Optimal Subband Analysis Filters Compensating for Quantization and Additive Noise

Anastasios Doulamis, Nikolaos Doulamis and Anastasios Delopoulos
Dept. of Electrical and Computer Engineering,
National Technical University of Athens,
Iroon Polytechniou 9, Zografou, Athens 15773, GREECE
Tel: +30 1 7722491; fax: +30 1 7722492
e-mail: adoulam@image.ntua.gr

ABSTRACT

In this paper, we present an analysis filter design technique which optimally defines the proper decimator so that the quantization noise is compensated. The analysis is based on a distortion criterion minimization using the Lagrange multipliers. The optimal decimation filters are derived through a Ricatti solution which involves both the quantization and the interpolation filters. Experimental results are presented indicating the good performance of the proposed technique versus conventional subband filter banks in the presence of quantization noise.

1 INTRODUCTION

Subband processing, which is concerned with problems in which more than one signal rates are considered, has become a topic of extensive research in the recent years. It is already an important part of a wide variety of applications, such as audio and speech coding (e.g., the audio coding part of the MPEG algorithm [5]), progressive image coding and spectrum analysis. This is due to the fact that it permits the design of efficient systems for tuning the associated coding to arbitrary level [11]. Furthermore, the new forthcoming standards which are currently in progress for multimedia processing and very low bit rate coding, use techniques and theoretical aspects of this field for achieving better compression efficiency.

The design of decimation and interpolation filters (analysis/synthesis filter banks) is in the core of multirate signal processing. Extensive research has been carried out on the design of decimators/interpolators that achieve perfect reconstruction [8].

A perfect reconstruction filter bank reproduces the input signal exactly in the output. The associated analysis is based on the assumption that all subband components are available to the interpolation bank with infinite precision. However, in real life applications the aforementioned assumption is not fulfilled. For example, when filter banks are used for signal compression, a non linear quantization operation is performed in the subband domain. The quality of the reconstruction in this case depends on both filter and quantizer characteristics. In addition, external subband noise is possibly added to the signal during its transmission usually due to channel disturbances.

Most of the previously reported works refer to the design of appropriate compensators that preceding the synthesis stage suppress the effect of the noise [2,3,4,6,10]. In [9] necessary and sufficient conditions are described for the optimality of orthonormal perfect reconstruction filter banks. Optimal scalar quantization of the banks together with optimal bit allocation is assumed. An analysis in the frequency domain without the assumption of orthonormality is also given in [7]. A method which compensates quantization or any other type of additive noise by appropriately adjusting the synthesis filter bank is proposed in [1].

In this paper we propose a method for the design of the analysis filter banks so as to reduce the effect of noise for a given set of synthesis filters. Our method is purely developed in the time domain and uses the statistical average distortion power as a measure of similarity between the original and reconstructed signal. The advantage of adjusting the analysis bank rather than the synthesis one is in accordance to common encoding/decoding strategies that tend to pull the complexity towards the transmitter part of a system.

2 MULTIRATE REPRESENTATION

Subband processing assumes a mechanism that decomposes a given signal, say, $x(n)$ into P sequences, where $(P \geq 2)$, each of them contains different frequency components of the original signal $x(n)$. This is performed by digital filters, $h_0(n), h_1(n), \cdots, h_{P-1}(n)$, called decimators. In particular, the ith subband component is formed by passing the signal $x(n)$ through $h_i(n)$ and then subsampling by P. Thus, the decimation procedure is described as

$$y_i(n) = \sum_k h_i(Pn-k)x(k) \qquad (1)$$

where y_i is the output of the ith decimator as it is illustrated in Figure 1 in case $P = 2$.

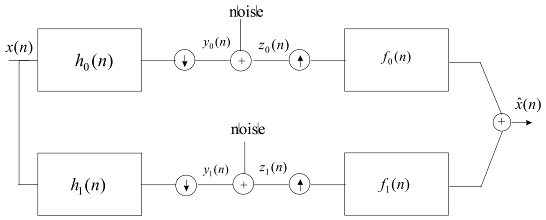

Figure 1: Decimator and interpolator filter banks with quantization noise

To reconstruct the signal $x(n)$ by its subband components, an inverse mechanism should be introduced using the so called $f_0(n), f_1(n), \cdots, f_{P-1}(n)$ interpolation filters. This is performed by inserting $P-1$ zeros between successive samples of the subband signal and then passing through the aforementioned interpolation filters. That is, the reconstructed signal $\hat{x}(n)$ of $x(n)$ can be derived by the following equation

$$\hat{x}(n) = \sum_{i=0}^{P-1} \sum_k f_i(n - Pk) y_i(k) \tag{2}$$

In case that appropriate decimation and interpolation filters are used we can guarantee perfect reconstruction i.e., $x(n) = \hat{x}(n)$. Since equations (1) is of convolution type, it can be written in a matrix form as follows

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x} \tag{3}$$

where $\mathbf{x} = [x(0), \cdots, x(N-1)]^T$ is a vector containing the N samples of the input $x(n)$ while $\mathbf{y}_i = [y(0), \cdots, y(N/2-1)]^T$. The $\mathbf{H}_i$ is a matrix the elements of which correspond to filter coefficients $h_i(n)$. That is,

$$\mathbf{H}_i(m,n) = h_i(Pm - n) \ m = 0, \cdots, N/P - 1 \\ \text{and } n = 0, \cdots, N-1 \tag{4}$$

Equation (3) is equivalent to (1) apart perhaps from the boundary of $\mathbf{y}_i$. However, for large number N this mismatching is neglitible. Let us denote as $\mathbf{y}$ a vector containing all subband components $\mathbf{y}_i$, $i = 0, 1, \cdots, P-1$, that is $\mathbf{y} = [\mathbf{y}_0^T, \cdots, \mathbf{y}_{P-1}^T]^T$. Then, based on (3) the vector $\mathbf{y}$ can be expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{x} \tag{5}$$

where matrix $\mathbf{H}$ is defined as

$$\mathbf{H} = [\,\mathbf{H}_0^T \cdots \mathbf{H}_{P-1}^T\,]^T \tag{6}$$

In a similar way, equation (2) can be expressed in a matrix form as follows

$$\hat{\mathbf{x}} = \mathbf{F}^T \mathbf{y} \tag{7}$$

where $\hat{\mathbf{x}}$ is a vector which corresponds to the reconstructed elements $\hat{x}(n)$, $n = 0, \cdots, N-1$ and matrix $\mathbf{F}$ is formed similarly to $\mathbf{H}$.

3 OPTIMAL DECIMATORS COMPENSATING FOR ADDITIVE NOISE

The use of perfect reconstruction filters, i.e., $\mathbf{F}^T\mathbf{H} = \mathbf{I}_N$, where $\mathbf{I}_N$ stands for the $N \times N$ identity matrix, guarantees that $x(n) = \hat{x}(n)$. However, in practice, there is a variety of reasons for deviation from this assumption. For example, a non linear quantization is usually performed in subband domain which disturbs signal $\mathbf{y}$. This means that the use of filters which satisfy the property $\mathbf{F}^T\mathbf{H} = \mathbf{I}_N$ does not deduct that the reconstructed signal $\hat{x}(n)$ is as much as possible close to $x(n)$.

3.1 Problem Formulation

In this paper a quantization noise is assumed that disturbs the signal $\mathbf{y}$. Let us denote as $\mathbf{z}$ the signal after the quantization. Then we can write that,

$$\mathbf{z} = Q(\mathbf{y}) \\ \text{with } Q(\mathbf{y}) = \mathbf{q}_i \ for \ \mathbf{q}_i \in R_i \tag{8}$$

R_i are regions of the N-dimensional space R^N while $\mathbf{q}_i$ correspond to the quantization levels. The set of R_i constitute a complete partition of the space R^N. Usually, we consider quantizers that use the same quantization decision levels for all samples of $\mathbf{y}$. In this most popular case R_i correspond to convex hyper cubes.

To optimally estimate the analysis filter $\mathbf{H}$ that compensates the effect of quantization noise a distortion criterion should be minimized with respect to the matrix $\mathbf{H}$. This criterion expresses the difference of the actual signal $\mathbf{x}$ and the reconstructed one $\hat{\mathbf{x}}$. The distortion criterion used in this paper is the following

$$D = \int_{\mathbf{x} \in R^N} d(\mathbf{x}, \hat{\mathbf{x}}) f_X(\mathbf{x}) d\mathbf{x} \tag{9}$$

where $f_X(\mathbf{x})$ is the joint probability density function (pdf) of the vector $\mathbf{x}$ and $d(\cdot)$ the distance between the two vectors, usually given by the Euclidean metric, $d(\mathbf{x}, \hat{\mathbf{x}}) = (\mathbf{x} - \hat{\mathbf{x}})^T \cdot (\mathbf{x} - \hat{\mathbf{x}})$. The integration in (9) is performed over the N-dimensional space R^N.

In the case of uniform distribution, the pdf factor of (9) reduces to an unimportant scalar, constant with respect to the matrix $\mathbf{H}$ and an analytical solution is possible. Since in most cases an appropriate transformation of the vector $\mathbf{x}$ prior to the analysis stage can be performed in order to convert the pdf of $\mathbf{x}$ to a uniform one, in the sequel we assume $f_X = 1$. As a result, the term corresponding to $f_X(\mathbf{x})$ can be omitted for the minimization procedure.

Equation (8) indicates that the noise, introduced by the quantizer, is correlated to the analysis filter. The optimal analysis matrix $\mathbf{H}$ is then derived by minimizing

(9). Based on both (5 and 8) and using the Euclidean metric for the distance $d(\cdot)$, eq. (9) can be expressed as follows

$$D = \int_{\mathbf{x}\in R^N} \left(\mathbf{x} - \mathbf{F}^T Q(\mathbf{H}\mathbf{x})\right)^T \cdot \left(\mathbf{x} - \mathbf{F}^T Q(\mathbf{H}\mathbf{x})\right) d\mathbf{x} \tag{10}$$

Using the partitions described in (8) and changing the variable of integration from $\mathbf{x}$ to $\mathbf{y}$, eq. (10) can be written as a sum of integrals over all partitions R_i. That is,

$$D = \sum_i \int_{\mathbf{y}\in R_i} \left(\mathbf{B}\mathbf{y} - \mathbf{F}^T\mathbf{q}_i\right)^T \cdot \left(\mathbf{B}\mathbf{y} - \mathbf{F}^T\mathbf{q}_i\right) |det(\mathbf{B})| d\mathbf{y} \tag{11}$$

provided that matrix $\mathbf{H}$ is of full rank and $\mathbf{B} = \mathbf{H}^{-1}$.

The optimal elements of $\mathbf{B}$ (and consequently of $\mathbf{H}$) are derived through the differentiation of (11) with respect to matrix $\mathbf{B}$. Minimization of (11) is in general involved for the optimal analysis filter design. It reduces, however, to a more tractable expression under the assumption that $\mathbf{H}$ is unitary i.e., under the constraint

$$\mathbf{H}^T\mathbf{H} = \mathbf{I} \tag{12}$$

3.2 Problem Solution

For the minimization of (11) subject to the constraint (12), the Lagrange multipliers are used. In particular, the solution is obtained through minimization of the following equation with respect to $\mathbf{B}$ and $\mathbf{\Lambda}$

$$L = D + tr\left(\mathbf{\Lambda}(\mathbf{B}^T\mathbf{B} - \mathbf{I})\right) \tag{13}$$

where $\mathbf{\Lambda}$ contains the elements of the lagrange multipliers and corresponds to a symmetric unknown matrix since the the constraint $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ is a set of symmetric equations. Estimation of the matrix $\mathbf{B}$ is performed by differizing (13) with respect to $\mathbf{B}$ and $\mathbf{\Lambda}$, that is $\partial L/\partial\mathbf{\Lambda} = 0$ and $\partial L/\partial\mathbf{B} = 0$. The first partial derivative concludes to the constraint (12) while the second one to

$$\frac{\partial L}{\partial \mathbf{B}} = \mathbf{A} + \mathbf{C} + 2\mathbf{B}\mathbf{\Lambda} \tag{14}$$

where matrices $\mathbf{A}$, $\mathbf{C}$ are expressed as

$$\mathbf{A} = |det(\mathbf{B})| \sum_i \int_{\mathbf{y}\in R_i} \left(2\mathbf{B}\mathbf{y}\mathbf{y}^T - 2\mathbf{F}^T\mathbf{q}_i\mathbf{y}^T\right) d\mathbf{y} \tag{15}$$

and

$$\mathbf{C} = \mathbf{B}^{-T}|det(\mathbf{B})| \sum_i \int_{\mathbf{y}\in R_i} ||\mathbf{B}\mathbf{y} - \mathbf{F}^T\mathbf{q}_i||_2 d\mathbf{y} \tag{16}$$

Matrices $\mathbf{A}$ and $\mathbf{C}$ of (14) involve integration over all quantization partitions R_i. Let us assume that the quantization levels are of hyper cubes form, since there is no reason to apply different quantization to some elements of $x(n)$. Then, $\int_{\mathbf{y}\in R_i} \mathbf{y}^T d\mathbf{y} = \mathbf{q}_i^T v_i$ where v_i stands for the hyper-volume of the ith partition. Consequently, the second term involved in matrix $\mathbf{A}$ can be written as

$$\sum_i \int_{\mathbf{y}\in R_i} -2\mathbf{F}^T\mathbf{q}_i\mathbf{y}^T d\mathbf{y} = -2\mathbf{F}^T \sum_i \mathbf{q}_i\mathbf{q}_i^T v_i \tag{17}$$

The $\sum_i \mathbf{q}_i\mathbf{q}_i^T v_i$ over all i can be expressed as a matrix, say, $\mathbf{Q}$.

$$\mathbf{Q} = \sum_i \mathbf{q}_i\mathbf{q}_i^T v_i \tag{18}$$

Thus, the RHS of eq. (17) can be written as $-2\mathbf{F}^T\mathbf{Q}$. It can be shown that all diagonal elements of $\mathbf{Q}$ are equal. As a result, matrix $\mathbf{Q}$ is characterized by only two different elements. If in addition the quantizer is symmetric w.r.t. the $\mathbf{Q} = \delta\mathbf{I}$.

The integral involved in the first term of $\mathbf{A}$ can be decomposed as $\int_{\mathbf{y}\in R_i} \mathbf{y}\mathbf{y}^T d\mathbf{y} = \mathbf{q}_i\mathbf{q}_i^T v_i + \text{diag}(\gamma_{i,k})v_i$ where $\text{diag}(\gamma_{i,k})$ represents a diagonal matrix with elements $\gamma_{i,k}$. The $\gamma_{i,k}$ corresponds to the quantization error of the kth element of the ith partition R_i. The index k takes value between 0 and $N-1$. It can be shown, using the quantization properties, that $\gamma_{i,k} = dx_{i,k}^2/12$, where $dx_{i,k}$ is the interval of the kth element of the ith partition. Based on the previous observations and since the first term of $\mathbf{A}$ involves summation over all partitions i, we have that

$$\sum_i \int_{\mathbf{y}\in R_i} 2\mathbf{B}\mathbf{y}\mathbf{y}^T d\mathbf{y} = 2\mathbf{B}\mathbf{Q} + 2\gamma\mathbf{B}\mathbf{I} \tag{19}$$

where the scalar $\gamma = \sum_i \gamma_{i,k} v_i$, is independent from k due to the quantization properties. Since $\mathbf{B}^T\mathbf{B} = \mathbf{I}$, $|det(\mathbf{B})| = 1$. Therefore, matrix $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = 2\mathbf{B}(\mathbf{Q} + 2\gamma\mathbf{I}) - 2\mathbf{F}^T\mathbf{Q} \tag{20}$$

In a similar way, we can show that the matrix $\mathbf{C}$ can be written as

$$\mathbf{C} = \mathbf{B}^{-T}\left(tr(\mathbf{Q} + \gamma\mathbf{I} + \mathbf{F}\mathbf{F}^T\mathbf{Q}) - 2tr(\mathbf{B}^T\mathbf{F}^T\mathbf{Q})\right) \tag{21}$$

Using the (20, and 21), eq. (14) transforms to

$$\mathbf{B}^T\mathbf{F}^T\mathbf{Q} + tr(\mathbf{B}^T\mathbf{F}^T\mathbf{Q}) = \mathbf{\Lambda} + \mathbf{Q} + \epsilon\mathbf{I} = \mathbf{\Lambda} + \mathbf{Q}' \tag{22}$$

where $\epsilon\mathbf{I} = \left(tr(\mathbf{Q} + \gamma\mathbf{I} + \mathbf{F}\mathbf{F}^T\mathbf{Q})/2 + \gamma\right)\mathbf{I}$ and $\mathbf{Q}' = \mathbf{Q} + \epsilon\mathbf{I}$.

Estimation of matrix $\mathbf{B}$ and consequently $\mathbf{H}$ is performed through equation (22) and the constraint (12). Let us first denote as $\mathbf{G}$ the matrix $\mathbf{B}^T\mathbf{F}^T\mathbf{Q}$, i.e., $\mathbf{G} = \mathbf{B}^T\mathbf{F}^T\mathbf{Q}$. Then, for a given quantizer and interpolation filter and a given matrix $\mathbf{\Lambda}$, there always exist only one matrix $\mathbf{G}$ and vice versa (22). That is, the non diagonal elements of $\mathbf{G}$, $g_{i,j}$ with $i \neq j$ are equal to $\lambda_{i,j} + q'_{i,j}$, while the diagonal ones of $\mathbf{G}$, say, $g_{i,i}$ are obtained by solving a linear system depending on the

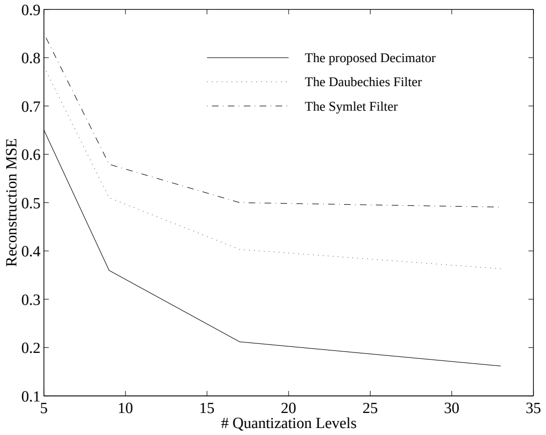

Figure 2: Reconstruction mean squared error versus number of quantization levels

elements $\lambda_{i,i} + q'_{i,i}$. The fact that matrices $\mathbf{\Lambda}$ and $\mathbf{Q}'$ are symmetric implies that matrix $\mathbf{G}$ is also symmetric. Then, due to eq. (12) matrix $\mathbf{G}$ is obtained via the following Ricatti equation

$$\mathbf{GRG} - \mathbf{I} = \mathbf{0} \tag{23}$$

where $\mathbf{R} = \left(\mathbf{F}^T\mathbf{Q}\right)^{-1}\left(\mathbf{F}^T\mathbf{Q}\right)^{-T}$. Matrix $\mathbf{R}$ is positive definite and thus there always exist one solution of (23). Having computed matrix $\mathbf{G}$ only one symmetric proper matrix $\mathbf{\Lambda}$ exists so that (22) is fulfilled. Matrix $\mathbf{H}$ which represents the decimator can be optimally obtained through the following equation

$$\mathbf{H} = \mathbf{G}^{-1}(\mathbf{FQ})^T \tag{24}$$

From the previous equation it is observed that in case that interpolating filters correspond to unitary $\mathbf{F}$, (i.e., $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ as in case of QMF filterbanks, then optimal decimator is the inverse matrix $\mathbf{HF}^{-T}$.

4 EXRERIMENTAL RESULTS

The performance of the proposed filters was tested using uniformly distributed random one-dimensional input signals $x(n)$. The quantizer is considered to be uniform while the number of quantization levels varies between successive experiments. The decimator filter, used in the process, has been derived by the biothogonal ones and it is assumed to be the same for all experiments while Several QMF filters have been used as decimators.

Figure 2 illustrates the mean squared error (MSE) of the original and the reconstructed signal at different quantization levels and for different QMF filters (Daubechies and Symlet). It is observed that the proposed method results in minimum MSE for a given quantizer (or equivalently a certain number of levels) for all unitary decimators filters.

5 CONCLUSIONS

In this paper, we derive an analytic method for obtaining optimal decimator filter which compensates the quantization noise. The method includes a minimization of a distortion criterion, possibly via solution of a Ricatti equation. The performance of the proposed technique is examined using uniformly distributed random input signal and quantizer with varying number of levels.

References

[1] A. N. Delopoulos and S. D. Kollias, "Optimal Filter Banks For Signal Reconstruction from Noisy Subband Components," *IEEE Trans. on Signal Processing,* vol., 44, no. 2, pp. 212-224, Feb. 1996.

[2] K. Gosse and P. Duhamel, "Perfect Reconstruction Versus MMSE Filter Banks in Source Coding," *IEEE Trans. Signal Processing,* vol. 45, no. 9, pp. 2188-2202.

[3] R. A. Haddad and N. Uzun, "Modeling, Analysis and Compensation of Quantization Effect in M-band Subband Codecs," *Proc. of ICASSP' 93,* Minneapolis, MN, vol. III, pp. 173-176.

[4] R. A. Haddad and K. Park, "Modeling, Analysis and Optimum Design of Quantized M-Bank Filter banks," *IEEE Trans. Signal Processing,* vol. 43, pp. 2540-2549, Nov. 1995.

[5] ISO/CD 11172-2, "Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbps," March 1991.

[6] J. Kovacevic, "Subband Coding Systems Incorporating Quantizer Models," *IEEE Trans. Image Processing,* vol. 4, pp.543-553, may 1995.

[7] A. Tabatabai, "Optimum Analysis/Synthesis Filter Banks Structures with Applications to Subband Coding Systems," *Proc. ISCAS,* Espoo, Finland, pp. 823-826, 1994.

[8] P. P. Vaidyanathan, "Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial," *Proceedings of the IEEE,* vol. 78, no. 1, January 1990.

[9] P. P. Vaidyanathan, "Theory of Optimal Orthonormal Filter Banks," *Proc. of ICASSP '96,* vol. III, Atlanta, GA, pp. 1487-1490.

[10] L. Vandendorpe, "Optimized Quantization for Image Subband Coding," *Signal Processing; Image Communication,* vol. 4, pp. 65-78, Nov. 1991.

[11] M. Veterli, "Multi-dimensional Sub-band Coding: Some Theory and Algorithms," *Signal Processing,* vol. 6, pp. 97-112, April 1984.