Statistically Optimum Pre- and Postfiltering in Quantization

Jamal Tuqan, *Student Member, IEEE,* and P. P. Vaidyanathan, *Fellow, IEEE*

***Abstract*—We consider the optimization of pre- and postfilters surrounding a quantization system. The goal is to optimize the filters such that the mean square error is minimized under the key constraint that the quantization noise variance is directly proportional to the variance of the quantization system input. 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 postfilters when the quantization system is a uniform quantizer. Using these optimum solutions, we obtain a coding gain expression for the system under study. The coding gain expression clearly indicates that, at high bit rates, there is no loss in generality in restricting the postfilter to be the inverse of the prefilter. We then repeat the same analysis with first-order pre- and postfilters in the form $1+\alpha z^{-1}$ and $1/(1+\gamma z^{-1})$. In specific, we study two cases: 1) FIR prefilter, IIR postfilter and 2) IIR prefilter, FIR postfilter. For each case, we obtain a mean square error expression, optimize the coefficients α and γ and provide some examples where we compare the coding gain performance with the case of $\alpha = \gamma$. In the last section, we assume that the quantization system is an orthonormal perfect reconstruction filter bank. To apply the optimum pre- and postfilters derived earlier, the output of the filter bank must be wide-sense stationary WSS which, in general, is not true. We provide two theorems, each under a different set of assumptions, that guarantee the wide sense stationarity of the filter bank output. We then propose a suboptimum procedure to increase the coding gain of the orthonormal filter bank.**

***Index Terms*—Half-whitening scheme, noise shaping, optimum pre- and postfiltering, subband coding.**

I. Introduction

CONSIDER the general scheme shown in Fig. 1 where the box labeled $\mathcal{QS}$ represents a quantization system. 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 $\hat{x}(n)$. The quantization system $\mathcal{QS}$ can be a simple uniform quantizer or a more sophisticated quantization system such as the M-channel uniform subband coder (SBC) shown in Fig. 2. 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.

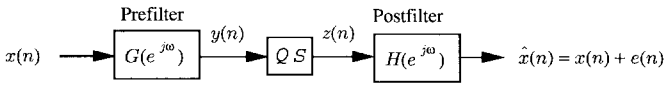

Fig. 1. A general pre- and postfiltering scheme.

The renewed interest in the above classic problem was motivated by its relation to some issues in the area of subband coding. To elaborate more, consider the M-channel uniform SBC of Fig. 2. The boxes labeled $\mathcal{Q}$ represent subband quantizers, a set of uniform quantizers which are modeled by additive noise sources. An equivalent representation of the uniform SBC is given in Fig. 3. It consists of two matrices $\mathbf{E}(e^{j\omega})$ and $\mathbf{R}(e^{j\omega})$, known, respectively, as the analysis and synthesis polyphase matrices. In the absence of quantizers, the filter bank (FB) is said to have the perfect reconstruction (PR) property if and only if $\mathbf{R}(e^{j\omega}) = \mathbf{E}^{-1}(e^{j\omega})$ [1]. A perfect reconstruction filter bank (PRFB) is also known as a biorthogonal FB. An important subclass of uniform PR filter banks is the class of orthonormal or paraunitary (PU) filter banks. In this case, the analysis polyphase matrix exhibits the lossless property, mathematically expressed as $\mathbf{E}(e^{j\omega})\mathbf{E}^{\dagger}(e^{j\omega}) = \mathbf{I}\ \ \forall\ \omega$, where the superscript † denotes the conjugate transpose operation. By choosing the synthesis polyphase matrix $\mathbf{R}(e^{j\omega})$ to be equal to $\mathbf{E}^{\dagger}(e^{j\omega})$, perfect reconstruction is guaranteed.

In the presence of quantizers, perfect reconstruction is not possible because quantization is a lossy process. The FB output $\hat{x}(n)$ in this case is the original input $x(n)$ plus a filtered version of the quantization noise denoted by $e(n)$. Recently, several authors have considered the optimization of filter banks when quantizers are present [2]–[5]. Given a fixed budget of b bits for the subband quantizers, the aim is to minimize the average variance of $e(n)$. This problem involves optimizing the analysis and synthesis filters and choosing a subband bit allocation strategy. For the sake of further discussions, we will from now on refer to the problem of optimizing a FB in the presence of quantizers as the subband coding problem. In a parallel fashion, interest in the so called energy compaction problem was growing [6]–[8]. Although the energy compaction problem might at first seem decoupled from the subband coding problem, Vaidyanathan [9] recently showed that the energy compaction problem and the subband coding problem for the case of an orthonormal SBC are actually highly connected. In fact, the orthonormal filter bank solution given in [9] for the subband coding problem turns out to be similar to the one given in [8] for the energy

Manuscript received June 12, 1995; revised July 8, 1996. This work was supported in part by the Office of Naval Research under Grant N00014-93-1-0231, by Tektronix, Inc., and by Rockwell International.

The authors are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: ppvnath@sys.caltech.edu).

Publisher Item Identifier S 1057-7130(97)07674-X.

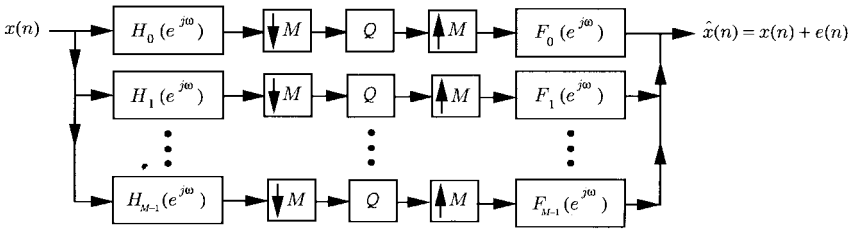

Fig. 2. An M-channel uniform maximally decimated subband coder (SBC).

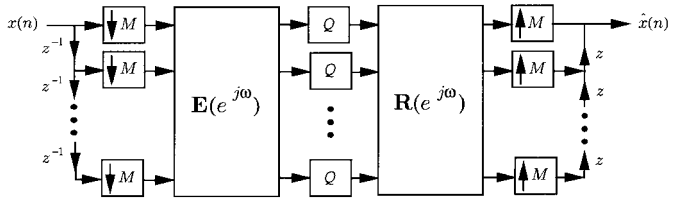

Fig. 3. The equivalent polyphase representation for a uniform M-channel subband coder (SBC).

compaction problem. Such filter banks are referred to as **optimum orthonormal filter banks**. We will use some of the results of optimum orthonormal filter banks later in Section V.

Although the subband coding problem was carefully analyzed and solved for the class of orthonormal FB [ideal filter case], the M channel $[M \neq 1]$ maximally decimated optimum biorthogonal FB is, at this point in time, an open problem. Only the solution of the one channel case is well established [10]. Furthermore, it is well known [11] 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. A similar observation was given by Gosse and Duhamel [4] calling this more general class of filter banks minimum mean-square-error (MMSE) filter banks. Kovacevic [2] also reaches the same conclusion for the case where the subband quantizer Q is modeled as a Lloyd–Max quantizer. While the synthesis bank was optimized in [11], Vaidyanathan and Chen did not address the issue of optimizing the analysis bank. Furthermore, their allocation of subband bits was done before optimizing the synthesis bank.

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 postfilters for the special case of $M = 1$. The system of Fig. 1 when the quantization system $\mathcal{QS}$ is a uniform quantizer 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. A summary of all the paper's results is given below.

A. Brief Overview of Past Related Work

The problem of finding optimum pre- and postfilters around a noisy processor has been considered by various researchers especially in the field of communication theory. Costas [12] has jointly optimized pre- and postfilters over an analog communication channel subject to a power constraint on the prefilter. Chan and Donaldson [13] considered the same problem with the input to the postfilter sampled every T seconds. Berger and Tufts [14] optimized transmission and receiving filters in PAM communication systems to minimize the mean square error (MSE) distortion resulting from channel noise and intersymbol interference. Malvar and Staelin [15] offered an iterative algorithm to design FIR pre- and postfilters in the presence of a downsampler and an upsampler.

The first fundamental difference between the above problems and the quantization problem under study in this paper is the nature of the noise variance. In specific, we will always assume throughout this paper that the quantization noise variance σ_q^2 is directly proportional to the variance of the input to the quantization system. Such a constraint describes in a fairly accurate manner the interaction between the quantization system granular noise output and the dynamic range of the quantization system input process. A simple example would be the relation $\sigma_q^2 = c2^{-2b}\sigma_y^2$ used in [10] for the case of a uniform quantizer. In a communication problem setting, the noise source variance is always assumed to be independent of the channel input signal statistics. The second main difference is that, in a communication problem, the prefilter is usually power constrained. This is not the case for the quantizer problem.

Taking a different approach than the one used in communications, Jayant and Noll analyzed the case where the quantization system $\mathcal{QS}$ is a simple uniform quantizer and the postfilter $H(e^{j\omega})$ is simply the inverse of the prefilter, i.e., $H(e^{j\omega}) = 1/G(e^{j\omega})$. Applying the Cauchy–Schwartz inequality, the magnitude response of the optimum filter can be found to be $|G_{\text{opt}}(e^{j\omega})| = 1/S_{xx}(e^{j\omega})^{1/4}$. The system was therefore called the half-whitening scheme [10] and represents an optimum one channel biorthogonal FB. Recently, Djokovic and Vaidyanathan [16] repeated the analysis for the case where the quantization system $\mathcal{QS}$ is a uniform orthonormal FB.

B. Main Results and Outline of the Paper

1) In the early sections of this paper, we will assume that the quantization system $\mathcal{QS}$ is a *uniform scalar quantizer*. With similar assumptions as the one used by Jayant and Noll in the derivation of the half-whitening solution, we derive optimum solution for the more general scheme of Fig. 1. In specific, closed-form expressions for the

optimum ideal pre- and postfilters are derived in Section II.

2) In Section III, using the optimum pre- and postfilters of Section II, we derive an expression for the so called coding gain of the scheme of Fig. 1. The beauty of this expression is that it clearly indicates that there is no loss of generality in using the half-whitening scheme if we are quantizing at high bit rate, a result that is intuitively very appealing.
3) In Section IV, we repeat the same type of analysis with first-order pre- and postfilters with monic polynomials. We derive an expression for the MSE for the cases of (a) FIR prefilter, IIR postfilter and (b) IIR prefilter, FIR postfilter. We then provide some examples where the coefficients of the filters can be computed numerically. We compare the coding gain of such cases with the one obtained from a first-order PR system. Our results indicate again that unless we are quantizing at a very low bit rate, the solution of the more general scheme of Fig. 1 tends to the half-whitening scheme.
4) In Section V, we assume that the quantization system $\mathcal{QS}$ is an orthonormal uniform PRFB. We do not however try to generalize the scheme proposed by Djokovic and Vaidyanathan [16]. Instead, we propose a suboptimum procedure. We first develop two theorems that give sufficient conditions for wide sense stationarity of the output noise of a nonuniform orthonormal PRFB. We then apply the optimum pre- and postfilters of Section II at the input and output of the FB, respectively, to improve the performance of the original orthonormal PRFB.

II. Optimum Unconstrained Pre- and Postfilters

The main goal of this section is to jointly optimize the prefilter $G(e^{j\omega})$ and postfilter $H(e^{j\omega})$ of Fig. 1 ($\mathcal{QS}$ is a uniform quantizer) to minimize the MSE $\triangleq E\{\hat{x}(n)-x(n)\}^2$ subject to the constraint

$$\sigma_q^2 = c2^{-2b}\sigma_y^2 \tag{2.1}$$

where σ_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 σ_y^2 is the variance of the quantizer input. Our main assumptions for this section are summarized as follows.

1) All random processes are zero mean, real and jointly wide sense stationary.
2) The input $x(n)$ and the quantization noise $q(n)$ are uncorrelated processes, i.e., $E\{x(n)q(m)\}=0 \ \ \forall n,m$.
3) The quantization noise $q(n)$ is white with variance σ_q^2 as in (2.1).
4) 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.
5) The power spectral density $S_{xx}(e^{j\omega})$ is positive for all ω. Furthermore, when deriving the optimum solution for the prefilter, we will also require $S_{xx}(e^{j\omega})$ and its first derivative to be continuous functions of frequency.

A. The Optimum Postfilter

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 postfilter solution is given in the following theorem.

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

$$H_{\rm opt}(e^{j\omega}) = \frac{1}{G(e^{j\omega})}\cdot\frac{S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})+\frac{c2^{-2b}}{|G(e^{j\omega})|^2}\int_{-\pi}^{\pi}S_{xx}(e^{j\omega})|G(e^{j\omega})|^2\frac{d\omega}{2\pi}} \tag{2.2}$$

Proof: For a fixed prefilter $G(e^{j\omega})$, the input to the postfilter $H(e^{j\omega})$ is a filtered version of the desired signal embedded in quantization noise. This is a classical Wiener filtering setting and hence, the optimum postfilter is given by [17] $H_{\rm opt}(e^{j\omega}) = S_{xz}(e^{j\omega})/S_{zz}(e^{j\omega})$ where $z(n) = y(n)+q(n)$ is the noisy input to the Wiener filter. Since $x(n)$ and $q(n)$ are assumed uncorrelated, it is easy to see that $S_{xz}(e^{j\omega}) = S_{xy}(e^{j\omega}) = G(e^{j\omega})^* S_{xx}(e^{j\omega})$ and $S_{zz}(e^{j\omega}) = S_{yy}(e^{j\omega})+\sigma_q^2 = |G(e^{j\omega})|^2 S_{xx}(e^{j\omega})+\sigma_q^2$ where the * denotes complex conjugation. Substituting in the above, we get

$$\begin{aligned} H_{\rm opt}(e^{j\omega}) &= \frac{G(e^{j\omega})^* S_{xx}(e^{j\omega})}{|G(e^{j\omega})|^2 S_{xx}(e^{j\omega})+\sigma_q^2} \\ &= \frac{1}{G(e^{j\omega})}\cdot\frac{S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})+\frac{\sigma_q^2}{|G(e^{j\omega})|^2}}. \end{aligned} \tag{2.3}$$

Substituting the constraint (2.1) in this last equation, we obtain the above solution. □

The optimum postfilter can be drawn as in Fig. 4. The Wiener filter of (2.2) is therefore expressed as a cascade of two filters: The first filter is the inverse of the prefilter $G(e^{j\omega})$. Its output is simply the original input $x(n)$ embedded in a filtered version of the quantization noise process. The power spectral density of the filtered quantization noise process is $\sigma_q^2/|G(e^{j\omega})|^2$. The second filter is the optimum Wiener filter for the output of the inverse filter.

Using the optimum postfilter solution (2.2) and the constraint (2.1), we can now derive an expression for the MSE only in terms of the prefilter $G(e^{j\omega})$:

$$\begin{aligned} \mathcal{E} &= E\{e^2(n)\} = E\{e(n)\cdot(\hat{x}(n)-x(n))\} \\ &= E\{e(n)\cdot x(n)\} = E\{(\hat{x}(n)-x(n))\cdot x(n)\} \\ &= R_{xx}(0) - \sum_{k=-\infty}^{\infty} h(k)\cdot E\{x(n)z(n-k)\} \\ &= R_{xx}(0) - \sum_{k=-\infty}^{\infty} h(k)R_{xz}(k). \end{aligned} \tag{2.4}$$

The second line is obtained from the first using the orthogonality principle [17]. By Parseval's relation, we can then write $\mathcal{E} = R_{xx}(0) - \int_{-\pi}^{\pi} S_{xz}^*(e^{j\omega})H_{\rm opt}(e^{j\omega})\frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) - S_{xz}^*(e^{j\omega})H(e^{j\omega})\frac{d\omega}{2\pi}$. Substituting with

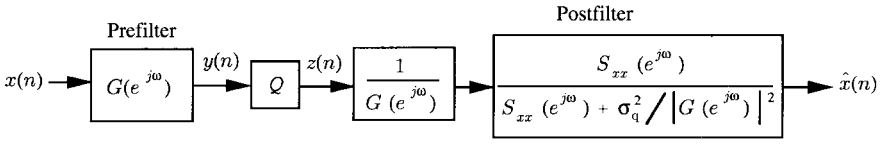

Fig. 4. The pre- and postfiltering scheme with a uniform quantizer and an optimum postfilter.

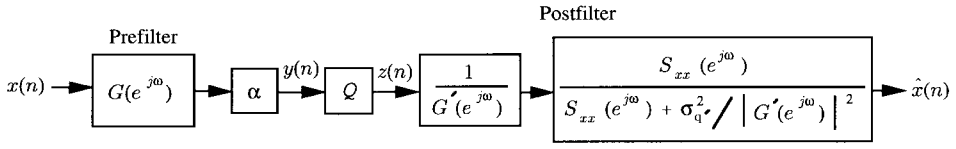

Fig. 5. Inserting a multiplier after the prefilter to study the effect of the quantizer input variance.

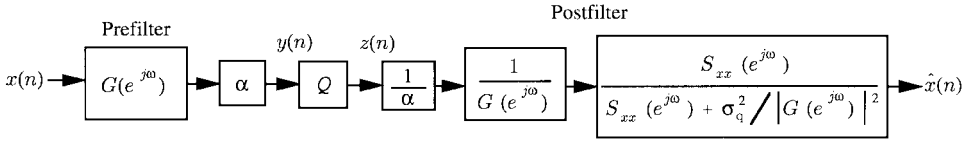

Fig. 6. The equivalent pre- and postfiltering scheme after the insertion of the multiplier.

$S_{xz}^*(e^{j\omega}) = S_{xy}^*(e^{j\omega}) = G(e^{j\omega})S_{xx}(e^{j\omega})$, we obtain

$$\mathcal{E} = \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})(1 - H_{\text{opt}}(e^{j\omega})G(e^{j\omega}))\frac{d\omega}{2\pi}. \tag{2.5}$$

We note that the previous (2.5) holds only for $H_{\text{opt}}(e^{j\omega})$. The reason is the use of the orthogonality principle in the derivation of (2.5). To obtain $\mathcal{E}$ only as a function of the prefilter $G(e^{j\omega})$, we substitute $H_{\text{opt}}(e^{j\omega})$ into (2.5) [see (2.6), given at the bottom of the page]. The problem now reduces to finding the prefilter $G(e^{j\omega})$ that minimizes $\mathcal{E}$ as given in (2.6). Two points are in order.

1) Since the MSE expression (2.6) 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 $G(e^{j\omega})$.
2) It is clear from (2.6) that trying to derive an optimum analytical expression for $|G(e^{j\omega})|^2$ can be quite tedious. Instead of attacking the problem as it is, the idea is to transform the above unconstrained integral (2.6) into another integral 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. It remains to show that the solution of both problems, the original one and the equivalent one, is the same. This is done in the following claim.

Theorem 2.2: The squared magnitude response $|G_{\text{opt}}(e^{j\omega})|^2$ that minimizes $\mathcal{E}(|G|, b)$, given as in (2.6), is also the solution of the following constrained optimization problem:

$$\min_{|G(e^{j\omega})|^2} \int_{-\pi}^{\pi} \frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} \frac{d\omega}{2\pi} \tag{2.7}$$

subject to

$$\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 1. \tag{2.8}$$

Proof: The role of the magnitude response of the prefilter is basically two fold: It affects the spectral shape of the quantizer input signal $y(n)$ and it changes the quantizer input variance σ_y^2 and therefore the noise variance. The idea is to insert a multiplier α directly before the quantizer. The insertion of this multiplier affect only the variance of the quantizer input. One can then show that the MSE at the output of this new system is unaffected by this multiplier. This, in turn, indicates that we can always fix the variance of the quantizer input signal $y(n)$ without changing the solution of our original problem. To prove the argument formally, we proceed as follows. Define

$$G'(e^{j\omega}) \triangleq \alpha G(e^{j\omega})$$

such that

$$\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G'(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 1. \tag{2.9}$$

Hence,

$$\begin{aligned}\sigma_q^2 &= c2^{-2b}\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 \frac{d\omega}{2\pi} \\ &= \frac{1}{\alpha^2}c2^{-2b}\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G'(e^{j\omega})|^2 \frac{d\omega}{2\pi} \\ &= \frac{1}{\alpha^2}c2^{-2b} = \frac{1}{\alpha^2}\sigma_{q'}^2\end{aligned} \tag{2.10}$$

$$\mathcal{E}(|G|, b) = \int_{-\pi}^{\pi} \frac{c2^{-2b}S_{xx}(e^{j\omega})\int_{-\pi}^{\pi} S_{xx}(e^{ju})|G(e^{ju})|^2 \frac{du}{2\pi}}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}\int_{-\pi}^{\pi} S_{xx}(e^{ju})|G(e^{ju})|^2 \frac{du}{2\pi}} \frac{d\omega}{2\pi}. \tag{2.6}$$

where $\sigma_{q'}^2$ is the quantization noise variance of the system of Fig. 5 and is equal to $c2^{-2b}$. The postfilter $H'_{\rm opt}(e^{j\omega})$ of the new system is given by (2.3) with $\sigma_{q'}^2$ and $G'(e^{j\omega})$ replacing σ_q^2 and $G(e^{j\omega})$, respectively. Substituting with $\sigma_{q'}^2$ as in (2.10) and with $G'(e^{j\omega})$ as in (2.9), it is easy to see that

$$H'_{\rm opt}(e^{j\omega}) = \frac{1}{\alpha} H_{\rm opt}(e^{j\omega}) \tag{2.11}$$

The filtering scheme of Fig. 5 can be redrawn as in Fig. 6. Following the same type of reasoning as before, the MSE of the scheme of Fig. 6 can be thus expressed as

$$\mathcal{E}' = \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})(1 - H'_{\rm opt}(e^{j\omega})G'(e^{j\omega}))\frac{d\omega}{2\pi} \tag{2.12}$$

By substituting $G'(e^{j\omega})$ and $H'_{\rm opt}(e^{j\omega})$ in (2.12) we can immediately see that $\mathcal{E}' = \mathcal{E}$. □

As a consequence of the above analysis, the MSE expression reduces to the integral in (2.7).

B. The Optimum Prefilter

The goal now is to find $|G(e^{j\omega})|^2$ that minimizes the functional (2.7) under the integral constraint (2.8). Since the magnitude squared response is always a nonnegative function of ω, the optimum minimizing solution we seek must be nonnegative. This implicit condition is incorporated in the optimization problem as a *pointwise inequality* constraint. The next theorem gives an expression for the optimum magnitude squared response of the prefilter.

Theorem 2.3: The prefilter $|G_{\rm opt}(e^{j\omega})|^2$ that minimizes (2.7) under the constraint $\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 1$ must have a magnitude squared response $|G_{\rm opt}(e^{j\omega})|^2$ in the following form:

$$|G_{\rm opt}(e^{j\omega})|^2 = \max\left(0, \frac{1}{\sqrt{S_{xx}(e^{j\omega})}}\left(\frac{1 + c2^{-2b}}{\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}} - \frac{c2^{-2b}}{\sqrt{S_{xx}(e^{j\omega})}}\right)\right) \quad \forall \omega \in [-\pi, \pi]. \tag{2.13}$$

Proof: The minimization of the functional (2.7) under the integral constraint (2.8) and the positivity condition belongs to a class of calculus of variation problems known as isoperimetric problems [18], [19]. An outline of the major steps of the proof with the corresponding equations is given below. For more details, we refer the reader to Appendix A.

Step 1—Problem Set Up: We transform the above constrained problem into an unconstrained one by lumping the integrand of (2.8) to the integrand of (2.7) by a parameter $\lambda(\omega)$. This leads to the following equation:

$$\mathcal{E}_{\rm new} = \int_{-\pi}^{\pi}\left(\frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} + \lambda S_{xx}(e^{j\omega})|G(e^{j\omega})|^2\right)\frac{d\omega}{2\pi}. \tag{2.14}$$

The parameter $\lambda(\omega)$ takes care of the integral constraint (2.8) that is independent of frequency. We can therefore treat $\lambda(\omega)$ as a constant λ. This last statement can be indeed proved formally [See page 175 of [20]]. The optimum magnitude response we seek must obviously be positive over all frequencies. To incorporate this constraint in our problem, we introduce an unspecified parameter $\beta(\omega)$ and consider now the problem of minimizing

$$\int_{-\pi}^{\pi}\left(\frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} + \lambda S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + \beta(\omega)|G(e^{j\omega})|^2\right)\frac{d\omega}{2\pi}. \tag{2.15}$$

The value of the parameter $\beta(\omega)$ is set in a way that assures that the positivity constraint is never violated. We note that, unlike the parameter λ, $\beta(\omega)$ in this case takes care of a pointwise constraint. It must therefore be a function of ω.

Step 2—Necessary Conditions for an Extremum: The key necessary condition for a calculus of variation problem is the Euler–Lagrange equation. For this problem, this is equivalent to requiring $|G(e^{j\omega})|^2$ to satisfy the following equation at all frequencies:

$$\frac{\partial}{\partial|G(e^{j\omega})|^2}\left(\frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} + \lambda S_{xx}(e^{j\omega})|G(e^{j\omega})|^2\right) = -\beta(\omega). \tag{2.16}$$

Solving the above equation leads to (2.13).

Step 3—Sufficient Condition for an Extremum: The derivation in Step 2 indicates that any minimizing curve for (2.7) under the integral constraint (2.8) and the implicit positivity constraint **must** have a magnitude response (2.13). Using the convexity of functionals, we finally prove that the solution (2.13) is not only necessary but also sufficient for a minimizing extremum. □

It follows immediately from this last theorem that the optimum prefilter $G_{\rm opt}(e^{j\omega})$ is not unique since its phase response can be arbitrary set. This is not the case for the optimum postfilter $H_{\rm opt}(e^{j\omega})$. From (2.5), we observe that the MSE is minimized with respect to the phase response of the filters if the product $G_{\rm opt}(e^{j\omega})H_{\rm opt}(e^{j\omega})$ has zero phase. The phase response of $G_{\rm opt}(e^{j\omega})$ must, therefore, be the complementary phase of $H_{\rm opt}(e^{j\omega})$. We also note that whenever $|G_{\rm opt}(e^{j\omega})|^2 = 0$, (2.2) simplifies to $H_{\rm opt}(e^{j\omega}) = 0$ as well. Finally, for an intuitive interpretation of the above result, we can see, from (A.5) in Appendix A, that the magnitude response of the prefilter is set to zero at those frequencies where the noise variance $\sigma_q^2 = c2^{-2b}$ exceeds $\gamma^2 S_{xx}(e^{j\omega})$, γ being a constant defined as in (A.7). *It is therefore better not to transmit the signal at those frequencies where the noise level is higher (by a certain threshold) than the signal level.*

III. Further Analysis of the Optimum One-Channel System

A. The Coding Gain Expression

Assume that we quantize $x(n)$ directly with b bits. We denote the corresponding MSE by $\mathcal{E}_{\rm direct}$. We then use the

optimum pre- and postfilters around the quantizer. With the rate of the quantizer fixed to the same value b, we denote the MSE in this case by $\mathcal{E}_{\text{new}}$. The ratio $\mathcal{G}_{\text{opt}} \triangleq \mathcal{E}_{\text{direct}}/\mathcal{E}_{\text{new}}$ is called the coding gain of the new system and, as the name suggests, is a measure of the benefits provided by the pre/postfiltering operation. The coding gain expression for the system of Fig. 1 with the optimum pre- and postfilters is given in the following theorem.

Theorem 3.1: With the optimal choice of pre- and postfilters, the coding gain expression for the scheme of Fig. 1 is

$$\mathcal{G}_{\text{opt}} = (1 + c2^{-2b})\mathcal{G}_{\text{hw}} \tag{3.1}$$

as long as the right-hand side of (2.13) is nonnegative $\forall\ \omega$. Here $\mathcal{G}_{\text{hw}}$ is the coding gain of the half-whitening scheme and is given by

$$\frac{\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}\right)^2}. \tag{3.2}$$

Proof: Following the above definition, the coding gain of the system of Fig. 1 can be expressed as

$$\mathcal{G}_{\text{opt}} = \frac{c2^{-2b}\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}}{c2^{-2b}\int_{-\pi}^{\pi}\frac{S_{xx}(e^{j\omega})}{(S_{xx}(e^{j\omega})|G_{\text{opt}}(e^{j\omega})|^2 + c2^{-2b})}\frac{d\omega}{2\pi}} \tag{3.3}$$

assuming that the right-hand side of (2.13) is always positive. From (2.13), one can then write

$$|G_{\text{opt}}(e^{j\omega})|^2 S_{xx}(e^{j\omega}) + c2^{-2b} = \frac{(1 + c2^{-2b})\sqrt{S_{xx}(e^{j\omega})}}{\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}}. \tag{3.4}$$

Substituting (3.4) into (3.3) and simplifying, we obtain (3.1). □

In the case where the right-hand side of (2.13) is set to zero at certain frequencies, we obtain the following coding gain expression:

$$\mathcal{G}_{\text{opt}} = (1 + c2^{-2b})\frac{\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}}{\left[\left(\int_{\Omega}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}\right)^2 + \left(1 + \frac{1}{C2^{-2b}}\right)\int_{\Omega'} S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}\right]}$$

where Ω and Ω' are the set of frequencies over which the right-hand side of (2.13) is >0 and <0, respectively. We still expect the filtering scheme under study to outperform the half-whitening scheme in this case but it is not clear how one can compare analytically $\mathcal{G}_{\text{opt}}$ to $\mathcal{G}_{\text{hw}}$.

Example 1—White Input Still Produces Gain: In this example, we assume that the input $x(n)$ is a white process with variance equal to one. It can be verified for this case that, $|G(e^{j\omega})|^2 = 1 \quad \forall\omega$. This is consistent with our earlier observation about the prefilter, namely, that it exploits the spectral shape of the input. The postfilter $H(e^{j\omega})$ is a constant, independent of frequency. The coding gain of the half-whitening scheme is one since it depends only on the spectral shape of the input. However, the more general system still produces a coding gain $(1 + c2^{-2b})$. The gain results from the "Wiener filter part" of the postfilter and, consequently, from the resulting prefilter expression.

Remarks on The Coding Gain Expression

1) **The coding gain expression for low bit rates.** It is quite clear from Theorem 3.1 that the system of Fig. 1 will always outperform the half-whitening scheme as long as the right-hand side in (2.13) remains nonnegative for all frequencies ω. The difference in performance is basically a function of the probabilistic distribution of the quantizer input and more important of the bit rate. Two points are in order: First, the reader should keep in mind that as we quantize at lower bit rates, the quantizer assumptions made at the beginning of this section become inaccurate questioning therefore the validity of the previous analysis. Second, even if those assumptions hold, the excess gain obtained by using the more general scheme is not worth the extra complexity. For example, for a Gaussian input source $x(n)$ and assuming an optimum uniform scalar quantizer, the factor $1 + c2^{-2b}$ provides an extra gain of 0.48 dB at $b = 2$, 0.16 dB at $b = 3$ and 0.05 dB at $b = 4$.
2) **The coding gain expression for high bit rates.** By letting b go to infinity, one can easily check that *the right-hand side* of (2.13) becomes

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

and is positive $\forall\omega$. Therefore, the coding gain expression derived in Theorem 3.1 can be used and as b goes to infinity, $\mathcal{G}_{\text{opt}}$ becomes equal to $\mathcal{G}_{\text{hw}}$. At high bit rate $(b \geq 4)$, the half-whitening scheme is good enough. A similar observation was first mentioned by Goodman and Drouilhet [21]. Although the final conclusion is the same, there are main differences between their work and ours. First, Goodman and Drouillet did not derive any coding gain expression. It was quantitatively unclear how much we can benefit from using the more general system of Fig. 4. Second, whereas our system is a discrete time system, the system analyzed in [21] was continuous time pre- and postfilters surrounding a sampler and a quantizer. Moreover, Goodman and Drouillet assumed an additive white noise source model for the quantizer where the noise source is uniform and independent of the quantizer input and its statistics. Although this model is a valid one, we prefer to use the different noise model proposed in [10] by imposing the constraint (2.1) in the beginning of our study. Finally, Goodman and Drouillet replaced the sampler and the quantizer by an additive independent noise source. By doing so, the system becomes identical to the communication system analyzed by Costas [12]. The starting point of Goodman and Drouillet's correspondence is therefore Costas result. This is a different problem as we pointed out in the introduction of this paper. In our case, we cannot use Costas result directly. The use Theorem 2.2 is essential in our derivation and it is because of this theorem that the quantization problem under study becomes similar to a communication problem.

B. Analysis Under a Colored Quantization Noise Assumption

The previous analysis can be repeated assuming that the quantization noise is now colored. The noise power spectral density $S_{qq}(e^{j\omega})$ becomes a function of frequency. The remaining assumptions are kept the same. The optimum postfilter in this case can be easily rederived and is given by

$$H_{\text{opt}}(e^{j\omega}) = \frac{1}{G(e^{j\omega})} \cdot \frac{S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega}) + \frac{S_{qq}(e^{j\omega})}{|G(e^{j\omega})|^2}}. \tag{3.5}$$

The corresponding MSE expression can be found to be

$$\mathcal{E} = \int_{-\pi}^{\pi} \frac{S_{qq}(e^{j\omega}) S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + S_{qq}(e^{j\omega})} \frac{d\omega}{2\pi}. \tag{3.6}$$

We can again argue that the MSE at the output of the system does not change by inserting a multiplier before the quantizer. The same type of analysis can therefore be carried out producing the following expression for the magnitude response of the optimum prefilter:

$$|G_{\text{opt}}(e^{j\omega})|^2 = \max\left(0, \frac{1}{\sqrt{S_{xx}(e^{j\omega})}}\left(\frac{(1+c2^{-2b})S_{qq}(e^{j\omega})}{\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}} - \frac{S_{qq}(e^{j\omega})}{\sqrt{S_{xx}(e^{j\omega})}}\right)\right) \quad \forall \omega \in [-\pi, \pi]. \tag{3.7}$$

IV. Optimum Pre- and Postfiltering with First-Order Filters

The goal of this section is to try to mimic the same kind of analysis as before with finite-order filters. In specific, we will constrain $H(e^{j\omega})$ and $G(e^{j\omega})$ to be first-order causal filters with monic polynomials in the form $1-\alpha z^{-1}$ and $\frac{1}{1-\gamma z^{-1}}$. These first-order filters can provide substantial coding gain, are easy to track mathematically and are very economic to implement. The quantization system in Fig. 1 is still a uniform quantizer. We will again jointly optimize the first-order pre- and postfilters to minimize the MSE under the constraint (2.1). All the other assumptions of Section II are the same. We will consider two main cases: a) an FIR prefilter with an IIR postfilter and b) an IIR prefilter with an FIR postfilter. The choice of this combination is not as arbitrary as it may seem. The case where α is not equal to γ can be seen as the first-order "version" of the general system of Fig. 1 whereas the case of $\alpha = \gamma$ can be interpreted as the first-order "version" of the half-whitening scheme. In the case of ideal filters, interchanging $G(e^{j\omega})$ and $H(e^{j\omega})$ is merely a change of notation but when dealing with finite-order filters, the performance of an FIR prefilter (postfilter) is not in general similar to the performance of an IIR prefilter (postfilter). The two cases must be considered and different results can occur as we will observe through some examples.

A. The FIR Prefilter-IIR Postfilter Case

In this subsection, the prefilter is in the form $1-\alpha z^{-1}$. The postfilter takes the form $\frac{1}{1-\gamma z^{-1}}$. Under the constraint (2.1), the MSE expression is derived. It is a function of two variables α and γ and the goal is to jointly optimize these coefficients to minimize the MSE. The next theorem gives the expression of the MSE.

Theorem 4.1: Assume that the prefilter $G(e^{j\omega})$ is $1-\alpha e^{-j\omega}$ and that the postfilter $H(e^{j\omega})$ is $\frac{1}{1-\gamma e^{-j\omega}}$. The MSE as a function of α and γ, under the constraint (2.1), is given by

$$\mathcal{E}(\alpha,\gamma) = \frac{c2^{-2b}((1+\alpha^2)R_{xx}(0) - 2\alpha R_{xx}(1))}{(1-\gamma^2)} + \frac{(\alpha-\gamma)^2}{(1-\gamma^2)}\left(R_{xx}(0) + 2\sum_{m=1}^{\infty}\gamma^m R_{xx}(m)\right). \tag{4.1}$$

Proof: See Appendix B. □

By using $\sigma_q^2 = c2^{-2b}\sigma_y^2$ where $\sigma_y^2 = (1+\alpha^2)R_{xx}(0) - 2\alpha R_{xx}(1)$, the MSE expression of Theorem 4.1 can be rewritten as follows:

$$\mathcal{E}(\alpha,\gamma) = \frac{\sigma_q^2}{(1-\gamma^2)} + \frac{(\alpha-\gamma)^2}{(1-\gamma^2)} \times \left(R_{xx}(0) + 2\sum_{m=1}^{\infty}\gamma^m R_{xx}(m)\right) \tag{4.2}$$

The first term in (4.2) disappears when we do not quantize the signal. In this case, the MSE can be reduced to zero by setting α equal to γ, i.e., the postfilter is the inverse of the prefilter. However, in the presence of the quantizer, the choice of $\alpha = \gamma$ is not the best since the choice of γ affects the two terms in (4.2) in different ways. Equation (4.2) also suggests that, at high bit rate, the contribution of the first term in the equation will be almost negligible compared to the second term. Hence, as b increases, we should expect the optimum coefficients α_{opt} and γ_{opt} to numerically approach each other.

Even in this very simple case, the problem is highly nonlinear in the filter coefficients α and γ. Closed-form expressions for the coefficients of the filters in terms of only the second-order statistics of the signal cannot be obtained. However, minimization of the MSE can be done numerically using for example MATLAB's optimization toolbox.

B. The IIR Prefilter-FIR Postfilter Case

We can easily derive, from (4.2), the MSE for the dual case, namely when the prefilter is $\frac{1}{1-\gamma z^{-1}}$ and the postfilter is $1-\alpha z^{-1}$. To see this, assume first that there is no quantization. It is then clear that the second term in (4.2), the error due to the mismatch of the coefficients, will not change by switching the position of the filters. When quantization is present, the noise term becomes $(1+\alpha^2)\sigma_q^2$ where the noise variance $\sigma_q^2 = c2^{-2b}\sigma_y^2 = c2^{-2b}\frac{1}{1-\gamma^2}(R_{xx}(0) + 2\sum_{m=1}^{\infty}\gamma^m R_{xx}(m))$. The MSE expression is therefore given by

$$\mathcal{E}(\alpha,\gamma) = c2^{-2b}\frac{(1+\alpha^2)}{(1-\gamma^2)}\left(R_{xx}(0) + 2\sum_{m=1}^{\infty}\gamma^m R_{xx}(m)\right) + \frac{(\alpha-\gamma)^2}{(1-\gamma^2)}\left(R_{xx}(0) + 2\sum_{m=1}^{\infty}\gamma^m R_{xx}(m)\right). \tag{4.3}$$

C. The Special Case of First-Order Filters with Equal Coefficients

The FIR Prefilter-IIR Postfilter Case: When α is equal to γ, the MSE becomes a function of one parameter α. The coding gain can be then expressed as follows:

$$\mathcal{G}_{\text{opt}} = \frac{R_{xx}(0)(1-\alpha_{\text{opt}}^2)}{(1+\alpha_{\text{opt}}^2)R_{xx}(0) - 2\alpha_{\text{opt}}R_{xx}(1)}. \tag{4.4}$$

If $R_{xx}(1) = 0$, then, the above coding gain expression becomes $\frac{(1-\alpha^2{}_{\text{opt}})}{(1+\alpha^2{}_{\text{opt}})}$. It is then quite clear that the optimum coefficient α_{opt} is equal to zero. No pre- and postfiltering can enhance the reconstructed output and the coding gain is simply unity. On the other hand, If $R_{xx}(1) = R_{xx}(0)$, then, the coding gain expression becomes $\frac{(1+\alpha_{\text{opt}})}{(1-\alpha_{\text{opt}})}$. As α_{opt} approaches 1, the coding gain grows unbounded. The tradeoff is the stability of the inverse filter.

Having taken care of these two extreme cases, we now assume that $0 < |R_{xx}(1)| < R_{xx}(0)$ and introduce the following notation: $\frac{R_{xx}(1)}{R_{xx}(0)} \triangleq \rho$ where $-1 < \rho < 1$. The problem expressed in this form was considered by Jayant and Noll [10]. We will therefore only give their final results.

1) The optimum coefficient α_{opt} that minimizes the MSE expression is given by

$$\alpha_{\text{opt}} = \frac{1}{\rho}\left(1 - \sqrt{1-\rho^2}\right) \quad \text{if } -1 < \rho < 1. \tag{4.5}$$

2) The coding gain expression as a function of ρ can be found to be

$$\mathcal{G}_{\text{opt}} = \frac{1}{\sqrt{1-\rho^2}}. \tag{4.6}$$

We note that the coding gain expression (4.6) is also the coding gain of a two-channel Karhunen–Loeve transform (KLT) under the assumption of optimum bit allocation. This is then a case of a one channel biorthogonal FB that is as good as a 2×2 KLT [an example of a two channel orthonormal filter bank]. This is interesting in view of the fact that the asymptotic coding gain of a KLT is higher than that of a half-whitening filter. A natural question then arises: how does the coding gain of a KLT of block length $(M+1)$ compare to the coding gain of a half-whitening-like scheme using filters $A(z)$ and $1/A(z)$ of order M? The coding gain of a KLT of block length $(M+1)$ is well established [10]. For the half whitening like scheme, since the postfilter is assumed to be the inverse of the prefilter $A(z)$, the MSE expression is due only to the noise component and can be expressed as follows:

$$\mathcal{E} = c2^{-2b}(a^T\mathbf{R}a)\int_{-\pi}^{\pi}\frac{1}{|A(e^{j\omega})|^2}\frac{d\omega}{2\pi}$$

where $a^T = (1 \;\; a_1 \;\; \cdots \;\; a_{M-1})$, $A(z) = 1 + a_1z^{-1} + \cdots + a_{M-1}z^{-(M-1)}$ and $\mathbf{R}$ is the $(M+1)\times(M+1)$ autocorrelation matrix. The integral in the above expression has a well known closed-form expression in terms of the reflection coefficients k_i (See for example [10], [22]). The following closed-form expression for the coding gain can be therefore obtained

$$\mathcal{G}_{\text{opt}} = \frac{R_{xx}(0)\prod_{i=1}^{M}(1-k_i^2)}{a^T\mathbf{R}a}. \tag{4.7}$$

The reflection coefficients are related in a nonlinear fashion to the coefficients of the filter $A(z)$ [17]. For the first-order case, k_1 is equal to a_1 and (4.7) simplifies to (4.4). The maximization of (4.7) is however beyond the scope of this paper.

The IIR Prefilter-FIR Postfilter Case: When α is equal to γ, the MSE is then given by

$$c2^{-2b}\frac{(1+\alpha^2)}{(1-\alpha^2)}\left(R_{xx}(0) + 2\sum_{m=1}^{\infty}\alpha^m R_{xx}(m)\right). \tag{4.8}$$

In this case, the problem is highly nonlinear in the filter coefficient α and an analytical solution is difficult to obtain. On the other hand, the minimization of the MSE can be easily done numerically. Results are illustrated in the next subsection for some specific examples.

D. Examples of Optimum Filters for Specific Inputs

The examples given in this subsection correspond, respectively, to the cases of a MA(1), an AR(1) and an AR(5) input process $x(n)$. In each case, we compare the coding gain of the general first-order system $[\alpha \neq \gamma]$ to the coding gain of the first-order system with α equal to γ at various bit rates. The optimization of the coefficients is done numerically using MATLAB's optimization toolbox whenever an analytical expression is difficult to obtain. We also include in our comparison the half-whitening coding gain $\mathcal{G}_{\text{hw}}$ and the coding gain of the system of Fig. 1, $\mathcal{G}_{\text{opt}}$. $\mathcal{G}_{\text{hw}}$ establishes a theoretical bound on the coding gain of the first-order system with α equal to γ whereas $\mathcal{G}_{\text{opt}}$ represents the theoretical bound for the more general system $[\alpha \neq \gamma]$.

Example 2—Case of a MA(1) Process: Assume that the input $x(n)$ is a zero mean Gaussian MA(1) process with an autocorrelation sequence in the form

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

It is well known that a MA(1) process has to have $R_{xx}(1)/R_{xx}(0) \leq 1/2$ to ensure that the power spectral density is indeed nonnegative. We therefore restrict θ to be between -1 and 1. For the FIR prefilter-IIR postfilter case, when α is equal to γ, the ratio $R_{xx}(1)/R_{xx}(0)$ is now equal to $\theta/(1+\theta^2)$. We therefore simply replace ρ in (4.5) and (4.6) by $\theta/(1+\theta^2)$ to obtain expressions for the optimum coefficient α_{opt} and the optimum coding gain $\mathcal{G}_{\text{opt}}$. The power spectrum of the MA(1) process is given by

$$S_{xx}(e^{j\omega}) = 1 - 2\frac{\theta}{(1+\theta^2)}\cos(\omega) \tag{4.9}$$

Substituting (4.9) in (3.2), the coding gain expression of the half-whitening scheme for a $MA(1)$ process is given by

$$\mathcal{G}_{\text{hw}} = \frac{(1+\theta^2)}{\left(\int_{-\pi}^{\pi} \sqrt{(1+\theta^2-2\theta\cos(\omega))}\frac{d\omega}{2\pi}\right)^2}. \tag{4.10}$$

The integral in (4.10) is equal to $F(-0.5,-0.5;1;\theta^2)$ where $F(a,b;c;d)$ is Gauss' hypergeometric function. From [23], $F(-0.5,-0.5;1;\theta^2)$ can be rewritten as $(1+\theta)F(-0.5,0.5;1;4\theta/(1+\theta)^2)$. This, in turn, can be simplified to $(1+\theta)\frac{2}{\pi}E(2\sqrt{(|\theta|)}/(1+\theta))$ where $E(\cdot)$ is the complete elliptic integral of the second kind. Finally, $\mathcal{G}_{\text{opt}}$ is given by (3.1).

The optimization of the coefficients for the FIR prefilter-IIR postfilter case and the IIR prefilter-FIR postfilter with $\alpha \neq \gamma$ were all done numerically using MATLAB's optimization toolbox routine "fmins.m." The plots of the coding gain for the FIR/IIR case are illustrated in Figs. 7 and 8 for b equal to 2 and 3, respectively. Similarly, the plots of the coding gain for the IIR/FIR case are shown in Figs. 9 and 10. The dotted curve is the coding gain obtained by the first-order equal coefficients scheme whereas the dashed curve is the coding gain of the unequal coefficients case. Also included is the half-whitening coding gain, $\mathcal{G}_{\text{hw}}$, denoted by the dash-dot curve and the coding gain of the system of Fig. 1, $\mathcal{G}_{\text{opt}}$, denoted by the solid line curve. From these figures, we can observe that as the bit rate increases, there is no loss of generality in assuming α to be equal to γ.

Example 3–Case of an AR(1) Process: Assume that the input $x(n)$ is a zero mean Gaussian AR(1) process with an autocorrelation sequence in the form $R_{xx}(k) = \rho^{|k|}$ where ρ must be between -1 and 1.

For the FIR prefilter-IIR postfilter case, when α is equal to γ, the ratio $R_{xx}(1)/R_{xx}(0)$ is equal to ρ. α_{opt} is therefore given by (4.5) and the coding gain $\mathcal{G}_{\text{opt}}$ is given by (4.6). The power spectrum of the AR(1) process is

$$S_{xx}(e^{j\omega}) = \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\omega)}. \tag{4.11}$$

Substituting (4.11) in (3.2), the half-whitening coding gain expression for the AR(1) process is as follows:

$$\mathcal{G}_{\text{hw}} = \frac{1}{(1-\rho^2)\left(\int_{-\pi}^{\pi} \frac{1}{\sqrt{(1+\rho^2-2\rho\cos(\omega))}}\frac{d\omega}{2\pi}\right)^2}. \tag{4.12}$$

The integral in (4.12) is equal to $\frac{2}{\pi}K(\rho)$ where $K(\rho)$ is the complete elliptic integral of the first kind [23]. The coding gain of the system of Fig. 1, $\mathcal{G}_{\text{opt}}$, is again given by (3.1). The optimization of the coefficients for the FIR prefilter-IIR postfilter case and the IIR prefilter-FIR postfilter with $\alpha \neq \gamma$ were all done numerically using the same MATLAB's optimization toolbox routine "fmins.m." The plots of the coding gain are illustrated in Figs. 11 and 12 for the FIR/IIR case and in Figs. 13 and 14 for the IIR/FIR case as the bit rate b varies from 2 to 3. The same curve notation as in the previous MA(1) example is used and the same conclusion can be reached.

TABLE I
THE CODING GAIN OBTAINED FROM FIRST-ORDER FILTERS FOR THE $AR(5)$ CASE OF EXAMPLE 4.3.3

	b = 1	b = 2	b = 3
FIR/IIR $\alpha \neq \gamma$	3.05	2.96	2.94
FIR/IIR $\alpha = \gamma$	2.92	2.92	2.92
IIR/FIR $\alpha \neq \gamma$	3.76	3.52	3.42
IIR/FIR $\alpha = \gamma$	3.37	3.37	3.37

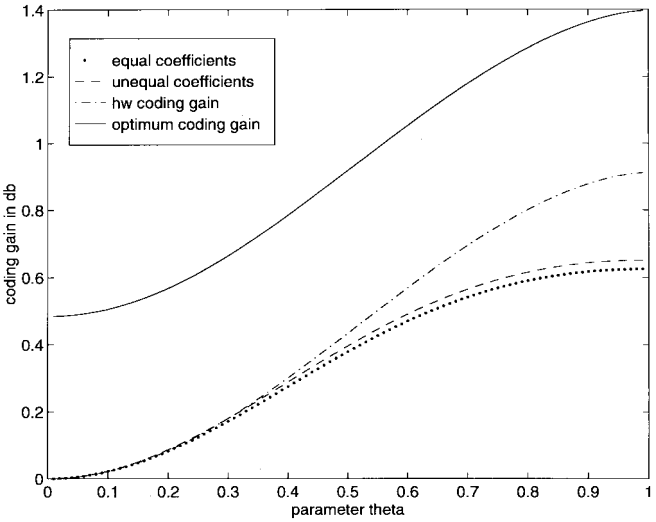

Fig. 7. Coding gain curves for the MA(1) case: FIR prefilter, IIR postfilter and $b = 2$.

Example 4—Case of an $AR(5)$ Process: The autocorrelation function of such a process extends to infinity and doesn't have a simple closed-form expression. The main problem is the infinite summation in the form $\sum_{m=1}^{\infty}\gamma^m R_{xx}(m)$ found in (4.1), (4.3) with $\alpha \neq \gamma$ and (4.8) with $\gamma = \alpha$. Our approach is to truncate this infinite summation with the assumption that after a certain lag m, the correlation coefficients are negligible. For this AR(5) process, we set $R(0) = 1$, $R(1) = 0.86$, $R(2) = 0.64$, $R(3) = 0.4$, $R(4) = 0.26$, $R(5) = 0.2$ and $R(m) = 0, \quad \forall m \geq 6$. The values of the correlation coefficients are obtained from [10, p. 37]. Table I summarizes our coding gain results in decibels 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 as we have mentioned previously.

V. REPLACING THE QUANTIZER SYSTEM BY AN ORTHONORMAL UNIFORM PRFB

Consider the M channel maximally decimated uniform SBC of Fig. 2. The boxes labeled $\mathcal{Q}$ are modeled by additive noise sources in the manner described in the introduction. Throughout this section, we will assume that the subband

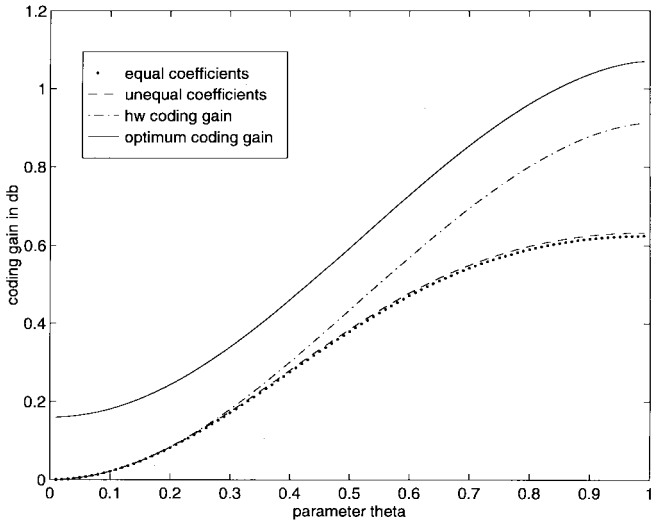

Fig. 8. Coding gain curves for the MA(1) case: FIR prefilter, IIR postfilter and $b = 3$.

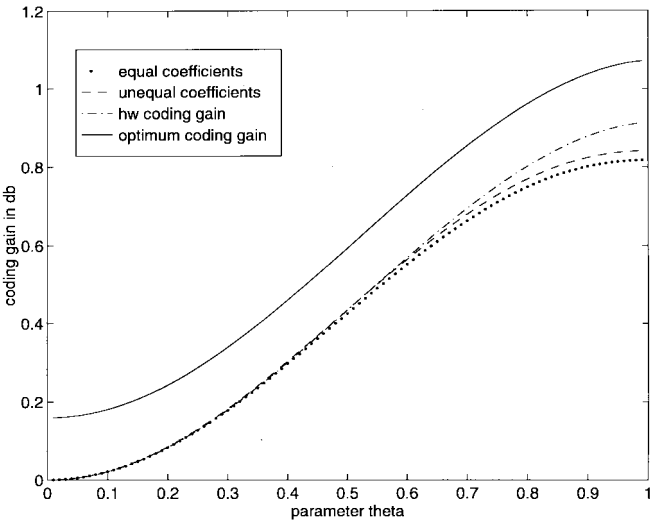

Fig. 10. Coding gain curves for the MA(1) case: IIR prefilter, FIR postfilter and $b = 3$.

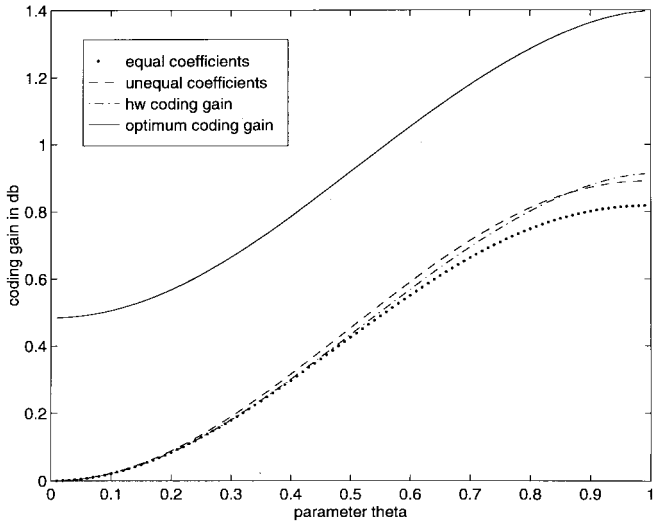

Fig. 9. Coding gain curves for the MA(1) case: IIR prefilter, FIR postfilter and $b = 2$.

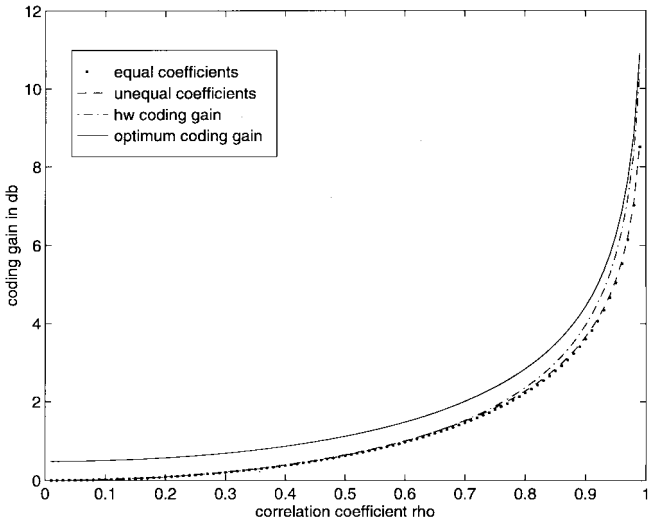

Fig. 11. Coding gain curves for the AR(1) case: FIR prefilter, IIR postfilter and $b = 2$.

quantization noise sources are white and pairwise uncorrelated. If we interpret this FB as a sophisticated quantizer, the use of pre- and postfilters around the FB can very well increase the coding gain. In a recent paper, Djokovic and Vaidyanathan [16] analyzed the system of Fig. 1 where the quantization system $\mathcal{QS}$ is a uniform orthonormal PRFB and the postfilter is the inverse of the prefilter. The authors gave a formula for the optimum allocation of bits in the subbands. Furthermore, they showed that minimizing the MSE of the so called prefiltered paraunitary (PPU) PRFB can be done by separately optimizing the pre/postfiltering scheme and the orthonormal filter bank. Their proposed solution was a half-whitening scheme surrounding an optimum orthonormal PRFB. A generalization of the scheme of Djokovic and Vaidyanathan would be again to relax the assumption that the postfilter is the inverse of the prefilter. An analytical *optimum* solution, if it exists, must incorporate the joint optimization of the orthonormal PRFB and the pre- and postfilters. It is not clear that a separate optimization of the pre- and postfilters and the orthonormal PRFB still holds in this case. Furthermore, any developed optimum bit allocation formula must include the pre- and postfiltering operation.

In the remainder of this section, we will not attempt to solve the problem described above. We will instead provide a *suboptimum* procedure that relies on the results derived in Section II. We will see that even in this simpler case, two theorems must be first established. The first step in the procedure is to optimize the orthonormal uniform PRFB for a certain WSS input $x(n)$. Vaidyanathan has recently shown [9] that the optimum orthonormal uniform PRFB, the one that maximizes the coding gain as defined in Section III, will consist of antialias filters. A discrete time filter is said to be an antialias(M) filter if its output can be decimated M-fold without aliasing. Since this requires infinite attenuation in the stopbands, antialias filters are therefore a class of ideal filters. The second step in the procedure is to perform the

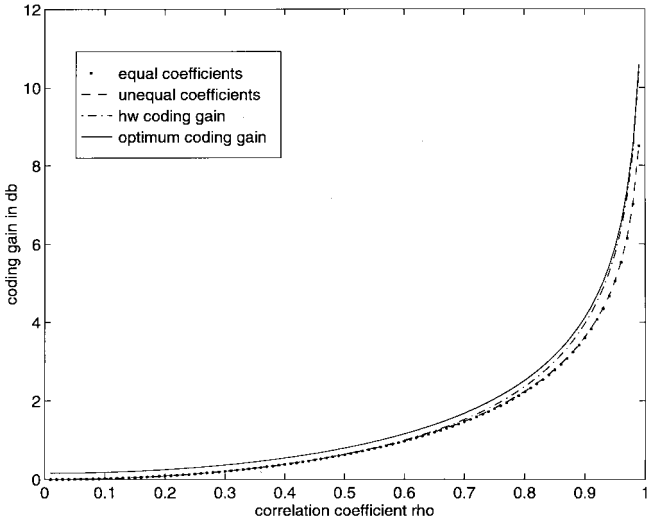

Fig. 12. Coding gain curves for the AR(1) case: FIR prefilter, IIR postfilter and $b = 3$.

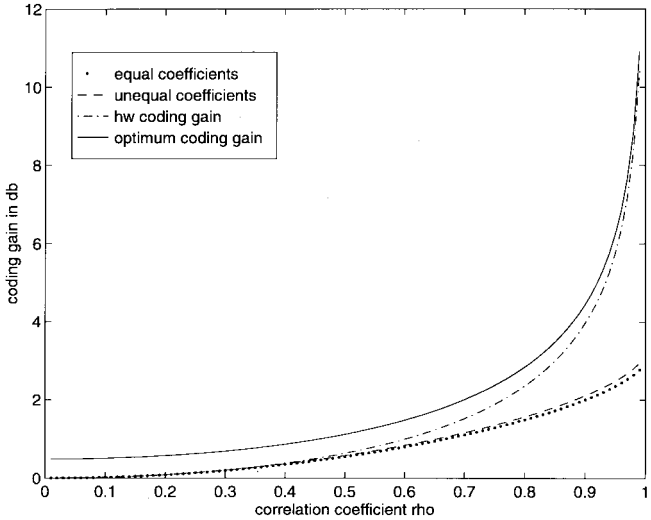

Fig. 13. Coding gain curves for the AR(1) case: IIR prefilter, FIR postfilter and $b = 2$.

optimum bit allocation operation in the usual way [24]. After optimally allocating the bits, we would like to apply the pre- and postfilters derived previously in Section II. In order to do this, we need first to replace the whole optimum orthonormal PRFB by an additive noise source, say $v(n)$. This noise source $v(n)$ must be WSS and uncorrelated with the prefilter output $y(n)$. Second, the variance of the input process σ_x^2 must be related to the PRFB output noise variance $\mathcal{E}_{\mathrm{SBC}}$ in a similar fashion as in (2.1). A major problem is the following: In the presence of quantizers, it is well known that the output of a uniform/nonuniform PRFB is in general a cyclo-wide sense stationary (CWSS) process. The cyclo-wide sense stationarity is due to the passage of the quantization noise through the interpolators [25]. We provide two results describing important cases that guarantee the wide sense stationarity of the quantization noise of a uniform **orthonormal** PRFB. Since the results hold for the nonuniform decimation case, the proofs will assume a nonuniform maximally decimated orthonormal

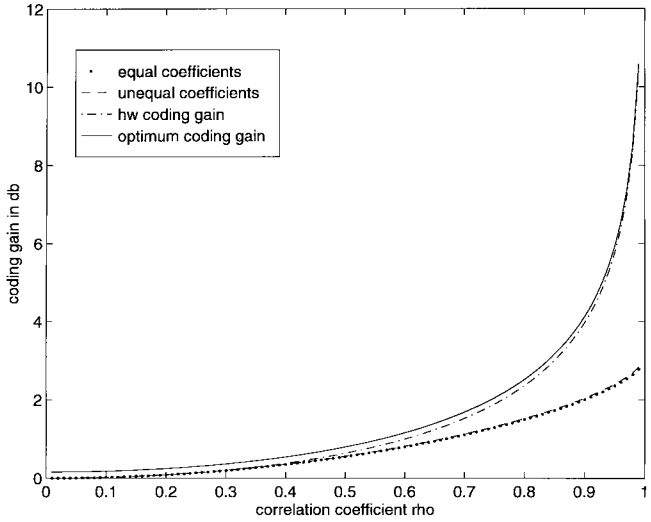

Fig. 14. Coding gain curves for the AR(1) case: IIR prefilter, FIR postfilter and $b = 3$.

PRFB case. A nonuniform SBC, shown in Fig. 15, is a SBC with unequal subband decimation ratios n_k. The boxes labeled $\mathcal{Q}$ represent, as before, uniform quantizers that are modeled by additive noise sources.

Theorem 5.1: Under optimum bit allocation, the output noise of a [possibly nonuniform] orthonormal PRFB is WSS provided the subband quantization noise sources are white, uncorrelated and zero mean (WUZE assumptions).

Proof: The proof is now established through the following series of steps:

1) Soman and Vaidyanathan [24] showed that for a nonuniform orthonormal PRFB, the variances of the subband quantization noises should be equal under optimum bit allocation. Because we are assuming optimum bit allocation in our theorem, we can immediately conclude that the noise variances in the nonuniform orthonormal PRFB should be equal to each other.
2) It is well known [26]–[28] that an M-channel nonuniform FB can be redrawn as an L-channel maximally decimated uniform FB where $L = n_k p_k$. The set of M analysis and synthesis filters $\{H_k(z), F_k(z)\}$ are replaced by the set of L filters $\{H'_k(z), F'_k(z)\}$ in the uniform system where $L \geq M$. The main goal at this point is to develop the form of the power spectral density matrix of the subband quantization noise $\mathbf{S_{qq}}(e^{j\omega})$ in the equivalent L channel maximally decimated uniform FB. We first observe that the white noise assumption guarantees that, for the kth channel, the quantization noises in its corresponding p_k channels are uncorrelated. Furthermore, the variance of the quantization noise is the same in all the p_k channels. Combining this observation with the conclusion of step 1, it is easy to see that $\mathbf{S_{qq}}(e^{j\omega})$ should be equal to $\sigma_q^2 I$ where $\mathbf{S_{qq}}(e^{j\omega})$ is an $L \times L$ matrix.
3) Since the nonuniform maximally decimated FB is orthonormal and exhibits the perfect reconstruction (PR) property, then, it follows that the analysis polyphase matrix $\mathbf{E}'(e^{j\omega})$ of the equivalent L channel uniform FB

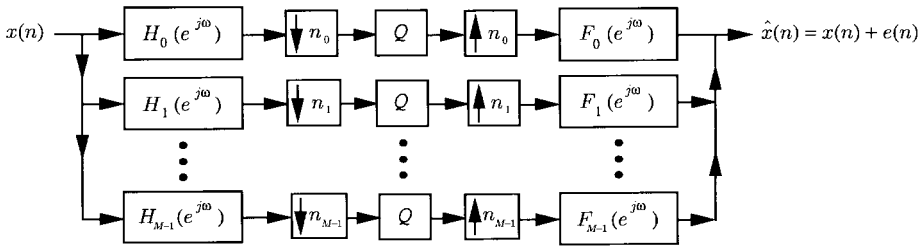

Fig. 15. An M-channel nonuniform subband coder (SBC).

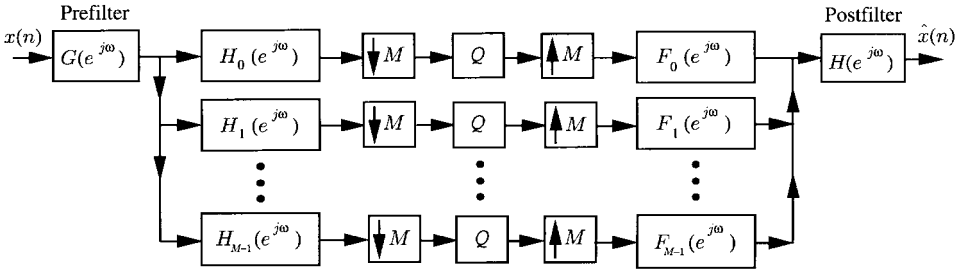

Fig. 16. The optimum uniform orthonormal FB with the general pre- and postfiltering scheme.

is lossless, i.e., $\mathbf{E}'(e^{j\omega})\mathbf{E}'^{\dagger}(e^{j\omega}) = \mathbf{I}$ (orthonormality) and the synthesis polyphase matrix $\mathbf{R}'(e^{j\omega})$ of the equivalent L channel uniform FB is equal to $\mathbf{E}'^{\dagger}(e^{j\omega})$ (PR property) [27]. The power spectral density matrix of the output quantization noise $\mathbf{S_{vv}}(e^{j\omega})$ is equal to $\mathbf{R}'(e^{j\omega})\mathbf{S_{qq}}(e^{j\omega})\mathbf{R}'^{\dagger}(e^{j\omega})$ which can be evaluated as $\sigma_q^2 I$ using the above properties. This means that the output noise $v(n)$ is an interleaved version of L uncorrelated white noise sources, each of variance σ_q^2. So, $v(n)$ itself is white with variance σ_q^2. □

Since the above theorem holds for a nonuniform maximally decimated orthonormal PRFB, it includes the uniform decimation case. The output quantization noise $v(n)$ in Theorem 5.1 is white with variance σ_q^2. Furthermore, $v(n)$ and $y(n)$ are uncorrelated. The problem with the optimum bit allocation is that it yields noninteger solution for the bits. If we use a simple rounding procedure or a more sophisticated algorithm [29] to obtain integer solutions, the assumption of equal quantizer noise variances is not valid any more. Nevertheless, in the next theorem, we prove that even with the more practical assumption of different quantization noise variances, the output of a nonuniform orthonormal PRFB with **antialias** filters will be wide sense stationary.

Theorem 5.2: The output noise of a [possibly nonuniform] orthonormal PRFB consisting of **antialias** filters is WSS provided the subband quantization noise sources are zero mean and pairwise uncorrelated.

Proof: Consider the synthesis bank of a nonuniform PRFB. The quantization noise sources $q_k(n)$ at the input of the interpolators are assumed to be WSS with power spectrum $S_{q_k}(e^{j\omega})$ and are pairwise uncorrelated. Since the filters $F_k(e^{j\omega})$ are antialias for all k, then, each upsampled and filtered noise sequence $v_k(n)$ is WSS [25]. Furthermore, since the interpolated noise sequences $v_k(n)$ are linear combinations of the input noise sources $q_k(n)$, the uncorrelatedness property is preserved. This can be verified by writing the output vector $\mathbf{v}(n)$ as a time varying linear combination of the vector $\mathbf{q}(n)$ and taking expectations. The interpolated noise sources $v_k(n)$ are therefore jointly wide sense stationary which implies that their sum $v(n)$ is WSS. □

We emphasize the fact that neither *the whiteness of the noise sources nor the equal variance* assumptions are required for this theorem to hold. We note that the output quantization noise $v(n)$ in Theorem 5.2 is still uncorrelated with the prefilter output $y(n)$. However, in this case, $v(n)$ is not white. If the subband quantization noise sources are white, it is easy to see that the power spectral density $S_{vv}(e^{j\omega})$ of the PRFB output noise is piecewise constant. The magnitude of each piece of $S_{vv}(e^{j\omega})$ is equal to $\sigma_{q_k}^2$ for some k. The location of the constant piece is determined by the passband of the corresponding synthesis filter $F_k(e^{j\omega})$. The variance of the output noise $v(n)$ is the average of the individual noise variances $\frac{1}{M}\sum_{k=0}^{M-1}\sigma_{q_k}^2$.

The above two theorems permits the continuation of our suboptimum procedure. The optimum bit allocation [without including the pre- and postfilters] allows us to relate the variance of the input process σ_x^2 to the FB output noise variance $\mathcal{E}_{\mathrm{SBC}}$ by $\mathcal{E}_{\mathrm{SBC}} = \frac{c2^{-2b}}{\mathcal{G}_{\mathrm{PU}}}\sigma_x^2$. The optimum orthonormal FB is a special case of a nonuniform PRFB with antialias filters for which Theorem 5.2 applies. The FB can be therefore modeled as a WSS noise source that is uncorrelated with the prefilter output sequence $y(n)$ and has a variance proportional to σ_y^2. This is the perfect setting for our previous pre- and postfiltering analysis. The complete system is shown in Fig. 16. The expressions for the optimum postfilter and the magnitude response of the optimum prefilter are given, respectively, by (2.3) and (2.13) if the noise $v(n)$ is white [case of Theorem 5.1] or by (3.5) and (3.7) if the noise $v(n)$ is colored [case of Theorem 5.2]. For either cases, the coding gain of the system of Fig. 16 can be easily obtained as $(1 + \frac{c2^{-2b}}{\mathcal{G}_{\mathrm{PU}}})\mathcal{G}_{\mathrm{lw}}\mathcal{G}_{\mathrm{PU}}$ provided the right-hand side of (2.13) or (3.7) is always positive. The next example illustrates the above procedure and provide some numerical results.

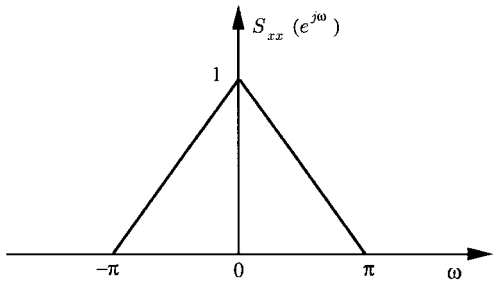

Fig. 17. The power spectral density for the input of Example 5.1.

Example 5: We assume that the input $x(n)$ is a zero mean, real, WSS random process with a triangular power spectral density $S_{xx}(e^{j\omega})$ as shown in Fig. 17. The optimum orthonormal FB in this case is the well-known contiguous ideal brick wall FB [6]–[9]. The coding gain of an orthonormal FB after the optimum allocation of subband bits is in general given by [24]

$$\mathcal{G}_{\rm PU} = \frac{\sigma_x^2}{\left(\prod_{k=0}^{M-1}\sigma_{x_k}^2\right)^{1/M}}$$

where σ_x^2 is the variance of $x(n)$ and $\sigma_{x_k}^2$ is the variance of the kth subband signal. For the ideal brick wall FB and a triangular power spectral density, the above coding gain expression can be simplified to the following expression:

$$\mathcal{G}_{\rm PU} = \frac{0.5}{\left(\prod_{k=0}^{M-1}(1-(2k+1)/2M)\right)^{1/M}}$$

We then apply the optimum pre- and postfilters at the input and output of the FB, respectively. For an average bit rate $b = 3$, the constant $c = 0.75$ and the number of channels $M = 2$, it can be verified that the optimum prefilter [in both cases of white and colored noise] is never set to zero at any frequency and, therefore, we can use the formula $\mathcal{G}_{\rm opt} = (1 + \frac{c^{-2b}}{\mathcal{G}_{\rm PU}})\mathcal{G}_{\rm hw}\mathcal{G}_{\rm PU}$. Using the above data, we obtain $\mathcal{G}_{\rm PU} = \frac{2}{\sqrt{3}}$ and $\mathcal{G}_{\rm hw} = \frac{9}{8}$. Finally, the theoretical bound on the coding gain, namely the prediction gain, is given by [1]:

$$\mathcal{G}_{\rm th} = \frac{\sigma_x^2}{\exp\left\{\int_{-\pi}^{\pi}\ln(S_{xx}(e^{j\omega}))\frac{d\omega}{2\pi}\right\}}$$

For this case, $\mathcal{G}_{\rm th}$ is equal to $\frac{e}{2}$. Expressing the above quantities in decibels, we get $\mathcal{G}_{\rm PU} = 0.625$ dB, $\mathcal{G}_{\rm hw} = 0.51$ dB, $\mathcal{G}_{\rm opt} = 1.19$ dB, and $\mathcal{G}_{\rm th} = 1.33$ dB. It is important to observe the relative gain obtained using the pre- and postfiltering operation rather than the absolute value of the coding gain. Clearly, we get a substantial increase by using the pre- and postfilters as $\mathcal{G}_{\rm opt}$ approaches $\mathcal{G}_{\rm th}$.

VI. Concluding Remarks

In this paper, we have studied pre- and postfiltering around a quantization system $\mathcal{QS}$ under the key assumption that the quantization noise variance σ_q^2 is proportional to the variance of the quantization system input. For the case where $\mathcal{QS}$ is a uniform quantizer, we provided joint optimum solutions for the ideal pre- and postfilters. Using these solutions, we then derived a coding gain expression for the system of Fig. 1. The importance of this expression is that it clearly indicates that, at high bit rate, there is no substantial loss of coding gain if we set the postfilter to be the inverse of the prefilter. We then considered two cases of first-order pre- and postfilters: FIR/IIR and IIR/FIR. For each case, we obtained a mean square expression, optimized the coefficients α and γ and compared coding gain performances with the case of α equal to γ through a set of examples. Finally, we considered the case where the quantization system $\mathcal{QS}$ is an orthonormal FB. To be able to apply the previously optimized pre- and postfilters at the input and output of the FB, respectively, we developed two theorems that guaranteed the wide sense stationarity of the filter bank output. We emphasize again that applying the pre- and postfilters in the manner described in Section V is suboptimum. While this paper deals with some of the current issues in the subband coding area, it opens up other interesting and challenging problems. One example that quickly comes to mind is the extension of this work to the M channel case. A globally optimum solution should include a strategy for the allocation of the subband bits as well as a joint optimization of the analysis and synthesis sections of the SBC. Another problem is the *optimum* generalization of Djokovic and Vaidyanathan's scheme. If fully optimized, the more general scheme of Fig. 16 should outperform the scheme proposed by Djokovic and Vaidyanathan. An easy way to see this is to simply put a Wiener filter at the output of the half-whitening filter surrounding an optimum orthonormal PRFB. Even in this suboptimum procedure, the mean square reconstruction error cannot increase.

Appendix A

Step 1: We have argued in the proof of Theorem 2.3 that the parameter $\lambda(\omega)$ is independent of frequency. We proceed to prove that it is a positive constant. Assume for the moment that $\beta(\omega)$ is equal to zero in (2.16) and denote the integrand of (2.16) by $F + \lambda W$ where

$$F = \frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}}$$

and

$$W = S_{xx}(e^{j\omega})|G(e^{j\omega})|^2$$

From the theorem on [18, p. 43], we see that if $|G(e^{j\omega})|^2$ is an extremum of (2.16) with $\beta(\omega) = 0$ (but is not in the same time an extremum of W), then, there exists a constant parameter λ such that $\partial F/\partial|G(e^{j\omega})|^2 + \partial W/\partial|G(e^{j\omega})|^2 = 0$ for all ω. Since $|G(e^{j\omega})|^2$ is not an extremal for W, then there is a ω_o such that $\partial W/\partial|G(e^{j\omega})|^2 \neq 0$ at $\omega = \omega_o$. This yields

$$\lambda = -\left.\frac{\partial F/\partial|G(e^{j\omega})|^2}{\partial W/\partial|G(e^{j\omega})|^2}\right|_{\omega=\omega_o}. \tag{A.1}$$

The numerator and denominator of (A.1) are found to be

$$\frac{\partial F}{\partial|G(e^{j\omega})|^2} = -\frac{c2^{-2b}S_{xx}(e^{j\omega})^2}{\left(S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}\right)^2}$$

and

$$\frac{\partial W}{\partial |G(e^{j\omega})|^2} = S_{xx}(e^{j\omega}). \tag{A.2}$$

Substituting (A.2) into (A.1), we obtain the following:

$$\lambda = \frac{c2^{-2b} S_{xx}(e^{j\omega})}{\left(S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}\right)^2} \tag{A.3}$$

which in particular shows that $\lambda > 0$.

Step 2: The necessary conditions for $|G(e^{j\omega})|^2$ to be a minimum of (2.15) are summarized next.

1) Define G' to be the derivative of $|G(e^{j\omega})|^2$ with respect to ω. The Legendre condition [18],

$$\frac{\partial^2}{\partial G'^2}(F + \lambda W) \geq 0 \quad \forall\, \omega$$

must be satisfied. In our case, this condition is satisfied trivially because neither the functional (2.7) nor the constraint (2.8) are functions of the derivative of $|G(e^{j\omega})|^2$.

2) $|G(e^{j\omega})|^2$ must satisfy the Euler–Lagrange equation for the functional (2.15) i.e. $|G(e^{j\omega})|^2$ must satisfy (2.16).

The Euler–Lagrange equation (2.16) is a pointwise relation that must be satisfied at all frequencies. The value of the unknown parameter $\beta(\omega)$ in the right-hand side is therefore set according to two criterions: First, the choice of $\beta(\omega)$ should not violate the Euler–Lagrange equation at any frequency. Second, the choice of $\beta(\omega)$ should insure the positivity of the solution at all frequencies. There are therefore two possible values for $\beta(\omega)$.

Case of $\beta(\omega) = 0$ Assume first that $\beta(\omega) = 0$. The left-hand side of (2.16) is now equal to zero and, in this case, (2.16) can be interpreted as the Euler–Lagrange equation for an exactly similar problem *without a positivity constraint* on the solution. Therefore, for those frequencies where $\beta(\omega) = 0$, the positivity constraint is actually ineffective and the solution we obtain must be ≥ 0 at those frequencies. The optimum magnitude squared response $|G_{\text{opt}}(e^{j\omega})|^2$ in this case is determined from (2.16) with the right-hand side set to zero. Perform the partial differentiation in (2.16) and equating the result to zero, the following equation can be obtained:

$$\left(S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}\right)^2 = \frac{c2^{-2b}}{\lambda} S_{xx}(e^{j\omega}) \tag{A.4}$$

Taking the square root of (A.4) and simplifying, we get: where $\gamma = \sqrt{\lambda}$.

$$|G(e^{j\omega})|^2 = \frac{1}{\gamma}\sqrt{\frac{c2^{-2b}}{S_{xx}(e^{j\omega})}} - \frac{c2^{-2b}}{S_{xx}(e^{j\omega})} \tag{A.5}$$

where $\gamma = \sqrt{\lambda}$. Substituting $|G(e^{j\omega})|^2$ as in (A.5) into the constraint (2.8), we obtain

$$\frac{1}{\gamma}\int_{-\pi}^{\pi} \sqrt{\frac{c2^{-2b}}{S_{xx}(e^{j\omega})}} S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} = 1 + \int_{-\pi}^{\pi} c2^{-2b} \frac{d\omega}{2\pi}. \tag{A.6}$$

Hence, the constant γ is given by

$$\gamma = \frac{\sqrt{c2^{-2b}}}{1 + c2^{-2b}} \int_{-\pi}^{\pi} \sqrt{S_{xx}(e^{j\omega})} \frac{d\omega}{2\pi} \tag{A.7}$$

Substituting γ as in (A.7) in (A.5), we therefore obtain part of (2.13), namely that:

$$|G(e^{j\omega})|^2 = \frac{1}{\sqrt{S_{xx}(e^{j\omega})}} \left(\frac{1 + c2^{-2b}}{\int_{-\pi}^{\pi} \sqrt{S_{xx}(e^{j\omega})} \frac{d\omega}{2\pi}} - \frac{c2^{-2b}}{\sqrt{S_{xx}(e^{j\omega})}} \right). \tag{A.8}$$

for all frequencies for which the right-hand side of (A.8) is nonnegative.

Case of $\beta(\omega) \neq 0$*:* At some particular frequency, the solution obtained in case 1 might turn out to be negative. The positivity constraint is obviously violated. At such a frequency, $\beta(\omega)$ should not bet set to zero anymore. Since the Euler–Lagrange equation must be satisfied at all times, we must set $\beta(\omega)$ to be equal to

$$-\frac{\partial}{\partial |G(e^{j\omega})|^2} \left(\frac{c2^{-2b} S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} + \lambda S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 \right).$$

The sign of $\beta(\omega)$ in this case is important to make sure that the positivity constraint is not violated. For our problem, $\beta(\omega)$ should be **nonpositive**. Finally, it remains to find the value of $|G_{\text{opt}}(e^{j\omega})|^2$ when $\beta(\omega) \neq 0$. The Euler–Lagrange equation cannot be used anymore because it determines the unknown parameter $\beta(\omega)$. However, we can simply observe that $|G_{\text{opt}}(e^{j\omega})|^2$ cannot be greater than zero. According to the first case, if $|G_{\text{opt}}(e^{j\omega})|^2$ is set to any value greater than zero, $\beta(\omega)$ should be zero. The only possible remaining value for $|G_{\text{opt}}(e^{j\omega})|^2$ is therefore zero. This argument establishes the complete form of (2.13).

From the above construction, we see that $|G_{\text{opt}}(e^{j\omega})|^2$ is a smooth function of ω (i.e. it is continuous with continuous first-order derivative) everywhere except at the frequencies where it has to be forced to zero (so it does not turn negative). The frequencies ω_k at which $|G_{\text{opt}}(e^{j\omega})|^2$ is set to zero are called corner points. To be an acceptable piecewise smooth solution, $|G_{\text{opt}}(e^{j\omega})|^2$ must satisfy the so called Weistrass–Erdmann conditions at those frequencies. In our case, the Weistrass–Erdmann conditions reduce to the requirement that the integrand in (2.16) be a continuous function of ω at the corner frequencies ω_k. This requirement is indeed satisfied because the integrand is a continuous function of $|G(e^{j\omega})|^2$ which in turn is continuous in ω even at the corner points ω_k

Step 3: We would like now to prove that the magnitude response expression (2.13) is not only necessary but also sufficient for the optimality of the prefilter. We introduce the following notation by rewriting (2.15) as follows:

$$\int_{-\pi}^{\pi} (f_1(\omega, y(\omega)) + f_2(\omega, y(\omega)) + f_3(\omega, y(\omega))) \frac{d\omega}{2\pi} \tag{A.9}$$

where

$$\begin{aligned} f_1(\omega, y(\omega)) &= \frac{\sigma_q^2 S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})y(\omega) + \sigma_q^2} \\ f_2(\omega, y(\omega)) &= \lambda S_{xx}(e^{j\omega})y(\omega) \\ f_3(\omega, y(\omega)) &= \beta(\omega)y(\omega) \\ y(\omega) &= |G(e^{j\omega})|^2 \\ \sigma_q^2 &= c2^{-2b}. \end{aligned} \tag{A.10}$$

Before proceeding further, we can now summarize the following useful facts from [19]:

Fact 1: The sum of a convex function with one or more convex functions is again convex.

Fact 2: If $f(\omega, y(\omega))$ is convex on $[-\pi, \pi] \times \mathcal{D}$, then, $J[y(\omega)] = \int_{-\pi}^{\pi} f(\omega, y(\omega)) \frac{d\omega}{2\pi}$ is convex on $\mathcal{D}$. Hence, each $y(\omega) \in \mathcal{D}$ that satisfies the necessary conditions of Step 2 minimizes $J[y(\omega)]$ on $\mathcal{D}$.

From the above two facts, it is then clear that to prove that the solution (2.13) is a minimizing curve, we simply need to prove the convexity of $f_i(\omega, y(\omega))\ \forall i$. The convexity of $f_i(\omega, y(\omega))$ on $[-\pi, \pi] \times \mathcal{D}$ can be established by using **anyone** of the following two conditions:

1. The following inequality must be satisfied $\forall\ (\omega, y(\omega))$ and $\forall\ (\omega, y(\omega) + v(\omega)) \in [-\pi, \pi] \times \mathcal{D}$:

$$\begin{aligned} &f_i(\omega, y(\omega) + v(\omega)) - f_i(\omega, y(\omega)) \\ &\quad \geq \left(\frac{\partial}{\partial y(\omega)} f_i(\omega, y(\omega)) \right) \cdot v(\omega). \end{aligned} \tag{A.11}$$

.

2. The matrix of second partial derivatives

$$\begin{bmatrix} f_{yy} & f_{yy'} \\ f_{yy'} & f_{y'y'} \end{bmatrix} \tag{A.12}$$

must be positive semidefinite on $[-\pi, \pi] \times \mathcal{D}$.

In the above two conditions, all the partial derivatives are assumed to be continuous on $[-\pi, \pi] \times \mathcal{D}$. The notation y' is used for the derivative of $y(\omega)$ with respect to ω. We use condition (A.11) to prove the convexity of $f_2(\omega, y(\omega))$ and $f_3(\omega, y(\omega))$ and condition (A.12) to prove the convexity of $f_1(\omega, y(\omega))$.

Convexity of $f_2(\omega, y(\omega))$ and $f_3(\omega, y(\omega))$: Assume first that $f(\omega, y(\omega)) = f_2(\omega, y(\omega))$ in (A.11). It is then easy to check, in this case, that the right-hand side of the equation is equal to the left-hand side. In fact, both sides will be equal to $\lambda \cdot S_{xx}(e^{j\omega}) \cdot v(\omega)$. Similarly, when $f(\omega, y(\omega)) = f_3(\omega, y(\omega))$, the right-hand side of (A.11) is equal to the left-hand side of the same equation. The two sides are, in turn, equal to $\beta(\omega) \cdot v(\omega)$. This establish the convexity of both $f_2(\omega, y(\omega))$ and $f_3(\omega, y(\omega))$.

Convexity of $f_1(\omega, y(\omega))$: When $f(\omega, y(\omega)) = f_1(\omega, y(\omega))$, then, we first observe that the matrix in (A.12) can be simplified to the following form:

$$\begin{bmatrix} f_{yy} & 0 \\ 0 & 0 \end{bmatrix}. \tag{A.13}$$

For this matrix to be positive semidefinite, the principal minors should be nonnegative. From (A.13), this is equivalent to proving that $f_{yy} \geq 0\ \forall \omega$. Differentiating $f_1(\omega, y(\omega))$ twice with respect to $y(\omega)$, we obtain the following equation:

$$f_{yy} = \frac{\sigma_q^2 S_{xx}^3(e^{j\omega})}{\left(S_{xx}(e^{j\omega})y(\omega) + \sigma_q^2\right)^3} \tag{A.14}$$

Since all quantities in (A.14) are positive, then, the condition (A.12) is indeed satisfied and $f_1(\omega, y(\omega))$ is convex. Using the convexity of the above functions and facts one and two, we conclude that the solution (2.13) is a minimizing extremum. □

APPENDIX B

Using the following set of equations:

$$\begin{aligned} e(n) &= x(n) - \hat{x}(n) \\ \hat{x}(n) &= z(n) \otimes h(n) = \sum_{k=0}^{\infty} \gamma^k z(n-k) \\ z(n) &= y(n) + q(n) \\ y(n) &= x(n) - \alpha x(n-1) \end{aligned}$$

we can easily verify by direct substitution that the error process at the output of the postfilter is given by

$$\begin{aligned} e(n) &= x(n) - \sum_{k=0}^{\infty} \gamma^k x(n-k) \\ &\quad + \alpha \sum_{k=0}^{\infty} \gamma^k x(n-k-1) - \sum_{k=0}^{\infty} \gamma^k q(n-k) \\ &= \left(\frac{\alpha}{\gamma} - 1 \right) \sum_{k=1}^{\infty} \gamma^k x(n-k) - \sum_{k=0}^{\infty} \gamma^k q(n-k). \end{aligned} \tag{B.1}$$

The MSE expression is defined to be $\mathcal{E} \triangleq E\ \{e^2(n)\}$. This, in turn can be written as

$$\begin{aligned} \mathcal{E} = E\Bigg\{ &\left(\left(\frac{\alpha}{\gamma} - 1 \right) \sum_{k=1}^{\infty} \gamma^k x(n-k) - \sum_{k=0}^{\infty} \gamma^k q(n-k) \right) \\ &\cdot \left(\left(\frac{\alpha}{\gamma} - 1 \right) \sum_{l=1}^{\infty} \gamma^l x(n-l) - \sum_{l=0}^{\infty} \gamma^l q(n-l) \right) \Bigg\}. \end{aligned} \tag{B.2}$$

This last equation can be simplified using the following assumptions about the noise process $q(n)$:

1) *White noise assumption*: $E\{q(n) \cdot q(n-k)\} = \sigma_q^2 \delta(n-k)$.
2) *Variance constraint assumption*: The noise variance σ_q^2 is equal to $c2^{-2b}\sigma_y^2$ where σ_y^2 is the variance of the quantizer input. Hence, $\sigma_q^2 = c2^{-2b}(R_{xx}(0)(1+\alpha^2) - 2\alpha R_{xx}(1))$.
3) *Uncorrelatedness with $x(n)$*: The sequence $x(n)$ and $q(n)$ are assumed to be uncorrelated. Hence,

$$E\{x(n) \cdot q(n-k)\} = E\{q(n) \cdot x(n-k)\} = 0 \qquad \forall k.$$

Based on the above assumptions, (B.2) can be therefore simplified. The result gives the following expression for the MSE:

$$\mathcal{E} = c2^{-2b}((1+\alpha^2)R_{xx}(0) - 2\alpha R_{xx}(1))\sum_{k=0}^{\infty}\gamma^{2k} + \left(\frac{\alpha}{\gamma} - 1\right)^2 \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\gamma^{l+k}R_{xx}(k-l). \quad \text{(B.3)}$$

This last expression consists of two terms and can be further simplified. The first term of (B.3) can be rewritten as follows:

$$c2^{-2b}((1+\alpha^2)R_{xx}(0) - 2\alpha R_{xx}(1))\frac{1}{(1-\gamma^2)}. \quad \text{(B.4)}$$

The second term can be divided into two subterms, one for $k = l$ and the other for $k \neq l$ to obtain

$$\left(\frac{\alpha}{\gamma} - 1\right)^2 \sum_{\substack{k=1\\k=l}}^{\infty}\gamma^{2l}R_{xx}(0) + \left(\frac{\alpha}{\gamma} - 1\right)^2 \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\gamma^{l+k}R_{xx}(k-l) \quad \text{(B.5)}$$

which in turn can be rewritten as

$$\left(\frac{\alpha-\gamma}{\gamma}\right)^2 \frac{\gamma^2}{1-\gamma^2}R_{xx}(0) + 2\left(\frac{\alpha-\gamma}{\gamma}\right)^2 \frac{\gamma^2}{1-\gamma^2}\sum_{m=1}^{\infty}\gamma^m R_{xx}(m). \quad \text{(B.6)}$$

Adding (B.4) and (B.6) we obtain the MSE expression of Theorem 4.1. □

REFERENCES

[1] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[2] J. Kovacevic, "Subband coding systems incorporating quantizer models," *IEEE Trans. Image Processing*, vol. 4, pp. 543–553, May 1995.

[3] N. Uzun and R. A. Haddad, "Cyclostationary modeling, analysis and optimal compensation of quantization errors in subband codecs," *IEEE Trans. Signal Processing*, vol. 43, pp. 2109–2119, Sept. 1995.

[4] K. Gosse and P. Duhamel, "Perfect reconstruction versus MMSE filter banks in source coding," *IEEE Trans. Signal Processing*, vol. 45, pp. 288–2202, Sept. 1997.

[5] A. Delopoulos and S. Kolias, "Optimal filterbanks for signal reconstruction from noisy subband components,"*IEEE Trans. Signal Processing*, vol. 44, no. 2, pp. 212–224, Feb. 1996.

[6] M. Unser, "On the optimality of ideal filters for pyramid and wavelet signal approximation," *IEEE Trans. Signal Processing*, pp. 3591–3596, Dec. 1993.

[7] P. Delsarte, B. Macq, and D. Slock, "Signal-adapted multiresolution transform for image coding," *IEEE Trans. Inform. Theory*, vol. 38, pp. 897–904, Mar. 1995.

[8] M. K. Tsatsanis and G. B. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Signal Processing*, vol. 43, pp. 1766–1777, Aug. 1995.

[9] P. P. Vaidyanathan, "Optimal orthonormal filterbanks," to be published.

[10] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice Hall, 1984.

[11] P. P. Vaidyanathan and T. Chen, "Statistically optimal synthesis banks for subband coders reconstruction," in *Proc. 28th Annu. Asilomar Conf. Sig., Sys. and Comp.*, Oct.–Nov. 1994.

[12] J. P. Costas, "Coding with linear systems," *Proc. IRE*, vol. 40, pp. 1101–1103, Sept. 1952.

[13] D. Chan and R. W. Donaldson, "Optimum pre- and postfiltering of sampled signals with application to pulse modulation and data compression systems," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 141–157, Apr. 1971.

[14] T. Berger and D. Tufts, "Optimum pulse amplitude modulation—Part I: Transmitter–receiver design and bounds from information theory," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 196–208, Apr. 1967.

[15] H. S. Malvar and D. H. Staelin, "Optimal FIR pre- and postfilters for decimation and interpolation of random signals," *IEEE Trans. Commun.*, vol. 36, pp. 67–74, Jan. 1988.

[16] I. Djokovic and P. P. Vaidyanathan, "On optimal analysis/synthesis filters for coding gain maximization," *IEEE Trans. Signal Process.*, vol. 44, pp. 1276–1279, May 1996.

[17] M. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: Wiley, 1996.

[18] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. Englewood Cliffs, NJ: Prentice Hall, 1963.

[19] J. L. Troutman, *Variational Calculus with Elementary Convexity*. New York: Springer-Verlag, 1983.

[20] D. E. Kirk, *Optimal Control Theory: An Introduction*. Englewood Cliffs, NJ: Prentice Hall, 1970.

[21] L. M. Goodman and P. R. Drouillet, "Asymptotically optimum preemphasis and de-emphasis networks for sampling and quantizing," *Proc. IEEE*, vol. 51, pp. 795–796, May 1963.

[22] C. J. Demeure and C. T. Mullis, "The Euclid algorithm and the fast computation of cross-covariance and autocovariance sequences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 545–552, Apr. 1989.

[23] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 5th ed. San Diego, CA: Academic, 1994.

[24] A. K. Soman and P. P. Vaidyanathan, "Coding gain in paraunitary analysis/synthesis systems," *IEEE Trans. Signal Processing*, vol. 41, pp. 1824–1835, May 1993.

[25] V. S. 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.

[26] J. Kovacevic and M. Vetterli, "Perfect reconstruction filter banks with rational sampling rate changes," in *Proc. ICASSP*, Toronto, Canada, 1991, pp. 1785–1788.

[27] P. P. Vaidyanathan, "Orthonormal and biorthonormal filter banks as convolvers, and convolutional coding gain," *IEEE Trans. Signal Processing*\, vol. 41, pp. 2110–2130, June 1993.

[28] P.-Q. Hoang and P. P. Vaidyanathan, "Non-uniform multirate filter banks: Theory and design," in *Proc. ISCAS*, Portland, OR, 1989, pp. 371–374.

[29] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice Hall, 1989.

Jamal Tuqan (S'91) was born in Cairo, Egypt, in 1966. He received the B.Sc. degree in electrical engineering from Cairo University, Egypt, in 1989, the M.S.E.E. degree from the Georgia Institute of Technology, Atlanta, in 1992, and the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1997.

From September 1989 to July 1990, he worked at the IBM Research Center in Cairo, writing software for Arabic speech compression based on linear predictive coding techniques. From September 1993 to June 1997 he has been a teaching assistant at the California Institute of Technology, Pasadena. Since July 1994, he has been a research assistant at the same university. His research interests are primarily in the compression area with emphasis on statistical optimization of multirate systems and filter banks. He is also interested in multirate signal processing applications in digital communications. He is a student member of the IEEE Signal Processing, Circuits and Systems, and Information Theory Societies.

P. P. Vaidyanathan (S'80–M'83–SM'88–F'91) was born in Calcutta, India, on Oct. 16, 1954. He received the B.Sc. (Hons.) degree in physics and the B.Tech. and M.Tech. degrees in radiophysics and electronics, all from the University of Calcutta, in 1974, 1977, and 1979, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California at Santa Barbara, CA, in 1982.

He was a Post-Doctoral Fellow at the University of California, Santa Barbara, from September 1982 to March 1983. In March 1983, he joined the Electrical Engineering Department of the California Institute of Technology as an Assistant Professor, and since 1993 has been Professor of Electrical Engineering there. His main research interests are in digital signal processing, multirate systems, wavelet transforms, and adaptive filtering. He has authored a number of papers in IEEE journals and is the author of the book *Multirate systems and filter banks*. He has written several chapters for various signal processing handbooks.

Dr. Vaidyanathan served as Vice Chairman of the Technical Program Committee for the 1983 IEEE International symposium on Circuits and Systems and as the Technical Program Chairman for the 1992 IEEE International symposium on Circuits and Systems. He was an Associate Editor for the IEEE Transactions on Circuits and Systems for the period 1985–1987 and is currently an Associate Editor for the journal *IEEE Signal Processing Letters* and a Consulting Editor for the journal *Applied and Computational Harmonic Analysis*. He was a recipient of the Award for excellence in teaching at the California Institute of Technology for the years 1983–1984, 1992–93, and 1993–94. He also received the NSF's Presidential Young Investigator Award in 1986. In 1989, he received the IEEE ASSP Senior Award for his paper on multirate perfect-reconstruction filter banks. In 1990, he was the recipient of the S. K. Mitra Memorial Award from the Institute of Electronics and Telecommunications Engineers, India, for his joint paper in the IETE journal. He was also the coauthor of a paper on linear-phase perfect reconstruction filter banks in the IEEE SP Transactions for which the first author (Truong Nguyen) received the *Young Outstanding Author* award in 1993. He received the 1995 F. E. Terman Award of the American Society for Engineering Education, sponsored by Hewlett Packard Co., for his contributions to engineering education, especially the book *Multirate systems and filter banks* published by Prentice Hall in 1993. He has been chosen a Distinguished Lecturer for the IEEE Signal Processing Society for the year 1996–97.