

## OPTIMUM PRE- AND POSTFILTERS FOR QUANTIZATION

Jamal Tuqan and P.P. Vaidyanathan  
 Department of Electrical Engineering 136-93  
 California Institute of Technology  
 Pasadena, CA 91125, USA.

E-mail : tuqan@systems.caltech.edu, ppvnath@sys.caltech.edu

**Abstract.** We consider the optimization of pre- and post filters surrounding a uniform quantizer such that the mean square error due to quantization is minimized. Unlike some previous work, the postfilter is not restricted to be the inverse of the prefilter. With no order constraint on the filters, we present closed form solutions for the optimum pre- and post filters. Using these optimum solutions, we obtain a coding gain expression for the system under study. We then repeat the same analysis with first order pre- and post filters in the form  $1 + \alpha z^{-1}$  and  $1/(1 + \gamma z^{-1})$  providing some examples where we compare coding gain performance with the case of  $\alpha = \gamma$ .

## I. INTRODUCTION

Consider the general scheme shown in Fig. 1 where the box labeled  $\mathcal{Q}$  represents a uniform quantizer. The input sequence  $x(n)$  is passed through a prefilter  $G(e^{j\omega})$  and produces an output  $y(n)$ . The sequence  $y(n)$  is then quantized and filtered with a postfilter  $H(e^{j\omega})$  to reproduce an estimate of the input denoted by  $\hat{x}(n)$ . Assuming that the quantization system is constrained to have a budget of  $b$  bits, the main theme in this paper is to jointly optimize the prefilter  $G(e^{j\omega})$  and the postfilter  $H(e^{j\omega})$  such that the mean square value  $E\{e^2(n)\}$  of the reconstruction error where  $e(n) \triangleq \hat{x}(n) - x(n)$  is minimized.

The renewed interest in the above classic problem [1] was motivated by the growing activity in optimizing perfect reconstruction filter banks in the presence of quantizers [2], [3], [4], [5]. When quantizers are present, the FB output  $\hat{x}(n)$  is the original input  $x(n)$  plus a filtered version of the quantization noise  $e(n)$ . Searching over the class of perfect reconstruction filter banks (PRFB), the goal is to find a set of analysis filters to minimize the quantization noise  $e(n)$ . Although this problem was solved for the class of orthonormal FB [ideal filter case] [6], [7], [8] [9], the  $M$  channel maximally decimated

optimum biorthogonal FB remains for example an open problem. Only the solution of the one channel case is well established [10]. Furthermore, it is well known [11], [2] that, in the presence of quantizers, the synthesis polyphase matrix is not necessarily the inverse of the analysis polyphase matrix. Restricting ourselves to the class of biorthogonal FB when quantizers are present is therefore a loss of generality. The joint optimization of the analysis bank and the synthesis bank together with the allocation of subband bits is quite a challenging problem. In this paper, we will provide a joint optimum solution of the pre- and post filters for the special case of  $M = 1$ . The system of Fig. 1 can indeed be seen as the one channel case of the more general and difficult  $M$  channel problem. It is also a generalization of the so-called half-whitening scheme [10] where the postfilter is assumed to be the inverse of the prefilter.

## II. THE OPTIMUM POST FILTER

In this section, we assume that all random processes are zero mean, real and jointly wide sense stationary. The input  $x(n)$  and the quantization noise  $q(n)$  are uncorrelated, i.e.,  $E\{x(n)q(m)\} = 0 \forall n, m$ . The quantization noise  $q(n)$  is white with variance

$$\sigma_q^2 = c2^{-2b}\sigma_y^2 \quad (2.1)$$

where  $\sigma_q^2$  is the quantization noise variance,  $c$  is a constant that depends on the statistical distribution of  $y(n)$  and the overflow probability, and  $\sigma_y^2$  is the variance of the quantizer input. The filters  $H(e^{j\omega})$  and  $G(e^{j\omega})$  are not constrained to be rational functions, i.e., the optimum  $H(e^{j\omega})$  and  $G(e^{j\omega})$  can be ideal filters. Furthermore, no causality constraint is imposed. To develop optimum closed form solutions for both filters, we first fix the prefilter  $G(e^{j\omega})$  and optimize  $H(e^{j\omega})$ . The optimum post filter solution is given in the following theorem. The proof of all four theorems included in this paper can be found in [12].

**Theorem 2.1.** For a fixed prefilter  $G(e^{j\omega})$ , the optimum postfilter  $H_{opt}(e^{j\omega})$  is the well-known Wiener filter and is given by :

$$H_{opt}(e^{j\omega}) = \frac{1}{G(e^{j\omega})} \cdot \frac{S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega}) + \sigma_q^2} \quad (2.2)$$

$$\text{where } \sigma_q^2 = \frac{c2^{-2b}}{|G(e^{j\omega})|^2} \cdot \int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) |G(e^{j\omega})|^2 \frac{d\omega}{2\pi}$$

Using the optimum post filter solution (2.2), the orthogonality principle, Parseval's relation and the constraint (2.1), we can derive the following expression for the mean square error  $\mathcal{E}$  only in terms of the prefilter  $G(e^{j\omega})$ :

$$\mathcal{E}(|G|^2) = \int_{-\pi}^{\pi} \frac{\sigma_q^2 S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega}) |G(e^{j\omega})|^2 + \sigma_q^2} \frac{d\omega}{2\pi} \quad (2.3)$$

$$\text{where } \sigma_q^2 = \frac{c2^{-2b}}{|G(e^{j\omega})|^2} \cdot \int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) |G(e^{j\omega})|^2 \frac{d\omega}{2\pi}$$

The problem now reduces to finding the prefilter  $G(e^{j\omega})$  that minimizes  $\mathcal{E}$ . Since the mean square error expression (2.3) is a function of  $|G(e^{j\omega})|^2$  only, we will be actually seeking an expression for the squared magnitude response of the prefilter rather than  $G(e^{j\omega})$ . Instead of attacking the problem directly, the idea is to transform the above unconstrained integral (2.3) to a communication-like problem, i.e., an integral objective function with a power constraint on the prefilter output. The problem then becomes more mathematically tractable and a closed form expression for  $|G(e^{j\omega})|^2$  can be obtained. The equivalence of the two problems is established by the following claim [12].

**Theorem 2.2.** The squared magnitude response  $|G_{opt}(e^{j\omega})|^2$  that minimizes  $\mathcal{E}(|G|, b)$  is also the solution of the following constrained optimization problem:

$$\min_{|G(e^{j\omega})|^2} \int_{-\pi}^{\pi} \frac{c2^{-2b} S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega}) |G(e^{j\omega})|^2 + c2^{-2b}} \frac{d\omega}{2\pi} \quad (2.4)$$

subject to:

$$\int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) |G(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 1 \quad (2.5)$$

### III. THE OPTIMUM PRE-FILTER

The goal now is to find  $|G(e^{j\omega})|^2$  that minimizes the integral in (2.4) under the integral constraint (2.5). Since the magnitude squared response is always a

non negative function of  $\omega$ , the optimum minimizing solution we seek must be non negative. This implicit condition is incorporated in the optimization problem as a *pointwise inequality* constraint. The next theorem [12] gives an expression for the optimum magnitude squared response of the prefilter.

**Theorem 3.1.** The magnitude squared of the prefilter  $|G_{opt}(e^{j\omega})|^2$  that minimizes the integral in (2.4) under the constraint (2.5) must have the following form:

$$\frac{1}{\sqrt{S_{xx}(e^{j\omega})}} \cdot \left( \frac{1 + c2^{-2b}}{\int_{-\pi}^{\pi} \sqrt{S_{xx}(e^{j\omega})} \frac{d\omega}{2\pi}} - \frac{c2^{-2b}}{\sqrt{S_{xx}(e^{j\omega})}} \right) \quad (3.1)$$

for all  $\omega \in [-\pi, \pi]$  for which (3.1) is non negative. Otherwise, the value of the optimum magnitude squared response is set to zero.

### IV. THE CODING GAIN EXPRESSION

Assume that we quantize  $x(n)$  directly with  $b$  bits. We denote the corresponding mean square error (m.s.e) by  $\mathcal{E}_{direct}$ . We then use the optimum pre and post filters around the quantizer. With the rate of the quantizer fixed to the same value  $b$ , we denote the m.s.e in this case by  $\mathcal{E}_{new}$ . The ratio  $\mathcal{G}_{opt} \triangleq \mathcal{E}_{direct}/\mathcal{E}_{new}$  is called the coding gain of the new system and, as the name suggests, is a measure of the benefits provided by the pre/post filtering operation. The coding gain of the scheme of Fig. 1 is given in the following theorem [12].

**Theorem 4.1.** With the optimal choice of pre- and postfilters, the coding gain expression for the scheme of Fig. 1 is

$$\mathcal{G}_{opt} = (1 + c \cdot 2^{-2b}) \cdot \mathcal{G}_{hw} \quad (4.1)$$

as long as (3.1) is non-negative  $\forall \omega$ . Here  $\mathcal{G}_{hw}$  is the coding gain of the half whitening scheme and is given by

$$\frac{\int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi}}{\left( \int_{-\pi}^{\pi} \sqrt{S_{xx}(e^{j\omega})} \frac{d\omega}{2\pi} \right)^2} \quad (4.2)$$

It is quite clear from theorem 3.1 that the system of Fig. 1 will always outperform the half whitening scheme as long as equation (3.1) remains non negative for all frequencies. The difference in performance is essentially a function of the bit rate. The lower the bit rate the higher the coding gain will be. However, the problem is that as we quantize

at lower bit rates, the quantizer assumptions made at the beginning of this section fail and all the previous analysis is not valid anymore. On the other hand, as  $b$  becomes large, one can easily check that the pre-filter (3.1) becomes

$$\frac{1}{\sqrt{S_{xx}(e^{j\omega})}} \left( \frac{1}{\int_{-\pi}^{\pi} \sqrt{S_{xx}(e^{j\omega})} \frac{d\omega}{2\pi}} \right) \forall \omega \in [-\pi, \pi]$$

and is positive  $\forall \omega$ . Therefore, the coding gain expression derived in Theorem 4.1 can be used and in fact becomes equal to  $\mathcal{G}_{hw}$ . Hence, at high bit rate, there is no loss of generality in using the half whitening scheme. The same observation can be also found in the work of Goodman and Drouilhet [13] although their approach is different than ours.

## V. FIRST ORDER FILTERS

In this section, we will constrain  $H(e^{j\omega})$  and  $G(e^{j\omega})$  to be first order causal filters in the form  $1 - \alpha z^{-1}$  and  $\frac{1}{1 - \gamma z^{-1}}$ . We again jointly optimize the first order pre- and post filters to minimize the m.s.e under the constraint (2.1) and the other assumptions of section II. We consider two main cases: a) an FIR prefilter with an IIR postfilter and b) an IIR prefilter with an FIR postfilter. The mean square error and optimum coefficients expressions for each case are not included in this paper due to lack of space and can be found in [12]. We will instead directly provide some numerical results for two specific examples of input  $x(n)$  with comparison to the  $\alpha = \gamma$  case.

**Example 4.1.1** *Case of a MA(1) process.* Assume that the input  $x(n)$  is a zero mean MA(1) process with an autocorrelation sequence in the form

$$R_{xx}(k) = \begin{cases} 1 & k = 0. \\ \frac{\theta}{1 + \theta^2} & k = 1, -1. \\ 0 & \text{otherwise.} \end{cases}$$

The optimization of the coefficients for the IIR prefilter-FIR postfilter case and the FIR prefilter-IIR postfilter with  $\alpha \neq \gamma$  were all done numerically using MATLAB's optimization toolbox routine "fmins.m". The plots of the coding gain are illustrated in Fig. 2 (a) and (b) for the FIR/IIR case and in Fig. 2 (c) and (d) for the IIR/FIR case. The vertical axis represents the coding gain magnitude in db and the horizontal axis represents the parameter  $\theta$  defined above. From these figures, we can observe that as the bit rate increases, there is no loss of generality in assuming  $\alpha$  to be equal to  $\gamma$ . We also note that the coding gain obtained in the FIR/IIR case is higher than the dual case for

the same process and bit rate. This is primarily due to the fact that the optimum coefficients in the IIR/FIR case are numerically close to zero and the coding gain is therefore close to one.

**Example 4.1.2** *Case of an AR(5) process.* Table 1 summarizes our coding gain results in db for the different cases and bit rates. Again, as  $b$  increases, we observe that there is almost no loss in coding gain if we assume that  $\alpha = \gamma$ . We also observe that, at low bit rate, e.g.  $b = 1$ , the coding gain of the more general system is very small. This suggests that the gain obtained from searching over a more general class than the biorthogonal class may not be worth the added complexity. Finally, as was the case for the previous example, the FIR/IIR scheme outperforms substantially the dual case.

## References

- [1] Costas, J.P., "Coding with linear systems", Proceedings of the IRE, pp. 1101-1103, Sept. 1952.
- [2] Kovacevic, J., "Subband coding systems incorporating quantizer Models", IEEE Trans. on Image Processing, Vol. 4, pp. 543-553, May 1995.
- [3] Uzun, N. and Haddad, R.A., "Cyclostationary Modeling, analysis and optimal compensation of quantization errors in subband codecs", IEEE Trans. on SP, Vol. 43, pp. 2109-2119, Sept. 1995.
- [4] Tabatabai, A. "Optimum analysis/synthesis filter bank structures with application to subband coding systems", ISCAS 1988, pp. 823-826.
- [5] Delopoulos, A. and Koliass, S., "Optimal filterbanks for signal reconstruction from noisy subband components", Proc. 28th Annual Asilomar conf. Sig., Sys. and Comp., Oct-Nov. 1994.
- [6] Unser, M., "On the optimality of ideal filters for pyramid and wavelet signal approximation", IEEE Trans. on SP, pp. 3591-3596, December 1993.
- [7] Delsarte, P., Macq, B. and Slock, D., "Signal-adapted multiresolution transform for image coding", IEEE Trans. on Information Theory, Vol. 38, pp. 897-904, March 1995.
- [8] Tsatsanis, M.K. and Giannakis, G.B., "Principal component filter banks for optimal multiresolution analysis", IEEE Trans. on Signal Processing, pp. 1766-1777, Vol. 43, August 1995.
- [9] Vaidyanathan, P. P., "Optimal orthonormal filterbanks", submitted, IEEE Trans. on SP.
- [10] Jayant, N.S. and Noll, P. *Digital coding of waveforms*, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1984.
- [11] Vaidyanathan, P. P. and Chen, T., "Statistically Optimal synthesis banks for subband coders reconstruction", Proc. 28th Annual Asilomar conf. Sig., Sys. and Comp., Oct-Nov. 1994.

[12] Tuqan, J. and Vaidyanathan, P. P., "Statistically optimum pre- and post filtering in quantization", Submitted, IEEE Trans. on CAS.

[13] Goodman, L.M. and Drouillet, P.R., "Asymptotically optimum pre-emphasis and de-emphasis networks for sampling and quantizing", Proc. IEEE, pp. 795-796, May 1963.

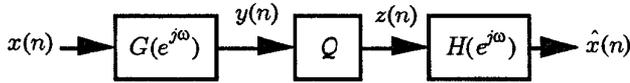


Fig. 1. A general pre- and post filtering scheme.

	b = 1	b = 2	b = 3
FIR/IIR $\alpha \neq \gamma$	2.008	1.972	1.963
FIR/IIR $\alpha = \gamma$	1.96	1.96	1.96
IIR/FIR $\alpha \neq \gamma$	1.091	1.087	1.086
IIR/FIR $\alpha = \gamma$	1.0852	1.0852	1.0852

Table 1. The coding gain obtained from first order filters for the AR(5) case.

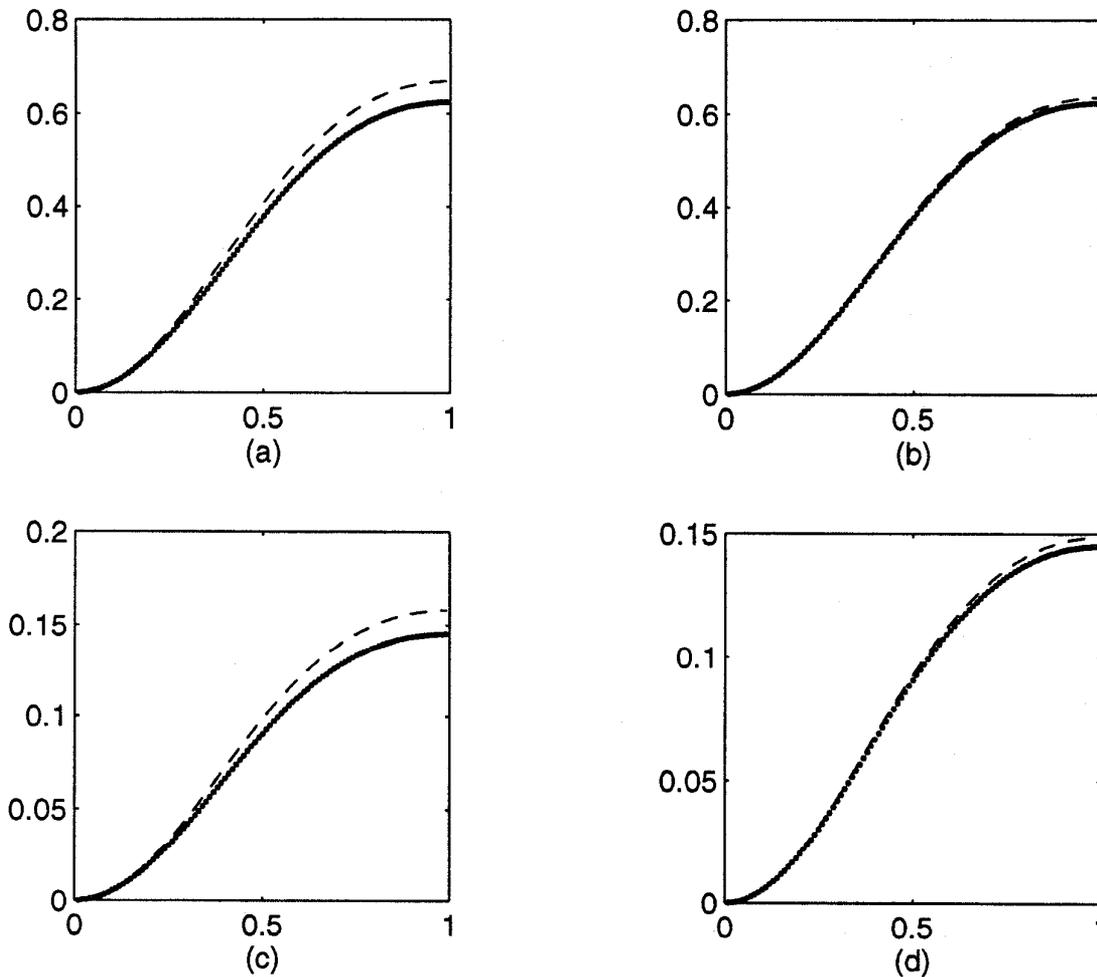


Fig. 2. (a) FIR prefilter, IIR postfilter, b = 1, (b) FIR prefilter, IIR postfilter, b = 2, (c) IIR prefilter, FIR postfilter, b = 1, (d) IIR prefilter, FIR postfilter, b = 2. For all plots, .... : equal coefficients and - - - : unequal coefficients.