Asilomar Conference, November 1996.

Optimum Quantization of a Class of Non Bandlimited Signals ‡

Jamal Tuqan and P. P. Vaidyanathan
Department of Electrical Engineering 136-93
California Institute of Technology
Pasadena, CA 91125
E-mail : tuqan@systems.caltech.edu, ppvnath@sys.caltech.edu

Abstract

We consider the quantization of a special class of non bandlimited signals, namely the class of discrete time signals that can be recovered from their decimated version. Similar to sigma-delta modulation ideas, we show that we can obtain a great reduction in the quantization noise variance due to the oversampled nature of these signals. We then consider noise shaping by optimizing a pre- and post filter around the quantizer and develop a closed form expression for the coding gain of the scheme under study. The use of an orthonormal filter bank as a sophisticated quantizer is investigated and several examples are provided.

1. Introduction

Walter [1] showed that, under some conditions, a class of non bandlimited continuous time signals can be reconstructed from uniformly spaced samples even though aliasing occurs. Vaidyanathan and Phoong [2], [3] developed the discrete time version of Walter's result from a multirate digital filtering perspective. In specific, they introduced the class of non bandlimited signals that can be modeled as the output of a single interpolation filter (single band model) as in Fig. 1.1 or as the output of the more general multiband model of Fig. 1.2. It can be shown that this class of non bandlimited signals can be recovered from its decimated version. As a quick example, assume that $x(n)$ is modeled as in Fig. 1.1 and consider $x(Mn)$, the M-fold decimated versions of $x(n)$. If $F(e^{j\omega})$ is a Nyquist(M) filter [4], then, $x(Mn)$ is equal to $y(n)$ and we have the relation $x(n) = \sum_k x(kM) f(n - kM)$. In other words, $x(n)$ is completely defined by the samples $x(Mn)$ even though the filter $F(e^{j\omega})$ is not necessarily ideal. More elaborate "sampling theorems" can be developed for Fig 1.1 and Fig. 1.2. The details can be found in [2] and [3].

In this paper, we restrict our analysis to the single band model of Fig. 1.1 and consider the efficient quantization of this class of non band-limited signals. To motivate such a study, consider the schematic of traditional sigma delta modulation shown in Fig. 1.3 where the box labeled $\mathcal{Q}$ represents a PCM quantizer. The signal $x(n)$ is assumed to be bandlimited (oversampled). After the quantization, the ideal low pass filter on the right chops off the noise in the stopband but does not change the signal component. The signal power is the same whereas the noise power decreases proportionally to the oversampling ratio. For bandlimited signals, we can therefore quantize the signal to very few bits, perhaps even one bit if we oversample the signal enough. In addition, we can generate a further decrease in the noise power by introducing *noise shaping* in the signal band, as for example in sigma-delta modulators, to allow higher resolution quantization of bandlimited signals.

The schematic description of Fig. 1.3 shows that we can take advantage of a signal model like Fig. 1.1 even though $x(n)$ is not bandlimited. Thus, consider Fig. 1.4 where the finite order filter $F(e^{j\omega})$ is assumed to be an optimum compaction filter. In specific, the filter maximizes the variance of its output signal under the constraint that its magnitude squared response $|F(e^{j\omega})|^2$ is Nyquist(M), that is, $(|F(e^{j\omega})|^2) \downarrow_M = 1$. The assumption is motivated by the fact that this particular choice of filters minimize the mean square reconstruction error between a signal, say $x(n)$, and its approximation modeled as in Fig. 1.1 [5]. With this last assumption, it can be shown that the signal $\hat{x}(n)$ in Fig. 1.4 is equal to $x(n)$ in the absence of the quantizer. The entire scheme of Fig. 1.4 behaves similar to Fig. 1.3, except that the low pass filtering is *multirate* and *non ideal*. Thus, generally speaking, if a non bandlimited signal can be reconstructed from its samples $x(Mn)$ because it satisfies a model like Fig. 1.1, then, a low precision quantizer should allow us to produce a high precision version $\hat{x}(n)$.

The quantization advantage offered by Fig. 1.4 can be

‡ Work supported in parts by Office of Naval Research grant N00014-93-1-0231, Tektronix, Inc., and Rockwell International.

useful, for example, in the following realistic engineering scenario. Suppose $x(n)$ is generated at a point where we cannot afford very complex signal processing (e.g., in deep space) and needs to be transmitted to a distant place (e.g., earth station). If we have the knowledge that $x(n)$ admits a satisfactory model like Fig. 1.1, we can compress it using a very simple low pass filter $P(e^{j\omega})$ with one or two multipliers and then quantize the output before transmission. The post filter $1/P(e^{j\omega})$ and the expensive multirate filter are at the receiver end, where the complexity is acceptable. In the sequel we shall find an expression for the theoretically best $P(e^{j\omega})$ without constraint on order. This will give an upper bound on the gain obtainable with a practical inexpensive $P(e^{j\omega})$.

2. Exploiting the signal model

Consider the set up shown in Fig. 1.4 where the input signal $x(n)$ satisfies the the single band model of Fig. 1.1. Our assumptions are as follows : the driving signal $y(n)$ in Fig. 1.1 is a zero mean wide-sense stationary (WSS) random process. The filter $F(e^{j\omega})$ is FIR and has the property that $|F(e^{j\omega})|^2$ is Nyquist(M) for the reasons described above. Because the model filter $F(e^{j\omega})$ is not ideal, the input signal $x(n)$ in Fig 1.4 is a cyclo wide-sense stationary signal of period M $\left((CWSS)_M\right)$ [6]. The box labeled $\mathcal{Q}$ represents a scalar uniform (PCM) quantizer and is modeled as an additive zero mean white noise source $q(n)$. We will design the quantizer $\mathcal{Q}$ as follows : since the input to the quantizer $x(n)$ is a $(CWSS)_M$ process, its variance $\sigma_x^2(n)$ is a periodic function of n with period M. Define σ_x^2 to be the average variance of $x(n)$, i.e., $\sigma_x^2 = \frac{1}{M}\sum_{n=0}^{M-1}\sigma_x^2(n)$. Then, choose the fixed step size Δ in the uniform quantizer such that the quantization noise variance σ_q^2 is directly proportional to the variance of the quantizer input $x(n)$. In specific, we design the uniform quantizer such that the following relation holds

$$\sigma_q^2 = c2^{-2b}\sigma_x^2 \qquad (2.1)$$

where σ_q^2 is the quantization noise variance, c is a constant that depends on the statistical distribution of $x(n)$ and the overflow probability, and σ_x^2 is the average variance of the quantizer input. The above relation is justified for a PCM quantizer using 3 (or more) bits per sample (see chapter 4 in [7]). The next theorem gives an expression for the average mean squared error $\mathcal{E} \triangleq \frac{1}{M}\sum_{n=0}^{M-1} E\{\hat{x}(n) - x(n)\}^2$.

Theorem 2.1. *Consider the scheme of Fig. 1.4 under the above assumptions. The average mean square error $\mathcal{E}$ is equal to $\frac{1}{M}\sigma_q^2$.*

The proof can be found in [8]. The quantization noise variance σ_q^2 obtained by directly quantizing $x(n)$ as shown in Fig. 2.1 is now reduced proportionally to the oversampling factor M. The signal variance σ_x^2 on the other hand did not change. By expressing the interpolator M in the form 2^r, we can immediately see that we can get the same quantitative advantage of the oversampling PCM technique, namely, an increase in SNR by 3 db for every doubling of the oversampling factor. For example, if $M = 2$, we get an SNR increase of 3 db whereas if $M = 4$, the SNR increment is by 6 db. The result of theorem 2.1 can be intuitively explained. The signal $x(n)$, modeled as in Fig. 1.1 is oversampled and therefore, contains redundant information in the form of an excess of samples. It is by quantizing these extra samples that we obtain the reduction in the quantization noise variance (equivalently in the mean square error). We are therefore effectively quantizing with a higher number of bits per sample. This trade off, between the quantization noise variance (effective quantizer resolution) and the sampling rate is the underlying principle of oversampled A/D converters.

A consequence of the previous results and discussion is then the natural question: what if the discrete time filtering of the oversampled signal is not a major burden ? If we know that $x(n)$ can be modeled quite accurately by the filter $F(e^{j\omega})$ of Fig. 1.1, we can filter and downsample $x(n)$ to obtain $y(n)$ as shown in Fig. 2.2. We can then in principle quantize the decimated signal $y(n)$ with $\hat{b} = Mb$ bits per sample. This situation is equivalent to fixing the bit rate (number of bits per second) to be equal to b in order to trade quantization resolution with sampling rate. At this point, we will however assume that the goal is to actually obtain a reduction in the bit rate. To achieve this, we fix the number of bits per sample $\hat{b}$ to be equal to b. Since the quantizer resolution did not increase, the quantization noise variance should not differ from the direct quantization case of Fig. 2.1. This last statement is verified formally in the next theorem.

Theorem 2.2. *Consider the scheme of Fig. 2.2. With a fixed number of quantizer bits b, the average mean square error $\mathcal{E}$ is equal to σ_q^2, where σ_q^2 is the noise variance obtained from directly quantizing $x(n)$ using b bits.*

The proof can be found in [8].

3. Noise Shaping

Following the philosophy of sigma-delta modulators, we would like now to perform noise shaping with the hope of achieving a further reduction in the average mean square error. To accomplish this, we propose using LTI pre- and post filters around the PCM quantizer as shown in Fig. 3.1. The goal is to optimize the filter $P(e^{j\omega})$ such that the average

m.s.e at the output of Fig. 3.1. is minimized. At this point, no order constraint is imposed on the filters and non causal solutions are accepted.

Following (2.1), the quantizer noise variance in this case is given by $\sigma_q^2 = c2^{-2b}\sigma_z^2$ where σ_z^2 is the average variance of the process $z(n)$. We emphasize that $z(n)$ is a $(CWSS)_M$ process since the output of a linear time invariant filter driven by a $(CWSS)_M$ process is also $(CWSS)_M$ [6]. It is then possible to express σ_z^2 in terms of the prefilter $P(e^{j\omega})$ and the so called average power spectral density of the process $x(n)$, denoted by $\hat{S}_{xx}(e^{j\omega})$, as follows :

$$\sigma_z^2 = \frac{1}{M}\int_{-\pi}^{\pi} |P(e^{j\omega})|^2 \hat{S}_{xx}(e^{j\omega})\frac{d\omega}{2\pi} \qquad (3.1)$$

The average power spectral density is a familiar concept that arises when stationarizing a $(CWSS)_M$ process [9] and satisfies the well known properties of the power spectrum of a WSS process. It is defined to be the discrete time Fourier transform of the time averaged autocorrelation function $\hat{R}_{xx}(k)$ given by $\frac{1}{M}\sum_{n=0}^{M-1} E[x(n)x^*(n-k)]$.

Theorem 3.1. *Consider the scheme of Fig. 3.1 under the same assumptions of section II. The optimum prefilter $P(e^{j\omega})$ that minimizes the average mean square reconstruction error has the following magnitude squared response:*

$$|P_{opt}(e^{j\omega})|^2 = \frac{|F(e^{j\omega})|}{\sqrt{\hat{S}_{xx}(e^{j\omega})}} \qquad (3.2)$$

The proof of equations (3.1) and (3.2) can again be found in [8]. A number of observations should be made at this point. First, the optimum filter is not unique since the phase response is not specified. Second, the above derivation assumes that the input spectrum $\hat{S}_{xx}(e^{j\omega}) \neq 0$ for all ω. The assumption is a reasonable one because $x(n)$ is assumed to be non bandlimited and therefore $\hat{S}_{xx}(e^{j\omega})$ cannot be identically zero on a segment of $[0, 2\pi)$. If $\hat{S}_{xx}(e^{j\omega})$ has an isolated zero for some ω, then, the resulting prefilter will have a zero on the unit circle and is therefore unstable. In any case, a practical system would use only a stable rational approximation of the ideal solution. Using (3.2), we can derive an interesting expression for the coding gain of the scheme of Fig 3.1. The coding gain of a quantization scheme is defined to be the ratio $\mathcal{E}_{direct}/\mathcal{E}_{min}$ where $\mathcal{E}_{direct}$ is the mean square error obtained by quantizing $x(n)$ directly with b bits as shown in Fig. 2.1 and $\mathcal{E}_{min}$ is the minimum mean squared error obtained by using optimum pre and post filters around the quantizer under a fixed bit rate assumption.

Theorem 3.2. *With the optimum choice of the pre- and post filters, the coding gain expression for the scheme of 3.1 is*

$$\mathcal{G}_{opt} = \frac{M\int_{-\pi}^{\pi} S_{yy}(e^{j\omega})\frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}\frac{d\omega}{2\pi}\right)^2} = M\cdot\mathcal{G}_{hw} \qquad (3.3)$$

where $\mathcal{G}_{hw}$ is the half whitening coding gain [7] of the WSS process $y(n)$.

The factor M in (3.3) is again due to the oversampled nature of the signal $x(n)$. It is interesting to note that the noise shaping contribution to $\mathcal{G}_{opt}$ in (3.3), which we denote by $\mathcal{G}_{hw}$, is exactly the *coding gain we would obtain by half whitening the WSS process $y(n)$ in the usual way* [7]. By appealing to the Cauchy Schwartz inequality, we can show that $\mathcal{G}_{hw} \geq 1$ with equality iff the power spectral density $S_{yy}(e^{j\omega})$ is a constant, i.e., $y(n)$ is white noise. Therefore, for the particular system of Fig. 3.1, we will not get additional coding gain by noise shaping if the driving WSS process $y(n)$ in Fig. 1.1 is white noise.

Example 6.1. *Case of a MA(1) process $y(n)$.* Assume that the input $x(n)$ is modeled as in Fig. 1.1 where the upsampler $M = 2$, the filter $F(e^{j\omega})$ is the optimum orthonormal FIR filter of length two given by $\frac{1}{\sqrt{2}}(1+z^{-1})$ and the driving WSS signal $y(n)$ is a zero mean gaussian MA(1) process with an autocorrelation sequence in the form

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

It is well known that a MA(1) process has to have $\frac{|R_{yy}(1)|}{R_{yy}(0)} \leq 1/2$ to ensure that the power spectral density is indeed non negative. We therefore restrict θ to be between -1 and 1. The power spectrum of the MA(1) process is given by:

$$S_{yy}(e^{j\omega}) = 1 - 2\frac{\theta}{(1+\theta^2)}cos(\omega) \qquad (3.4)$$

Substituting (3.4) in (3.3) and after some manipulations, the coding gain of the scheme of Fig. 3.1 can be expressed as:

$$\mathcal{G}_{opt} = 2\frac{(1+\theta^2)}{[(1+\theta)\frac{2}{\pi}E(2\sqrt{(|\theta|)}/(1+\theta))]^2} \qquad (3.5)$$

where $E(.)$ is the complete elliptic integral of the second kind. The plot of the coding gain as a function of θ is shown in Fig. 3.2.

Example 6.2. *Case of an AR(1) process $y(n)$.* With the same assumptions as in example 6.1, let the driving signal $y(n)$ be a zero mean gaussian AR(1) process with an

autocorrelation sequence in the form $R_{yy}(k) = \rho^{|k|}$ where ρ is between 0 and 1. The power spectrum of the AR(1) process is

$$S_{yy}(e^{j\omega}) = \frac{1-\rho^2}{1+\rho^2-2\rho cos(\omega)} \qquad (3.6)$$

Substituting (3.6) in (3.3) and simplifying, the coding gain for the scheme of Fig. 3.1 can be expressed as follows:

$$\mathcal{G}_{opt} = \frac{2}{(1-\rho^2)(\frac{2}{\pi}K(\rho))^2} \qquad (3.7)$$

where $K(\rho)$ is the complete elliptic integral of the first kind. The plot of the coding gain as a function of ρ is shown in Fig. 3.3.

4. Using an orthonormal filter bank

Since the signal model $x(n)$ is $(CWSS)_M$ by construction, restricting ourselves to linear time invariant noise shaping filters and quantizers is a loss of generality. The optimum noise shaping filters for such processes should be linear periodically time varying $(LPTV)_M$ filters surrounding a $(PTV)_M$ quantizer. This is equivalent to M LTI analysis and synthesis filters surrounding M time invariant uniform scalar quantizers. We can further impose the perfect reconstruction condition in the absence of quantization by confining ourselves to the class of M LTI analysis and synthesis filters satisfying the biorthogonality condition: $(P_k(e^{j\omega})Q_m(e^{j\omega}))|\downarrow_M = \delta(m-k)$ for all k, m. The goal is then to find the set of M analysis and synthesis filters, $P_k(e^{j\omega})$ and $Q_k(e^{j\omega})$, that minimize the average mean square error at the output due to quantization noise. Because the general $(LPTV)_M$ problem is difficult to track analytically, we will restrict ourselves to the special class of orthonormal $(LPTV)_M$ filters, i.e., filters satisfying the following properties: $Q_k(e^{j\omega}) = \tilde{P}_k(e^{j\omega})$ for each k and $(P_k(e^{j\omega})\tilde{P}_m(e^{j\omega}))|\downarrow_M = \delta(m-k)$ for all k, m. The goal is to jointly allocate the subband bits b_k under the fixed bit rate constraint $b = \frac{1}{M}\sum_{k=0}^{M-1} b_k$ and optimize the filters $P_k(e^{j\omega})$ in order to minimize the average m.s.e.

Theorem 4.1. *Consider the scheme of Fig. 4.1 under the above assumptions. The synthesis section of the optimum orthonormal filter bank $\{P_k(e^{j\omega})\}$ corresponds to choosing one of the filters, say $\tilde{P}_0(e^{j\omega})$ to be equal to $\tilde{F}(e^{j\omega})$ and the remaining filters $\tilde{P}_k(e^{j\omega})$, $k = 1, \ldots, M-1$, to be orthogonal to $\tilde{P}_0(e^{j\omega})$. In this case, the optimum orthonormal filter bank reduces to Fig. 2.3 where the quantizer $\mathcal{Q}$ is now allocated Mb bits per sample.*

The proof can be found in [8]. The result of Theorem 4.1 is very intuitive and somewhat expected: Decimate the oversampled signal $x(n)$ according to its model and then quantize the decimated signal $y(n)$ in Fig. 2.2 with $\hat{b} = Mb$ bits per sample. This amounts to fixing the bit rate (number of bits per second) in order to trade quantization resolution with sampling rate. It is interesting though to see that this very intuitive scheme is equivalent to using an optimum orthonormal filter bank as a sophisticated quantizer to the input $x(n)$. Using (2.1), the coding gain expression can be derived and is equal to $2^{2b(M-1)}$. The coding gain depends on the bit rate b and can be quite large for moderate values of M and b.

References

[1] Walter, G.G., "A sampling theorem for wavelet subspaces", IEEE Trans. on Information Theory, pp. 881-884, March 1992.

[2] Vaidyanathan, P.P. and Phoong, S., "Reconstruction of sequences from non uniform samples", ISCAS Proc., pp. 601-604, Seattle 1995.

[3] Vaidyanathan, P.P. and Phoong, S., "Discrete time signals which can be recovered from samples", ICASSP Proc., pp. 1448-1451, Detroit 1995.

[4] Vaidyanathan, P.P., *Multirate systems and filter banks,* Prentice Hall, Inc., Englewood Cliffs, 1993.

[5] 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.

[6] Sathe, V.S. and Vaidyanathan, P.P., "Effects of multirate systems on the statistical properties of random signals", IEEE Trans. on Signal Processing, pp. 131-146, Vol. 41, January 1993.

[7] Jayant, N.S. and Noll, P. *Digital coding of waveforms,* Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1984.

[8] Tuqan, J. and Vaidyanathan, P.P., "Sigma-delta modulators like techniques for quantizing a class of non-bandlimited signals", to be submitted.

[9] Gardner, W.A. and Franks, L.E., "Characterization of cyclostationary random processes", IEEE Trans. on Information Theory, vol. 21, No. 1, pp.4-14, January 1975.

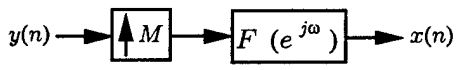

Fig. 1.1. The single band model .

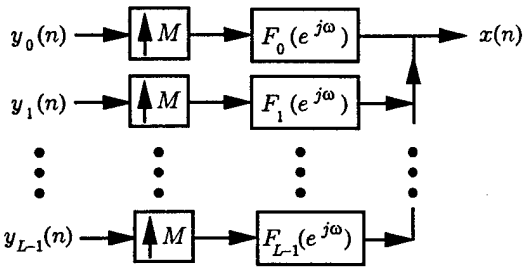

Fig. 1.2. The multiband model .

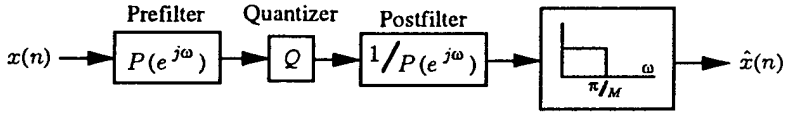

Fig. 1.4. Schematic of traditional sigma-delta modulator.

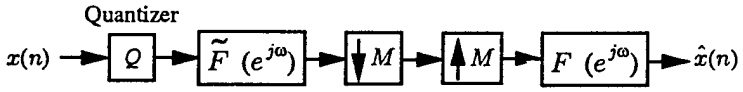

Fig. 2.1. Exploiting the signal model in quantization.

Fig. 2.2. Direct quantization of $x(n)$.

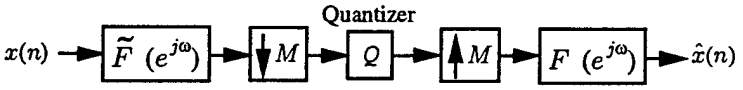

Fig. 2.3. Quantizing the lower rate signal $y(n)$.

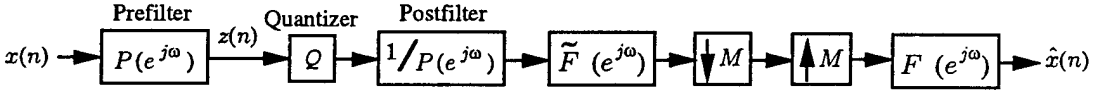

Fig. 3.1. Noise shaping using a LTI pre- and post filter.

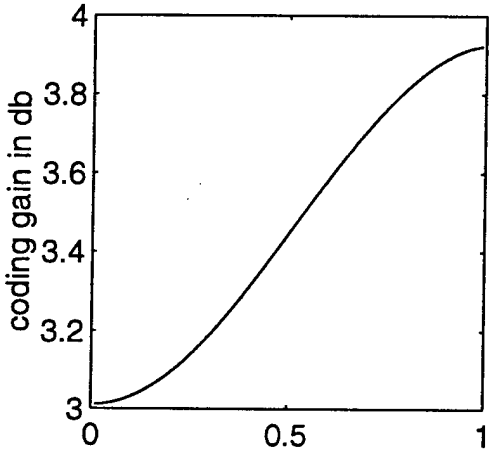

Fig. 3.2. Coding gain curve as a function of theta with $M = 2$ and $y(n)$ is an MA(1) process.

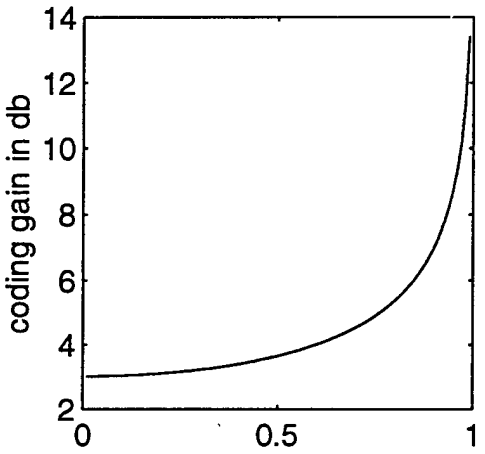

Fig. 3.3. Coding gain curve as a function of rho with $M = 2$ and $y(n)$ is an AR(1) process.

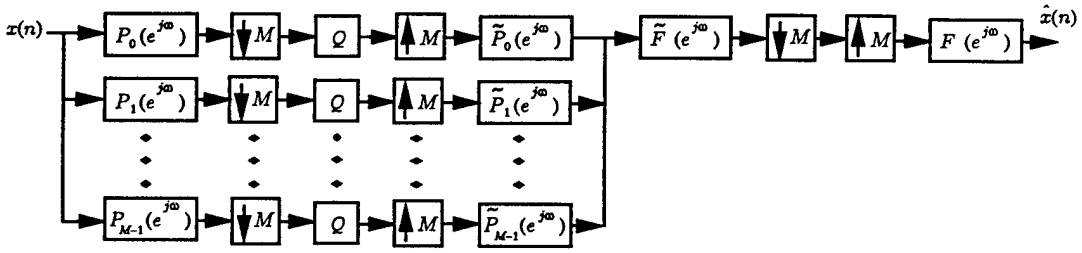

Fig. 4.1. Using an orthonormal filter bank as a sophisticated quantizer.