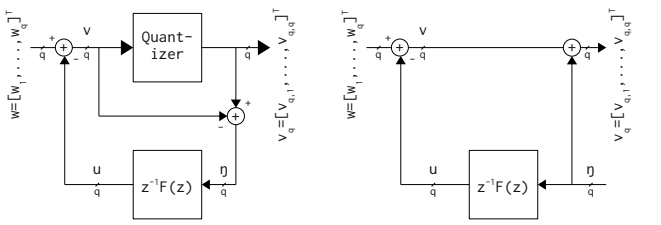
49th IEEE Conference on Decision and Control
December 15-17, 2010
Hilton Atlanta Hotel, Atlanta, GA, USA

Design of Optimal Feedback Filters with Guaranteed Closed-Loop Stability for Oversampled Noise-Shaping Subband Quantizers

Sander Wahls Holger Boche.

***Abstract*—We consider the oversampled noise-shaping subband quantizer (ONSQ) with the quantization noise being modeled as white noise. The problem at hand is to design an optimal feedback filter. Optimal feedback filters with regard to the current standard model of the ONSQ inhabit several shortcomings: the stability of the feedback loop is not ensured and the noise is considered independent of both the input and the feedback filter. Another previously unknown disadvantage, which we prove in our paper, is that optimal feedback filters in the standard model are never unique. The goal of this paper is to show how these shortcomings can be overcome. The stability issue is addressed by showing that every feedback filter can be "stabilized" in the sense that we can find another feedback filter which stabilizes the feedback loop and achieves a performance equal (or arbitrarily close) to the original filter. However, the other issues are inherent properties of the standard model. Therefore, we also consider a so-called extended model of the ONSQ, which includes the influence of the feedback filter on the noise power and indirectly also the influence of the inputs statistical properties. We show that the optimal feedback filter for this extended model is unique and automatically achieves a stable feedback loop. We also give an algorithm to compute it.**



(a) Real model: η is the qnt. error (b) Linearized model: η is white noise

Figure 1: Noise-Shaping Quantizer

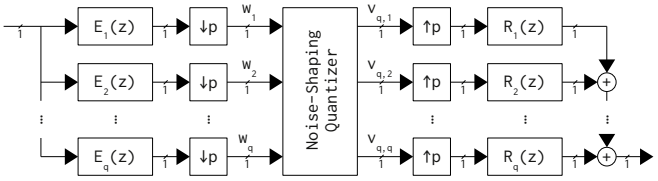


Figure 2: Oversampled Noise-Shaping Subband Quantizer

I. Introduction

The concept of an oversampled noise-shaping subband quantizer (ONSQ) was introduced by Bölcskei and Hlawatsch in [1]. The system model of a ONSQ is depicted in Fig. 2, and can be described as follows. The input signal of the ONSQ, which is modeled by a sequence of complex numbers, is first passed though the analysis filter bank $E_1(z), \ldots, E_q(z)$. The q outputs of the analysis filter bank are downsampled[1] by a factor p, then jointly quantized by a noise-shaping quantizer, and finally upsampled[2] again by the factor p. The output of the ONSQ is obtained after the results of these operations are passed through the synthesis filter bank $R_1(z), \ldots, R_q(z)$. The "oversampled" in the term ONSQ refers to the assumption that the size q of the filter banks is larger than the sampling factor p. The "perfect reconstruction" means that the output of the ONSQ should equal the input if the quantization error is zero (probably subject to a finite delay). It remains to have a closer look at the noise shaping quantizer in Fig. 2. The system model of a noise shaping quantizer is depicted in Fig. 1a. The input signal, which we describe with a vector-valued sequence, is first reduced by a feedback signal. The output is obtained when this difference is quantized by some memoryless quantizer with finitely many quantization levels. The aforementioned feedback signal is generated by passing the difference between the quantizers output and its input, i.e., the quantization error, into the feedback filter $z^{-1}F(z)$.

The design problem treated in this paper consists in designing the feedback filter $F(z)$ of the noise-shaping quantizer such that the quantization noise is maximally damped by the synthesis filter bank $R_1(z), \ldots, R_p(z)$. The first approach to this problem was given in [1]. There, the model of the noise-shaping quantizer was linearized by the assumption that the quantization noise η is white with constant variance (cf. Fig. 1b). We call this the *standard model*. The optimal finite impulse response (FIR) feedback filter for the standard model was derived in [1]. The standard model has three well-known disadvantages (also cf. [2]), viz.:

1) stability of the feedback loop is not taken into account,
2) the noise is considered independent of the input, and
3) the noise is considered independent of the feedback.

In order to address these problems, Quevedo et. al. proposed the so-called moving horizon subband quantizer (MHSQ) which uses the quantized symbol that minimizes the prediction of the quantization error over a finite time horizon [2]. While this approach often performs very good, there is the disadvantage of exponential complexity in both the prediction horizon and the number of quantizer outputs. Clearly, a lower computational burden is desirable in practical scenarios.

Both authors are with Technische Universität Berlin, Fachgebiet Informationstheorie und theoretische Informationstechnik, Einsteinufer 25, 10587 Berlin, Germany. Email: {sander.wahls,holger.boche}@mk.tu-berlin.de

This work has been supported by the German Research Foundation (DFG) under grant BO 1734/5-2.

[1]Downsampling by p means $\{a_k\}_{k=0}^{\infty} \mapsto \{a_{kp}\}_{k=0}^{\infty}$.

[2]Upsampling by p means $\{a_k\}_{k=0}^{\infty} \mapsto \{b_k\}_{k=0}^{\infty}$ where $b_k = a_{k/p}$ if k is a integer multiple of p and $b_k = 0$ otherwise.

Therefore, we revisit the white quantization noise assumption in this paper and show how the disadvantages given before can be addressed. First, we extend the results on the standard model from [1] (i.e., the optimal FIR feedback filter which ignores closed-loop stability) to the infinite impulse response (IIR) case. Second, we show how an optimal IIR feedback filter for the standard model can be converted into a filter which ensures stability of the feedback loop and at the same time achieves a performance equal or at least arbitrarily close to the initial filter (which did not necessarily achieve closed-loop stability). Thus, we show how the first disadvantage of the standard model can be removed *without sacrificing any performance*. However, the other two disadvantages are still around. Even worse, we show that computation of optimal feedback filters for the standard model is an ill-posed problem in the sense of Hadamard because its solutions are *never* unique. In order to address these issues, we pick up an idea proposed by Derpich et. al. in the context of scalar quantization [3]. They assumed a fixed signal-to-noise ratio (SNR) in the quantizer, and used this assumption to get an approximation of the quantization noise variance which incorporates the feedback filter. Obviously, the SNR of the quantizer depends on the statistics of the inputs. Thus, this *extended model* addresses the other two disadvantages of the standard model. We show that the optimal IIR filter regarding the extended model (ignoring closed-loop stability) *always exists and is unique*, and give a simple algorithm to compute it. Moreover, we prove that this filter *automatically achieves closed-loop stability* as soon as an optimal filter with closed-loop stability exists at all. For the rare cases where this is not the case, we derive suboptimal feedback filters with performance arbitrarily close-to-optimal and closed-loop stability. Thus, closed-loop stability can again be achieved *without sacrificing any performance*.

Notation: We denote the space of stable and causal rational $m \times l$ matrices by $\mathcal{RH}_\infty^{m\times l}$, i.e., all poles are contained in the open unit disc $|z| < 1$. The para-hermitian of any $A \in \mathcal{RH}_\infty^{m\times l}$ is defined as $A^\sim(z) := A(\bar{z}^{-1})^*$, the infinity norm as $\|A\|_\infty := \sup_{|z|=1} \sigma_{\max}[A(z)]$ where $\sigma_{\max}[\cdot]$ denotes the largest singular value. The Frobenius norm of a complex matrix $B \in \mathbb{C}^{m\times l}$ is defined via $\|B\|_F^2 := \operatorname{trace}\{BB^*\}$. Finally, we introduce the two norm of a rational matrix C, $\|C\|_2^2 := \int_{\theta=0}^{2\pi} \|C(e^{i\theta})\|_F^2 \frac{d\theta}{2\pi}$. The normal rank of C is given by $\max_{z\in\mathbb{C}} \operatorname{rank}[C(z)]$.

II. Problem Statement

The system model of an ONSQ has already been discussed in the introduction. However, direct use of the system model is intricate during computations. Hence, we use the equivalent polyphase representation of the ONSQ which is depicted in Fig. 3 (please see [1] for details). We assume that the polyphase matrices $E \in \mathcal{RH}_\infty^{q\times p}$ and $R \in \mathcal{RH}_\infty^{p\times q}$ of the synthesis and analysis filter banks are given and achieve perfect reconstruction, i.e., $R(z)E(z) = z^{-L}I$ for some decision delay $L > 0$. We want to design an "optimal" feedback filter $F \in \mathcal{RH}_\infty^{q\times q}$.

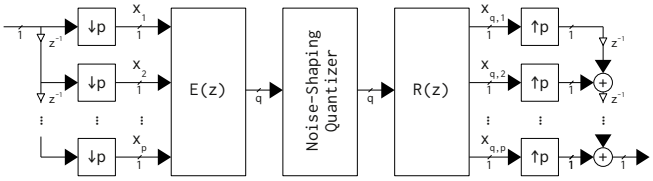


Figure 3: Polyphase representation of the ONSQ

In order to get a tractable formulation of "optimal" we model the quantization noise as temporally and spatially white noise with variance $\sigma_\eta^2 > 0$ (cf. Fig. 1b). The input-output relation in Figure 3 becomes

$$x_q = R[Ex + (I - z^{-1}F)\eta] = z^{-L}x + R(I - z^{-1}F)\eta.$$

Hence, we aim to minimize the variance of the expected reconstruction error $\sigma_e^2 := \sigma_\eta^2 \|R(I - z^{-1}F)\|_2^2$ of the linearized model. Two different approaches to the minimization of σ_e^2 can be found in the literature.

1) In [1], the variance σ_η^2 of the quantization error is considered independent of the feedback filter F. Then, one has to minimize $\|R(I - z^{-1}F)\|_2^2$. We call this the *standard model*.
2) In [3], the variance σ_η^2 of the quantization error is approximated by $\|E\|_2^2/(\gamma - \|F\|_2^2)$, where γ is the signal to noise ratio (SNR) of the quantizer. Then, minimization of σ_e^2 is equivalent to minimization of $\|R(I - z^{-1}F)\|_2^2/(\gamma - \|F\|_2^2)$ subject to $\|F\|_2^2 < \gamma$. We call this the *extended model*.

Finally, let us point out an important difference between the real model and the linearized model of the noise-shaping quantizer: In contrast to the real model (Fig. 1a) there is no feedback in the linearized model (Fig. 1b). Thus, we should explicitly take stability of the feedback loop into account when we design our filters. Following Derpich et. al. [3, §II-A3], we say that the quantizer is stable if a bounded input implies that all internal signals are bounded. Since the quantized signal v_q in Figure 1a is always bounded and $v = (I - z^{-1}F)^{-1}(w - z^{-1}Fv_q)$, $u = z^{-1}F(v_q - v)$, this means that we should additionally require $(I - z^{-1}F)^{-1} \in \mathcal{RH}_\infty^{q\times q}$.

III. Preliminaries

Let us shortly recall some basics from linear systems theory that will be used repeatedly in this paper.

A. Outer-Coinner Factorization

Let $0 \neq M \in \mathcal{RH}_\infty^{m\times l}$ have no unit circle zeros. Then, we can introduce the so-called outer-coinner factorization $M = M_o M_i$, where $M_o \in \mathcal{RH}_\infty^{m\times k}$, $M_i \in \mathcal{RH}_\infty^{k\times l}$, $M_o^+ M_o = I$ for some $M_o^+ \in \mathcal{RH}_\infty^{k\times m}$ and $M_i M_i^\sim = I$ [4]. Note that M_o is square if and only if M has normal rank m. Finally, let us repeat the well-known observation that multiplication from the right with a coinner function preserves the two norm.

Lemma 1. *Suppose $A \in \mathcal{RH}_\infty^{m\times l}$ and $B \in \mathcal{RH}_\infty^{l\times k}$ with B coinner, i.e., $BB^\sim = I$. Then, $\|AB\|_2^2 = \|A\|_2^2$.*

Proof: This follows immediately from $B^\sim(e^{i\theta}) = B(e^{i\theta})^*$ and the definitions of $\|\cdot\|_2$ and $\|\cdot\|_F$. ∎

B. Model-Matching

The (one-sided, inexact) model-matching problem, which often also is considered a $\mathcal{H}_2$ filtering problem, has been treated extensively in the literature (see, e.g., [5]). The next lemma provides solvability conditions.

Lemma 2. *Suppose $A \in \mathcal{RH}_\infty^{m\times l_1}$ and $B \in \mathcal{RH}_\infty^{m\times l_2}$, where B has no unit circle zeros. Then, there exists $K \in \mathcal{RH}_\infty^{l_2\times l_1}$ such that $\|A+BK\|_2^2 = \inf_{\tilde{K}\in\mathcal{RH}_\infty^{l_2\times l_1}} \|A+B\tilde{K}\|_2^2$. This K is unique if and only if B has normal rank $l_2 > 0$.*

Proof: See, e.g., Cor. 10.51 and Th. 10.54 in [5]. ■

IV. Optimal Feedback Filter: Standard Model

In this section, we derive optimal IIR feedback filters for the standard model. These can be seen as the limiting case of the FIR filters given in [1]. We also show how instable feedback loops, which can occur in the standard model, can be removed without loosing performance, and analyze the solution sets.

A. Unstable Feedback Loop

We start with with the optimal feedback filter with respect to the standard model, where we do not care for stability of the feedback loop.

Problem 3. Find $F \in \mathcal{RH}_\infty^{q\times q}$ such that $\|R(I-z^{-1}F)\|_2^2$ is minimized.

The FIR version of this problem, i.e., $F(z) = F_0 + F_1z^{-1} + \cdots + F_Nz^{-N}$, has been solved in [1]. We point out that the general case considered here is a simple model-matching problem (cf. Section III-B). However, the simplicity comes at a price. The next proposition shows that Problem 3 is an ill-posed problem (in the sense of Hadamard).

Proposition 4. *Problem 3 has infinitely many solutions.*

Proof: Lemma 2 implies existence of a solution because the perfect reconstruction property $R(z)E(z) = z^{-L}I$ and $R \in \mathcal{RH}_\infty^{q\times p}$ ensure that R has no unit circle zeros. Now, suppose F_{opt} is any solution to Problem 3. Let $E_\perp \in \mathcal{RH}_\infty^{q\times(q-p)}$ denote a basis of the kernel of R, i.e., $E_\perp$ has full column normal rank, and $RE_\perp = 0$. Then, for any $Q \in \mathcal{RH}_\infty^{(q-p)\times q}$, also $F_{opt}+E_\perp Q$ is a solution to Problem 3 because $\|R(I-z^{-1}(F_{opt}+E_\perp Q))\|_2^2 = \|R(I-z^{-1}F_{opt}) - z^{-1}RE_\perp Q\|_2^2$ and $RE_\perp = 0$. Since $E_\perp$ has full column normal rank, this amounts to infinitely many solutions. ■

B. Stable Feedback Loop

Next, we consider the optimal feedback filter with respect to the standard model that stabilizes the feedback loop. The problem can be formulated as follows.

Problem 5. Find $F \in \mathcal{RH}_\infty^{q\times q}$ such that $(I-z^{-1}F)^{-1} \in \mathcal{RH}_\infty^{q\times q}$ and $\|R(I-z^{-1}F)\|_2^2$ is minimized.

Solution of Problem 5 is more complicated than solution of Problem 3 because additionally the stability of the feedback loop has to be taken into account. To the best of our knowledge, no solution technique for Problem 5 is known in the literature. Therefore, we propose the following new approach, where we modify a solution to Problem 3 (which is simple to obtain) in order to compute an (probably approximate) solution to the more complicated Problem 5. We point out that Gerzon and Craven already observed in the scalar case that minimum-phaseness of the feedback filter can be enforced without loss of performance [6, p. 9]. They proposed to recursively reflect the non-minimum-phase zeros of $(I - z^{-1}F)$ with respect to the unit circle. Note that this is simple in the scalar case, but not in the matrix case considered here. Therefore, we do not try to reflect the zeros, but instead dislocate them using an inner factor.

Theorem 6. *Let $F_\star$ denote a solution to Problem 3. Then, the following holds.*

1) *Suppose $I - z^{-1}F_\star$ has no unit circle zeros, and introduce an outer-coinner factorization $I - z^{-1}F_\star = F_{\star o}F_{\star i}$. Then, $F_{opt} := I - F_{\star o}F_{\star o}^{-1}(\infty)$ is strictly proper, and $zF_{opt} \in \mathcal{RH}_\infty^{q\times q}$ solves Problem 5.*
2) *Suppose $I - z^{-1}F_\star$ has unit circle zeros, and introduce a family of outer-coinner factorizations $[I - z^{-1}F_\star, \epsilon I] = F_{\star o}^{(\epsilon)}F_{\star i}^{(\epsilon)}$ and a family of feedback filters $F_{opt}^{(\epsilon)}(z) := I - F_{\star o}^{(\epsilon)}(z)[F_{\star o}^{(\epsilon)}(\infty)]^{-1}$. Then, $zF_{opt}^{(\epsilon)}, (I - z^{-1}zF_{opt}^{(\epsilon)})^{-1} \in \mathcal{RH}_\infty^{q\times q}$ for all $\epsilon > 0$, and*

$$\lim_{\epsilon\searrow 0} \|R(I - z^{-1}zF_{opt}^{(\epsilon)})\|_2^2 = \|R(I - z^{-1}F_\star)\|_2^2.$$

Proof: "1)": The factorization is well-defined because $I - z^{-1}F_\star$ has no unit circle zeros (and full normal rank). By construction, we have $F_{opt} \in \mathcal{RH}_\infty^{q\times q}$. F_{opt} is strictly proper because $F_{opt}(\infty) = I - F_{\star o}(\infty)F_{\star o}^{-1}(\infty) = 0$. Since the matrix $F_{\star o}^{-1}(\infty)$ is invertible, we further see that $(I - z^{-1}zF_{opt})^{-1} = (I - (I - F_{\star o}F_{\star o}^{-1}(\infty)))^{-1} = (F_{\star o}F_{\star o}^{-1}(\infty))^{-1} = F_{\star o}(\infty)F_{\star o}^{-1} \overset{(F_{\star o}\text{ outer})}{\in} \mathcal{RH}_\infty^{q\times q}$. Finally, we also have $\|R(I - z^{-1}zF_{opt})\|_2^2 \overset{\text{(Def. } F_{opt})}{=} \|RF_{\star o}F_{\star o}^{-1}(\infty)\|_2^2 \overset{\text{(Lem. 18+19/Appdx.)}}{=} \int_{\theta=0}^{2\pi} \|R(e^{i\theta})F_{\star o}(e^{i\theta})F_{\star o}^{-1}(\infty)\|_F^2 \frac{d\theta}{2\pi} \overset{\text{(Lem. 1)}}{\leq} \int_{\theta=0}^{2\pi} \|R(e^{i\theta})F_{\star o}(e^{i\theta})\|_F^2 \frac{d\theta}{2\pi} = \|RF_{\star o}\|_2^2 = \|RF_{\star o}F_{\star i}\|_2^2 = \|R(I - z^{-1}F_\star)\|_2^2$. Since the optimal performance in Problem 5 is lower bounded by the optimal performance in Problem 3, i.e., $\|R(I - z^{-1}F_\star)\|_2^2$, and zF_{opt} achieves this lower bound, we see that zF_{opt} solves Problem 5. We skip the second part of the proof because of space limitations. ■

Finally, we point out that although the requirements in Problem 5 are stronger than in Problem 3 because we additionally require stability of the feedback loop, there still can be no unique solution. The problem is still ill-conditioned.

Proposition 7. *Problem 5 has either no or infinitely many solutions.*

Proof: Suppose F_{opt} is a solution to Problem 5. Then, there exists some $\delta > 0$ such that $(I - z^{-1}F_{opt} + X)^{-1} \in \mathcal{RH}_\infty^{q\times q}$ for all $X \in \mathcal{RH}_\infty^{q\times q}$ with $\|X\|_\infty < \delta$, because the invertible stable proper rational matrices form an open set in the topology induced by the infinity norm. (This follows from

the fact that the invertible elements in any Banach algebra with unit form an open set, where we choose $\mathcal{RH}_\infty^{q\times q}$ as the algebra.) Now, consider $X = E_\perp Q \in \mathcal{RH}_\infty^{q\times q}$ for any $Q \in \mathcal{RH}_\infty^{(q-p)\times q}$ with $\|Q\|_\infty < \delta \|E_\perp\|_\infty^{-1}$, where again $E_\perp \in \mathcal{RH}_\infty^{q\times(q-p)}$ denotes a basis of the kernel of R. (cf. the proof of Proposition 4). Then, $\|X\|_\infty \le \delta \|E_\perp\|_\infty^{-1}\|E_\perp\|_\infty = \delta$, and thus $(I - z^{-1}(F_{opt}+X))^{-1} \in \mathcal{RH}_\infty^{q\times q}$. We also have $\|R(I - z^{-1}(F_{opt}+X))\|_2^2 = \|R(I - z^{-1}F_{opt})\|_2^2$ because $RE_\perp = 0$. Hence, $F_{opt}+X$ is another solution to Problem 5. Since $E_\perp$ has full column normal rank, we can obtain infinitely many optimal solutions in this way. ∎

V. Optimal Feedback Filter: Extended Model

The previous section has shown that the standard noise model leads to ill-posed optimization problems with non-unique solutions. In this section, we consider an extended model proposed by Derpich et. al. in the context of non-oversampled noise shaping subband quantization with only one analysis/synthesis filter [3]. We derive the optimal feedback filter for our system model with respect to the extended model. We also show that the extended model does not suffer from non-unique solutions, or unstable feedback loops.

A. Unstable Feedback Loop

Like for the standard model, we initially do not require a stable feedback loop.

Problem 8. Find $F \in \mathcal{RH}_\infty^{q\times q}$ such that $\|F\|_2^2 < \gamma$, and $\|R(I - z^{-1}F)\|_2^2/(\gamma - \|F\|_2^2)$ is minimized.

In contrast to Problem 3, this is no longer a simple model matching problem. However, the next lemma will allow us to establish a relation to a family of model matching problems.

Lemma 9. *Let $F \in \mathcal{RH}_\infty^{q\times q}$, $C > 0$. Then,*

$$\frac{\|R(I - z^{-1}F)\|_2^2}{\gamma - \|F\|_2^2} = C \text{ and } \|F\|_2^2 < \gamma \tag{1}$$

$$\Leftrightarrow \left\| \begin{bmatrix} C^{-1}R \\ 0 \end{bmatrix} - \begin{bmatrix} z^{-1}C^{-1}R \\ I \end{bmatrix} F \right\|_2^2 = \gamma. \tag{2}$$

Proof: First, note that (1) implies $C^{-1}\|R(I - z^{-1}F)\|_2^2 + \|F\|_2^2 = \gamma$, which in turn is equivalent to (2). Thus, the "$\Rightarrow$" part of the Lemma is shown. For the "$\Leftarrow$" part it remains to show that $\|F\|_2^2 < \gamma$, but this follows immediately from $C^{-1}\|R(I - z^{-1}F)\|_2^2 + \|F\|_2^2 = \gamma$ because $C^{-1}\|R(I - z^{-1}F)\|_2^2 \ge C^{-1}C_R > 0$. Here, $C_R > 0$ is given in Lemma 16 in the appendix. ∎

The interesting aspect of this lemma is that the reformulation (2) of (1) automatically implies the constraint $\|F\|_2^2 < \gamma$. Let us introduce the following two auxiliary functions, which will allow us to further exploit this property.

Lemma 10. *The functions $\Psi, \Psi_\gamma : (0,\infty) \to (0,\infty)$,*

$$\Psi(C) := \inf_{F\in\mathcal{RH}_\infty^{q\times q}} \left\| \begin{bmatrix} C^{-1}R \\ 0 \end{bmatrix} - \begin{bmatrix} z^{-1}C^{-1}R \\ I \end{bmatrix} F \right\|_2^2,$$

$$\Psi_\gamma(C) := \inf_{\substack{F\in\mathcal{RH}_\infty^{q\times q} \\ \|F\|_2^2<\gamma}} \left\| \begin{bmatrix} C^{-1}R \\ 0 \end{bmatrix} - \begin{bmatrix} z^{-1}C^{-1}R \\ I \end{bmatrix} F \right\|_2^2,$$

satisfy $\Psi(C) \le \Psi_\gamma(C)$ and, if $\Psi(C) \le \gamma$, $\Psi(C) = \Psi_\gamma(C)$.

Proof: The fact that $\Psi(C) \le \Psi_\gamma(C)$ for all $C > 0$ follows trivially from the definitions of Ψ and Ψ_γ. Now, suppose that $\Psi(C) \le \gamma$ for some $C > 0$. Lemma 2 shows that the infimum in the definition of Ψ is achieved by some F. The same arguments as in the proof of Lemma 9 show that this F automatically satisfies $\|F\|_2^2 < \gamma$. But then $\Psi_\gamma(C) \le \Psi(C)$. ∎

The next lemma constitutes the last technical preliminary before the main result of this subsection. It establishes several useful properties of the function Ψ_γ.

Lemma 11. *The function Ψ_γ is strictly monotonously decreasing, continuous, and has the limits $\lim_{C\searrow 0}\Psi_\gamma(C) = \infty$ and $\lim_{C\nearrow\infty}\Psi_\gamma(C) = 0$.*

We skip the proof of this Lemma. We can now establish existence and uniqueness of the solution to Problem 8. Moreover, the next proposition also establishes a numerically simple to exploit characterization of the optimal solution.

Proposition 12. *Problem 8 has a unique solution, i.e.,*

$$F_{opt} = \operatorname*{argmin}_{F\in\mathcal{RH}_\infty^{q\times q}} \left\| \begin{bmatrix} C_{opt}^{-1}R \\ 0 \end{bmatrix} - \begin{bmatrix} z^{-1}C_{opt}^{-1}R \\ I \end{bmatrix} F \right\|_2^2. \tag{3}$$

Here, C_{opt} is the unique solution to $\Psi(C_{opt}) = \gamma$.

Proof: Let us consider $D := \inf_{F\in\mathcal{RH}_\infty^{q\times q}, \|F\|_2^2<\gamma} \|R(I - z^{-1}F)\|_2^2/(\gamma - \|F\|_2^2)$. Our first goal is to show that C_{opt} given in the proposition is unique and equals D. Lemma 9 shows that D is the infimum over all $C > 0$ for which there exists some $F \in \mathcal{RH}_\infty^{q\times q}$ that satisfies (2). One can show that this condition is equivalent to $\Psi(C) \le \gamma$, which, by Lemma 10, is equivalent to $\Psi_\gamma(C) \le \gamma$. Thus, $D = \inf\{C > 0 : \Psi_\gamma(C) \le \gamma\}$. Lemma 11 now implies that D is the unique solution to $\Psi_\gamma(D) = \gamma$, which, again by Lemma 10, also is the unique solution to $\Psi(D) = \gamma$. But this is exactly the C_{opt} given in the proposition. Next, we want to compute the optimal solution, i.e., we want to find $F_{opt} \in \mathcal{RH}_\infty^{q\times q}$ such that $F = F_{opt}$ satisfies (1) for $C = C_{opt}$. Lemma 9 shows this is true if and only if $F = F_{opt}$ and $C = C_{opt}$ solve (2). Since $\Psi(C_{opt}) = \gamma$, we see that F_{opt} must be a solution to the model matching problem (3). Lemma 2 shows that F_{opt} is the only solution. ∎

Proposition 12 shows that computation of the solution to Problem 8 is simple because it is the solution to the model-matching problem (3). However, (3) involves the quantity C_{opt}, which has to be computed in advance. Luckily, computation of C_{opt} is also simple. The Lemmas 10 and 11 show that $\Psi(C) > \gamma$ for all $C < C_{opt}$ and $\Psi(C) < \gamma$ for all $C > C_{opt}$. Thus, a simple bisection algorithm may be used to compute C_{opt}. Note that each evaluation of Ψ again amounts to a model-matching problem.

B. Stable Feedback Loop

Finally, we consider the optimal feedback filter with respect to the extended model that stabilizes the feedback loop. The exact formulation is as follows.

Problem 13. Find $F \in \mathcal{RH}_\infty^{q\times q}$ such that $(I - z^{-1}F)^{-1} \in \mathcal{RH}_\infty^{q\times q}$, $\|F\|_2^2 < \gamma$, and $\|R(I - z^{-1}F)\|_2^2/(\gamma - \|F\|_2^2)$ is minimized.

The next theorem shows that Problem 13 usually is very simple to solve because in most cases the solution to Problem 8 already solves Problem 13. In the rare cases where this is not true, we can get arbitrarily close-to-optimal performance with an approximation scheme.

Theorem 14. *Let $F_\star$ denote the solution to Problem 8. Then, the following holds.*

1) *Suppose $I - z^{-1}F_\star$ has no unit circle zeros. Then, $F_\star$ also solves Problem 13.*
2) *Suppose $I - z^{-1}F_\star$ has unit circle zeros, and introduce a family of outer-coinner factorizations $[I - z^{-1}F_\star, \epsilon I] = F_{\star o}^{(\epsilon)} F_{\star i}^{(\epsilon)}$ and a family of feedback filters $F_{opt}^{(\epsilon)}(z) := I - F_{\star o}^{(\epsilon)}(z)[F_{\star o}^{(\epsilon)}(\infty)]^{-1}$. Then, $zF_{opt}^{(\epsilon)}$, $(I - z^{-1}zF_{opt}^{(\epsilon)})^{-1} \in \mathcal{RH}_\infty^{q\times q}$ for all $\epsilon > 0$. Furthermore, we have $\|zF_{opt}^{(\epsilon)}\|_2^2 < \gamma$ for all $\epsilon < \sqrt{\gamma - \|F_\star\|_2^2}$, and*

$$\lim_{\epsilon \searrow 0} \frac{\|R(I - z^{-1}zF_{opt}^{(\epsilon)})^{-1}\|_2^2}{\gamma - \|zF_{opt}^{(\epsilon)}\|_2^2} = \frac{\|R(I - z^{-1}F_\star)^{-1}\|_2^2}{\gamma - \|F_\star\|_2^2}.$$

Proof: We only sketch the proof of part 1). Similarly to the proof of Theorem 6, one uses the solution $F_\star$ to Problem 8 in order to construct a solution F_{opt} to Problem 13 such that $\|R(I - F_{opt})\|_2^2 \le \|R(I - z^{-1}F_\star)\|_2^2$ and $\|F_{opt}\|_2^2 = \|F_\star\|_2^2$. But then does F_{opt} also solve Problem 8. The uniqueness result in Proposition 12 now shows $F_{opt} = F_\star$. ∎

We close this section with the observation that as soon as Problem 13 has a solution, this solution must be $F_\star$. The approximation scheme proposed in the second part of Theorem 14 only becomes necessary if an optimal solution does not exist. It then allows to obtain arbitrarily close-to-optimal suboptimal solutions. The proof will be skipped.

Proposition 15. *Let $F_\star$ denote the solution to Problem 8. Then, $F_\star$ is the unique solution to Problem 13 if and only if $I - z^{-1}F_\star$ has no unit circle zeros. Otherwise, there is no solution.*

VI. Numerical Examples

In this section, we illustrate possible performance gains of our new feedback filters derived in Section VI.

First example: We consider an example from [7, Sec. VI]. The analysis filter bank consists of a low-pass, a band-pass, and a high-pass Butterworth filter, $E(z) = [\frac{0.4208+0.4208z}{-0.1584+z}, \frac{-0.2452+0.2452z^2}{0.5095+z^2}, \frac{0.4208-0.4208z}{-0.1584-z}]^T$. Algorithm 2 in [8] gives us the optimal synthesis filter bank for the decision delay $L = 2$, $R(z) \approx [\frac{p_1(z)}{q(z)}, \frac{p_2(z)}{q(z)}, \frac{p_1(-z)}{q(z)}]$, where $p_1(z) := -0.005 - 0.006z + 0.189z^2 + 0.219z^3 + 0.453z^4 + 0.476z^5 + 0.096z^6$, $p_2(z) := -0.003 + 0.115z^2 + 0.070z^4 - 0.329z^6$, and $q(z) := 0.028z^2 + 0.396z^4 + z^6$. The quantizer used is $v_q = \operatorname{argmin}_{\tilde{v}_q \in \{-1,0,1\}^{3\times 1}} \|\tilde{v}_q - v\|_F^2$.

We compare the following feedback filter: direct quantization, i.e., $F(z) = 0$, the optimal order 45 FIR feedback filter from [1],[3] and the solution to Problem 13 for various assumed quantizer SNRs γ. Furthermore, we compare with the moving horizon subband quantizer from [2] (with horizon one). Per simulation, we quantized 1000 Gaussian random variables $[x_1, \ldots, x_{1000}]$ with zero mean and unit variance. The data vector was extended by L zeros in order to comprehend the decision delay. Our error measure was the mean square error $MSE = \frac{1}{1000}\sum_{k=1}^{1000} |x_{q,k+L} - x_k|^2$, where the $x_{q,k}$ denote the outputs of the subband quantizer (cf. Fig. 3 and remember that $p = 1$). We averaged our results over 1000 runs.

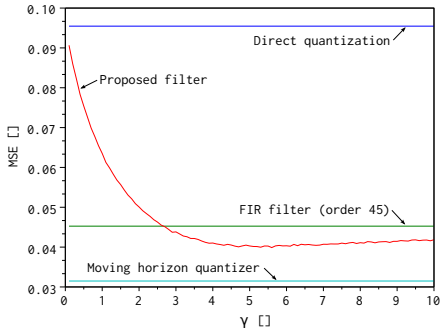


Figure 4: Simulation results (first example)

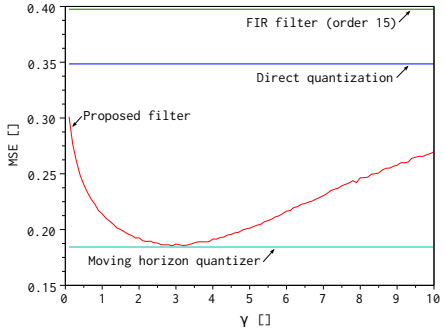


Figure 5: Simulation results (second example)

The simulation results are shown in Fig. 4. Our proposed feedback filter outperforms both direct quantization and the FIR filter from [1] (for $\gamma \approx 5$), but not the (more complex) moving horizon subband quantizer from [2], which still achieves an about 21% better performance. We have also plotted the MSEs of the optimal FIR filters from [1] as a function of the filter order in Fig. 6 because we wanted to show that the improved performance of our proposed filter is not a result of an insufficient FIR filter order but is due to the

[3]We point out that the optimal FIR filter in [1] usually is not unique because the linear system [1, (21)] often is rank deficient. This is also what Prop. 4 suggests. We can prove that the least-squares solution to [1, (21)] always gives the optimal FIR filter (regarding the standard criterion) with the smallest norm, and that this filter results in a stable feedback loop. Therefore, we have always used the pseudoinverse to get a least squares solution to [1, (21)] in our simulations. Unfortunately, this approach turned out to be numerically delicate. We often observed unstable feedback loops although in theory they should always be stable. Increasing filter orders seemed to improve the numerical properties of [1, (21)].

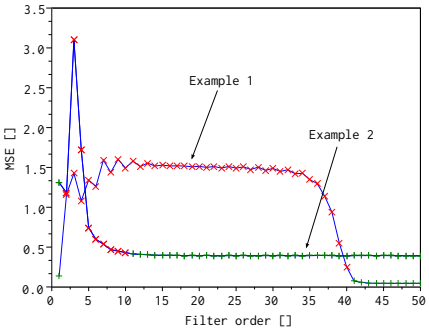


Figure 6: FIR filter performance. The "x" mark unstable feedback loops, while the "+" mark stable feedback loops.

extended model. The results are shown in Fig. 6. Note that the very high FIR filter order of 45 was indeed necessary in order to obtain satisfying performance. We have also marked the FIR filters in Fig. 6 which resulted in an unstable feedback loop. We again point out that the unstable feedback loops result from numerical inaccuracies (cf. Footnote 3).

Second example: Our second example is the random FIR synthesis filter bank $E(z) = [0.513 + 0.694z^{-1}, 0.742 + 0.635z^{-1}]^T$. The optimal synthesis filter bank for the decision delay $L = 4$ is $R(z) \approx \frac{1}{0.794z^3+z^4}[0.492 + 0.275z - 0.218z^2 + 0.173z^3 - 0.430z^4, 0.712 + 0.044z - 0.035z^2 + 0.028z^3 + 0.297z^4]$ [8, Alg. 2]. The other parameters are chosen as in the previous example.

The simulation results are shown in Fig. 5. Our proposed filter again outperforms both direct quantization and the optimal order 15 FIR filter from [1] (for $\gamma \approx 3$). Moreover, this time our proposed filter even achieves the performance of the more complex moving horizon subband quantizer (with horizon one) from [2]. It is interesting to note that the FIR filter performs even worse than direct quantization. Again, this is not due to an insufficient FIR filter order, and some of the FIR filters have stability issues (cf. Fig 6).

VII. Conclusion

We have proposed a new design method for feedback filters in oversampled noise-shaping subband quantizers. Our new method relies on an extended model of the quantizer, which allows us to take both the influence of the feedback filter on the quantization noise and the influence of the input data into account. We were able to show that in contrast to the standard model the optimal solutions regarding the extended model do not suffer from non-uniqueness (Props. 4, 7, 12, 15). A simple method to compute the optimal feedback filter regarding the extended model was given in Prop. 12. Furthermore, it was shown in Prop. 15 that our design method guarantees stability of the feedback loop. Numerical examples showed (sometimes significantly) increased performance compared to the FIR feedback filters for the standard model.

Appendix

Lemma 16. *There exists $C_R > 0$ such that $\|R(I - z^{-1}F)\|_2^2 \geq C_R$ for all $F \in \mathcal{RH}_\infty^{q\times q}$.*

Proof: Let us write $R(z) = \sum_{k=0}^{\infty} R_k z^{-k}$ and $(z^{-1}RF)(z) = \sum_{k=0}^{\infty} X_k z^{-k}$, and denote by κ the smallest k such that $R_k \neq 0$. Note that $R_k = X_k = 0$ for all $k < \kappa$ and $X_\kappa = 0$. Then, by Parsevals relation, we have $\|R(I - z^{-1}F)\|_2^2 = \sum_{k=\kappa}^{\infty} \|R_\kappa - X_k\|_F^2 = \|R_\kappa\|_F^2 + \sum_{k=\kappa+1}^{\infty} \|R_{\kappa+k} - X_k\|_F^2 \geq \|R_\kappa\|_F^2 > 0$. Since $\|R_\kappa\|_F^2$ is independent of F, the lemma follows. ∎

Lemma 17. *Suppose $M \in \mathcal{RH}_\infty^{m\times l}$. Then, $\|M(\infty) - M\|_2^2 = \|M\|_2^2 - \|M(\infty)\|_F^2$.*

Proof: Let us write $M(z) = \sum_{k=0}^{\infty} M_k z^{-k}$. Then, $\|M\|_2^2 = \sum_{k=0}^{\infty} \|M_k\|_F^2$ and the lemma follows from the observation that $M(\infty) = \lim_{z\to\infty} M(z) = M_0$. ∎

Lemma 18. *Let $M \in \mathcal{RH}_\infty^{m\times l}$ have full row normal rank, and no unit circle zeros. Suppose $M(\infty) = I$, and let $M = M_o M_i$ denote an outer-coinner factorization. Then, $M_o^{-1}(\infty) = UH^{-1}$, where $H^{-1} = H^{-*}$ has spectral radius at most one and U is an unitary matrix.*

Proof: The outer factor M_o also is a spectral factor of $MM^\sim$, i.e., $MM^\sim = M_o M_i M_i^\sim M_o^\sim = M_o M_o^\sim$ and $M_o, M_o^{-1} \in \mathcal{RH}_\infty^{m\times m}$. But these are unique up to multiplication with a unitary matrix from the right [9, Th. 2]. Thus, using the special outer factor given in [4, Th. 2] (subject to transposition because we need a outer-coinner and not a inner-outer factorization), we can write $M_o(z) = [I - C(zI - A)^{-1}F^*]H^*U$ for some unitary U, where $M(z) = D + C(zI - A)^{-1}B$ denotes a spate-space realization. The lemma now follows from the construction of H. We have $H = (DD^* + B^*XB)^{1/2} = (I + CXC^*)^{1/2} \geq I$ for some $X = X^* \geq 0$. ∎

Lemma 19. *Suppose $X = X^* \in \mathbb{C}^{q\times q}$ has a spectral radius at most one, and $B \in \mathbb{C}^{q\times q}$. Then, $\|BX\|_F^2 \leq \|B\|_F^2$.*

Proof: We skip the proof due to space limitations. ∎

References

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

[2] D. E. Quevedo, H. Bolcskei, and G. C. Goodwin, "Quantization of filter bank frame expansions through moving horizon optimization," *IEEE Trans. Signal Process.*, vol. 57, pp. 503–515, Feb. 2009.

[3] M. S. Derpich, E. I. Silva, D. E. Quevedo, and G. C. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Trans. Signal Process.*, vol. 56, pp. 3871–3890, Aug. 2008.

[4] V. Ionescu and C. Oara, "Spectral and inner-outer factorizations for discrete-time systems," *IEEE Trans. Autom. Control*, vol. 41, pp. 1840–1845, Dec. 1996.

[5] A. Saberi, A. A. Stoorvogel, and P. Sannuti, *Filtering Theory: With Applications to Fault Detection, Isolation, and Estimation*. Boston: Birkhäuser, 2007.

[6] M. A. Gerzon and P. G. Craven, "Optimal noise shaping and dither of digital signals," in *Proc. AES Conv.*, (New York, USA), Oct. 1989.

[7] L. Chai, J. Zhang, C. Zhang, and E. Mosca, "Frame-theory-based analysis and design of oversampled filter banks: Direct computational method," *IEEE Trans. Signal Process.*, vol. 55, pp. 507–519, Feb. 2007.

[8] S. Wahls and H. Boche, "Novel system inversion algorithm with application to oversampled perfect reconstruction filter banks," *IEEE Trans. Signal Process.*, vol. 58, June 2010.

[9] D. Youla, "On the factorization of rational matrices," *IRE Trans. Inf. Theory*, vol. 7, pp. 172–189, July 1961.