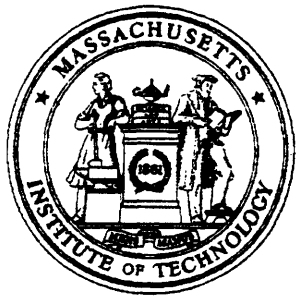

A Novel QMF Design Algorithm

RLE Technical Report No. 551

March 1990

Kambiz Casey Zangi

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

This work was supported by the Advanced Television Research Program and the Research Laboratory of Electronics at the Massachusetts Institute of Technology.

A Novel QMF Design Algorithm

by

Kambiz Casey Zangi

Submitted to the
Department of Electrical Engineering and Computer Science
on December 27, 1989 in partial fulfillment of the requirements for the Degree of
Master of Science in Electrical Engineering

Abstract. A new algorithm for designing two-band quadrature mirror filter (QMF) banks is presented in this thesis. It is shown that this algorithm is computationally more efficient than the existing algorithms by a factor of eight. Moreover, this algorithm is not sensitive to the choice of the initial guess. By using this algorithm, one can also design two-band QMF banks in which one of the filters is pre-described.

Thesis Supervisor: Dr. William F. Schreiber

Title: Professor of Electrical Engineering

Contents

Chapter 1

Introduction

Quadrature mirror filter (QMF) banks have been of great interest during the past decade, ever since their introduction by Croisier, Esteban, and Galand [1],[2]. These filters find applications where a discrete signal is to be split into a number of consecutive bands in the frequency domain, so that each subband signal can be processed in an independent manner. Typical processing includes undersampling the subband signals, encoding them and transmitting over a channel, or merely storing the coded signals. Eventually, at some point in the process, the subband signals should be recombined so that the original signal is properly reconstructed. Typical applications of such signal-splitting include subband coders for video signals [2],[3], digital transmultiplexers [4] used in FDM/TDM conversion, and frequency domain speech-scramblers.

In this thesis we look at the QMF design problem purely as a signal-reconstruction problem. The channel is therefore assumed to be noiseless, and the exact signal characteristics are not given. We treat the QMF design problem as a multivariable optimization over the filter coefficients. We further require all the filters to be linear phase and FIR, with even number of taps.

In this thesis we present a new algorithm for designing quadrature mirror filters in the time domain (called the jointly-optimal QMF design algorithm). Closed-form solutions for the optimal filters can not be derived, however the optimal analysis filters

for a given pair of synthesis filters can be analytically found as a solution to a system of linear equations. The computation of the optimal analysis filters for a given pair of synthesis filters is similarly possible. The joint optimization algorithm is an iterative procedure that alternates between computing a pair of optimal analysis filters and optimal synthesis filters, until the error reduction is negligible.

Using this algorithm, we would then design QMF banks in which one of the filters is pre-described. Such a QMF bank can not be designed by the existing algorithms, since they derive all the filters from the lowpass analysis filter. With these algorithms, fixing one of the filters would completely determine the QMF bank and no optimization can be done.

This thesis consists of five chapters. The first two chapters review the previous work done in the area of QMF design. Chapter three presents the jointly-optimal design algorithm in detail. Chapter four offers several design examples and compares our algorithm to the previous QMF design algorithms. Finally in chapter five, we give a short summary of the most important issues discussed in this thesis.

1.1 Decimators and Interpolators

The QMF bank is a multirate digital system. The term multirate signifies that the sampling rate is not constant throughout the system; there are decimators in the system which down-sample a sequence, and there are interpolators which perform up-sampling. Since decimators and interpolators are the building blocks of any multirate digital system, let us briefly review their characteristics [11].

Decimators

A two-fold decimator is shown in Figure 1.1 [5]. Its input sequence is $x(n)$, and the output sequence $y(n)$ is a compressed version of $x(n)$. More specifically, the output is obtained by retaining only those samples of $x(n)$ which occur at times that are

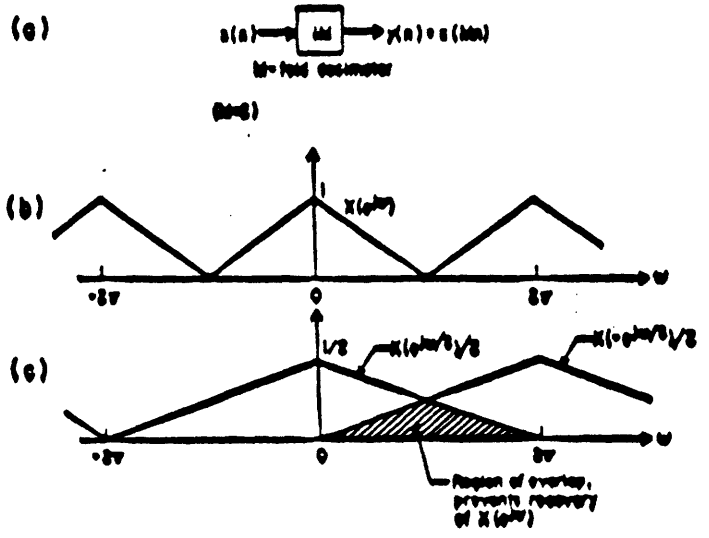

Figure 1.1: Two-fold decimator with a lowpass input.

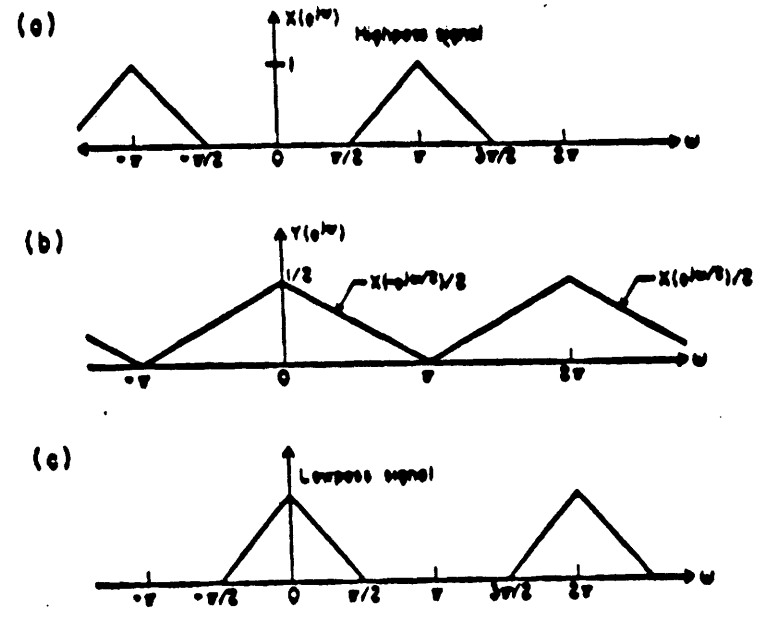

Figure 1.2: Two-fold decimator with a highpass input.

multiples of two. The input-output relation of this two-fold decimator is $y(n) = x(2n)$. Note, that the decimator is a linear, time-varying device. Accordingly, it can not be represented by a transfer function; although there are other methods to describe the decimator in the frequency domain. For a two-fold decimator, the quantity $X(e^{jw/2})$ is a stretched version of $X(e^{jw})$. Since $X(e^{jw/2})$ has a period of 4π rather than 2π, it is not a valid transform of a sequence. It can be verified that $Y(e^{jw})$ in fact has two terms. The first term is $X(e^{jw/2})$, and the second term is $X(e^{-jw/2})$ which is a shifted version of the first term (shifted by amount 2π). More formally, the input-output relation for a two-fold decimator can be written in the frequency domain as

$$Y(e^{jw}) = 0.5[X(e^{jw/2}) + X(-e^{jw/2})]. \tag{1.1}$$

Note that $Y(e^{jw})$ given as above does have a period of 2π. This is demonstrated in Fig. 1.1(b),(c) where $x(n)$ is assumed to be a lowpass signal. If the transform of $x(n)$ is not bandlimited to $-\pi/2 \leq w \leq \pi/2$, there is an overlap of the two terms in (1.1) as shown by the shaded area in Fig. 1.1(c). This overlap is the aliasing effect, caused by undersampling. There is no way we can get back the original signal $x(n)$ from $y(n)$, once aliasing has taken place.

What happens if a highpass signal is decimated by a factor of two? Since it is not bandlimited to $-\pi/2 \leq w \leq \pi/2$, the first impression is that there will be aliasing. Let us refer to Fig. 1.2. Assuming the highpass signal to be bandlimited to $\pi/2 \leq w \leq 3\pi/2$, the decimator's output $Y(e^{jw})$ is as shown in Fig. 1.2(b). Thus, the decimated version of the highpass signal looks like a lowpass signal. Notice that if a signal having a lowpass spectrum as in Fig. 1.2(c) were decimated, we have would obtained precisely the spectrum of Fig. 1.2(b); it is not possible to tell whether it came from Fig. 1.2(c) or Fig. 1.2(a). However, as long as we have the additional information as to whether $x(n)$ is lowpass or highpass signal, we can reconstruct the signal $x(n)$ from $y(n)$.

In practice, before a signal is passed through a decimator, it is first bandlimited by using a bandpass filter, to reduce the effects of aliasing. Such filters are called analysis filters or pre-filters in the QMF terminology.

Interpolators

A two-fold interpolator, schematically shown in Fig. 1.3, inserts a zero between adjacent samples of its input [5]. Its input-output relation is

$$y(n) = \begin{cases} x(n/2) & \text{if } n \text{ is even} \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

The effect of this stretching in the time domain is a compression in the frequency domain, as demonstrated by Fig. 1.3(b). Since $Y(e^{jw})$ has a replica (image) of the basic prototype spectrum, the interpolator is said to cause an imaging effect (which is the dual of the aliasing effect of a decimator). The frequency domain relation for a two-fold interpolator is

$$Y(e^{jw}) = X(e^{2jw}). \tag{1.3}$$

It can be verified through simple examples, that an interpolator is also a linear, time-varying system.

In practice, an interpolator is followed by a filter called the interpolation filter, which eliminates the images in Fig. 1.3 so that the final result is a simple bandpass signal (or lowpass signal if desired).

It is interesting to see what happens when a decimator and interpolator are cascaded, Fig. 1.4(a). The decimator causes stretching and aliasing, whereas the interpolator causes compression, all in the frequency domain. The end result is shown in Fig. 1.4(c). If the spectrum of $x(n)$ is bandlimited to $-\pi/2 \leq w \leq \pi/2$, then the output of the cascade is an imaged version of its input with no stretching or compression. A simple way to see that decimators and interpolators are time-varying systems is to observe that if the decimator and the interpolator are interchanged, the result will be an identity system; one which is obviously not equivalent to the original system. However, cascades of LTI systems can be interchanged without affecting the input-output relation. Therefore, decimators and interpolators must be time-varying systems.

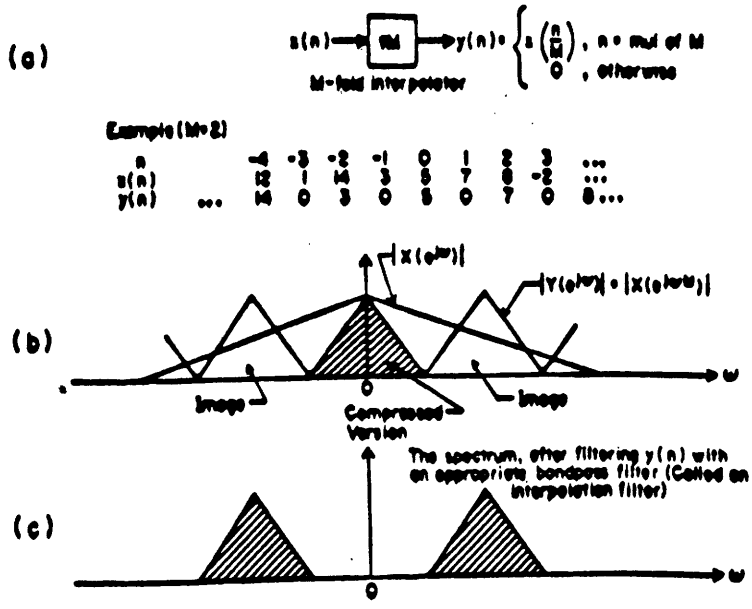

Figure 1.3: Two-fold interpolator with a lowpass input

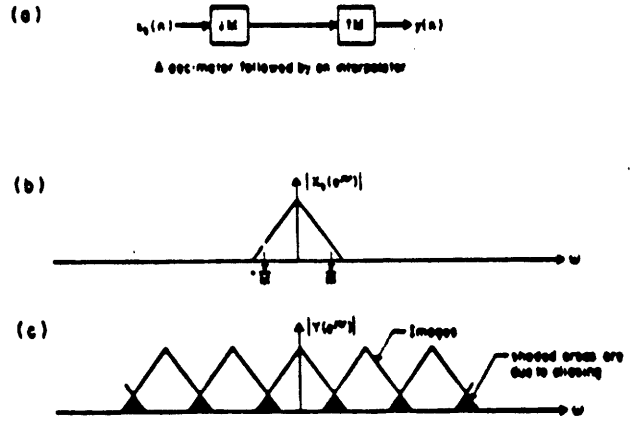

Figure 1.4: Cascade of a decimator and an interpolator

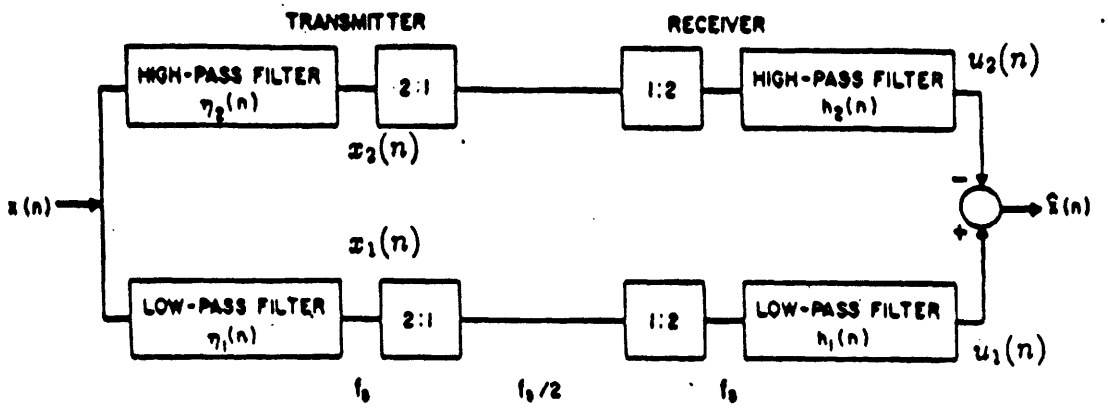

Figure 1.5: Two-Channel QMF bank

1.2 Two-Channel Quadrature Mirror Filter Bank

The two-channel QMF bank, shown in Fig. 1.5, is one of the earliest and most commonly used filter bank structures [5]. The analysis bank is composed of a lowpass filter η_1 and a highpass filter η_2, which split the input sequence $x(n)$ into two frequency bands. The lowpass signal $x_1(n)$ and the highpass signal $x_2(n)$ are then decimated by a factor of two. The decimated signals are typically coded and transmitted. At the the receiver, the signals are decoded, and passed through the interpolators. The decimator-interpolator cascade causes aliasing and imaging as discussed in the previous section; the purpose of the synthesis filters h_1 and h_2 is to eliminate these images. h_1 is a lowpass filter so that the highpass image of the interpolated lowpass signal $x_1(n)$ is suppressed. Similarly, $h_2(n)$ is a highpass filter so that the lowpass image of the interpolated highpass signal $x_2(n)$ is eliminated. As a result, the signals $u_1(n)$ and $u_2(n)$ are good approximations to $x_1(n)$ and $x_2(n)$, and the reconstructed signal $\hat{x}(n)$ hopefully resembles $x(n)$ closely.

What makes the QMF problem nontrivial and interesting? To avoid aliasing (due to decimation), the response of $N_1(e^{jw})$ and $N_2(e^{jw})$ must be disjoint. On the other hand, to have no frequency range left out in the reconstructed output, the responses

should be overlapping. The only obvious solution is to make the filter responses very sharp (approximately the ideal brickwall response). But it is well known that sharp filters require very high orders, are highly sensitive to quantization, and never exactly produce the ideal desired frequency response. Furthermore, pictures resulting from these filters have perceptually poor quality due to the ringing caused by these filters.

The philosophy adopted in the QMF bank to overcome this problem is to permit aliasing at the output of the decimators, by designing overlapping analysis filters, and then choosing the synthesis filters such that the imaging produced by the interpolators cancels the aliasing. In fact exact cancellation of aliasing is possible with the correct choice of filters.

Based on the relations developed previously for the two-fold decimators and interpolators, we can write

$$\begin{aligned} \hat{X}(e^{jw}) &= [N_1(e^{jw})H_1(e^{jw}) - N_2(e^{jw})H_2(e^{jw})]X(e^{jw}) + \\ &\quad [N_1(-e^{jw})H_1(e^{jw}) - N_2(-e^{jw})H_2(e^{jw})]X(-e^{jw}). \end{aligned} \tag{1.4}$$

Due to the second term in (1.4), the term involving $X(-e^{jw})$, we can not write down an expression for $\hat{X}(e^{jw})/X(e^{jw})$ that is independent of $X(e^{jw})$ itself. This is not surprising, since the QMF bank is a time-varying system. The second term in (1.4) represents the effects of aliasing and imaging. This term can be made to disappear simply by choosing the synthesis filters to be

$$H_1(e^{jw}) = N_2(-e^{jw}),\ H_2(e^{jw}) = N_1(-e^{jw}). \tag{1.5}$$

The time domain interpretation of (1.5) is that to eliminate aliasing completely we should choose the synthesis filters such that

$$h_1(n) = (-1)^n \eta_2(n),\ h_2(n) = (-1)^n \eta_1(n). \tag{1.6}$$

Note that to eliminate aliasing completely, the number of degrees of freedom in choosing the filters is reduced by a factor of two; choosing two filters would then

completely specify the QMF bank. If complete elimination of aliasing is not necessary, the designer can use these additional degrees of freedom to reduce some other unwanted characteristic of the QMF bank.

Once aliasing is canceled by the above choice of filters, the QMF bank becomes a linear,time-invariant system with transfer function

$$T(e^{jw}) = \frac{\hat{X}(e^{jw})}{X(e^{jw})} = [N_1(e^{jw})N_2(-e^{jw}) - N_2(e^{jw})N_1(-e^{jw})]. \tag{1.7}$$

Ideally, we would like $T(e^{jw}) = e^{-jwn_0}$ so that the reconstructed signal is a delayed version of $x(n)$. Since $T(n)$ in general is not a delay, it represents a distortion and is called the overall transfer function.

To reduce the number of variables involved in the problem, most QMF designers impose the additional constraint that synthesis filters and the analysis filters are the same. Obviously this would reduce the degrees of freedom in choosing the filters by another factor of two. These two constraints would then leave only one of the filters free to be chosen; choosing any of the four filters completely specifies the QMF bank in this case. These constraints lead to the structure shown in Fig. 1.6.

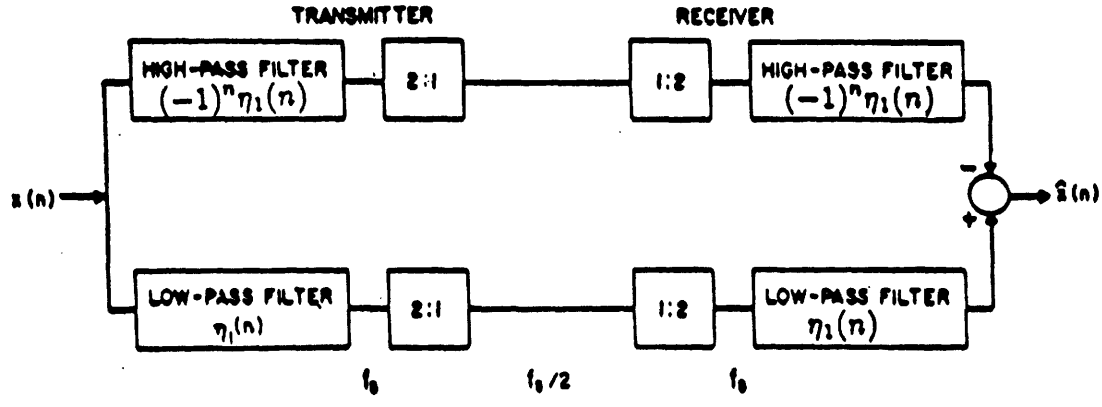

Figure 1.6: Alias-free two-Channel QMF bank

Elimination of aliasing and making the analysis filters the same as the synthesis filters would result in the following distortion function

$$T(e^{jw}) = N_1^2(e^{jw}) - N_1^2(e^{j(w-\pi)}). \tag{1.8}$$

Assuming that $N_1(e^{jw})$ is a linear phase filter, $T(e^{jw})$ given by (1.8) clearly has linear phase, and phase distortion is completely eliminated. By further assuming that $\eta_1(n)$ has N,even, number of taps, we can rewrite (1.8) as

$$T(e^{jw}) = e^{-jw(N-1)}[N_{1a}^2(e^{jw}) + N_{1a}^2(e^{j(w-\pi)}), \tag{1.9}$$

where $N_{1a}(e^{jw})$ is the real-valued amplitude response of $N_1(e^{jw})$.

Now comes the bad news: if two linear-phase transfer functions $H_1(e^{jw})$ and $H_2(e^{jw})$ are such that $|H_1(e^{jw})|^2 + |H_2(e^{jw})|^2$ is constant for all w, $H_1(e^{jw})$ and $H_2(e^{jw})$ must be trivial transfer functions [5] with frequency responses of the form $\cos^2(kw)$ and $\sin^2(kw)$. In other words, for the above choice of filters, there does not exist a non-trivial, linear-phase, FIR filter $\eta_1(n)$ which makes the overall response of the QMF bank distortionless. It is clear that the best we can do is to minimize the distortion in the QMF bank, since we can never completely eliminate it.

Chapter 2

Previous Work

This chapter reviews two papers which we consider representative of the previous work done in the area of quadrature mirror filter design. Both papers use optimization techniques to minimize the difference between the input and the reconstructed output in either the time domain or the frequency domain. Johnston uses a simple Hook and Jeaves optimization algorithm to minimize the distortion in the frequency domain [9]. On the other hand, Jain and Crochiere offer an iterative technique to minimize the distortion in the time domain [8]. Both papers assume FIR linear-phase filters with N, even, number of taps. They further assume that synthesis filters are the same as the analysis filters, and the high band filters are the mirror image of the low band filters. With these assumption, both techniques are guaranteed to have alias-free reconstruction at the cost of reducing the degrees of freedom by a factor of four.

2.1 Johnston's Work

Basic Assumptions and Definitions

Figure 2.1 shows the two-band QMF bank used by Johnston [9]. $\eta_1(n)$ is a linear-phase, lowpass, FIR filter with N, even, number of taps. All other filters are derived from

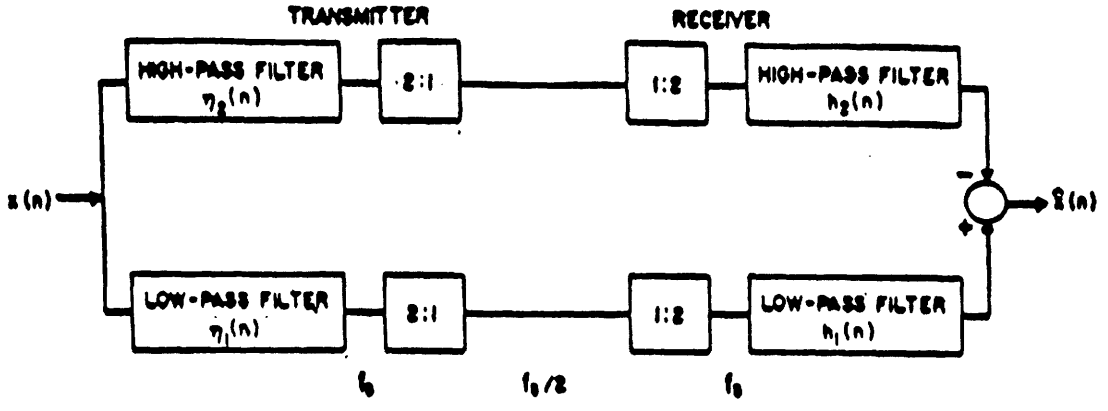

Figure 2.1: Two-channel QMF bank

$\eta_1(n)$ according to the following relations:

$$\eta_2(n) = (-1)^n \eta_1(n) \quad n = 0, \ldots, N-1 \tag{2.1}$$

$$h_1(n) = \eta_1(n) \quad n = 0, \ldots, N-1 \tag{2.2}$$

$$h_2(n) = (-1)^n \eta_1(n) \quad n = 0, \ldots, N-1. \tag{2.3}$$

The above relations guarantee that aliasing in the reconstructed output is completely eliminated. In absence of aliasing, the QMF bank becomes an LTI system characterized by its impulse response; that is,

$$\hat{x}(n) = t(n)x(n). \tag{2.4}$$

Optimization Criteria

Johnston's design algorithm minimizes a weighted sum of two errors. The first error corresponds to the flatness of $|T(e^{j\omega})|$, and is the determining factor in the accuracy of the reconstructed output. Another consideration in the QMF design is the frequency band separation of the filters. In subband coders, filtering is intended to split the input signal into various frequency bands so that each one is processed independently.

Therefore, the second error term is associated with the out of band rejection of the filters. Johnston constructed a metric which combines these two errors into a single scalar function of the filter taps.

The following scalar function is minimized by Johnston's design procedure

$$E = E_r + \alpha E_s(w_{sb}). \tag{2.5}$$

Corresponding to the flatness of the overall frequency response,

$$E_r = \sum_{k=0}^{N-1} [N_{1a}^2(e^{jw_k}) + N_{1a}^2(e^{j(w_k-\pi)}) - 1]^2, \tag{2.6}$$

where $N = 512$ and $w_k = \frac{2\pi}{N}k$. The user can specify the stopband frequency w_{sb} and the stopband error weight α. Measuring the out of band rejection of the filters,

$$E_s = \sum_{k=k_{sb}}^{N/2} N_{1a}^2(w_k), \tag{2.7}$$

where $N = 512$, $w_k = \frac{2\pi}{N}k$, and $k_{sb} = \frac{w_{sb}}{2\pi}N$. Johnston then used a simple optimization routine to minimize E as a function of the filter taps $\eta_1(n) \quad n = 0, \ldots, N-1$.

Hooks and Jeaves Optimization Technique

Having chosen the objective function to be minimized, Johnston then used the Hooks and Jeaves algorithm to carry out the optimization [9]. This algorithm requires a subroutine that takes the search variables as its inputs and returns the value of the objective function as its output [10]. Furthermore, it makes no assumption about the objective function to be minimized, and does not require calculation of the gradient of the objective function. This direct search method relies only on evaluating the objective function at a sequence of points to find a local minimum of the objective function.

To illustrate this technique, we shall consider the problem of minimizing $f(\vec{x})$. We choose an initial base point $\vec{b}_1$ and step length h_j for the respective variables x_j. The quantities

$$\left|f(\vec{b}_1 + h_j\vec{e}_j) - f(\vec{b}_1)\right| \tag{2.8}$$

are the magnitudes of the changes in $f(\vec{b}_1)$ due to change of one step length in each variable, where $\vec{e}_j$ is the unit vector in the direction of x_j. After $f(\vec{b}_1)$ has been evaluated, the method proceeds by a sequence of *exploratory* and *pattern moves*. If an exploratory move leads to a decrease in the value of $f(\vec{x})$, it is called a *success*; otherwise, it is called a *failure*. A pattern move is not tested for success or failure.

Exploratory Moves

The purpose of an exploratory move is to acquire information about $f(\vec{x})$ in the neighborhood of the current base point. The procedure for an exploratory move about the base point $\vec{b}_1$ is as follows:

1. Evaluate $f(\vec{b}1 + h_1\vec{e}_1)$. If the move from $\vec{b}_1$ to $[\vec{b}_1 + h_1\vec{e}_1]$ is a success, replace the base point $\vec{b}_1$ by $[\vec{b}_1 + h_1\vec{e}_1]$. If it is a failure, evaluate $f(\vec{b}_1 - h_1\vec{e}_1)$. If this is a success, replace $\vec{b}_1$ by $[\vec{b}_1 - h_1\vec{e}_1]$. If it is another failure, retain the original base point $\vec{b}_1$.

2. Repeat step (1) for the variable x_2 by considering variations $\pm h_2\vec{e}_2$ from the point which results from step (1). Apply this procedure to each variable in turn.

3. If $\vec{b}_2 = \vec{b}_1$, halve each of the step lengths h_j and return to step (1). If $\vec{b}_2 \neq \vec{b}_1$, make a pattern move from $\vec{b}_2$.

Pattern Moves and Subsequent Moves

A pattern move attempts to speed up the search by using the information already acquired about $f(\vec{x})$ by the exploratory moves. It is invariably followed by a sequence

of exploratory moves, with a view to find an improved direction of search in which to make another pattern move. We shall denote by $\vec{p}_1, \vec{p}_2, \ldots$ the points reached by successive pattern moves. It seems sensible to move from $\vec{b}_2$ in the direction $[\vec{b}_2 - \vec{b}_1]$, since a move in this direction has lead to a decrease in the value of $f(\vec{x})$. The procedure for a pattern move from $\vec{b}_2$ is therefore as follows:

1. Move from $\vec{b}_2$ to $\vec{p}_1 = [2\vec{b}_2 - \vec{b}_1]$ and continue with a new sequence of exploratory moves about $\vec{p}_1$.

2. If the lowest function value obtained during the pattern and exploratory moves is less than $f(\vec{b}_2)$, then a new base point $\vec{b}_2$ has been reached. In this case, return to exploratory moves. Otherwise, abandon the pattern move from $\vec{b}_2$ and continue with a new sequence of exploratory moves about $\vec{b}_2$.

2.2 Critique of Johnston's Work

The two error terms defined by Johnston are merely approximations. The exact expression for E_r, the reconstruction error, is

$$E_r = \int_{w=0}^{\pi} [N_{1a}^2(w) - N_{1a}^2(w-\pi) - 1]^2 dw. \tag{2.9}$$

The expression used by Johnston approximates the above integral by a sum. Similarly, the exact expression for E_s, the stopband error, is

$$E_s = \int_{w_{sb}}^{\pi} N_{1a}^2(w) dw. \tag{2.10}$$

To closely approximate the integrals in (2.9) and (2.10), the sums must be taken over a large number of terms. However, summing over a large number of terms in each iteration makes the algorithm very slow to execute.

Johnston's method reduces the degrees of freedom in choosing the filters by a factor of four to eliminate aliasing; however, it was shown in the chapter 1 that a reduction by

a factor of two is sufficient to do this. These additional degrees of freedom can be used to achieve the same performance with fewer filter taps, or to achieve better performance with the same number of filter taps. A more flexible design algorithm would allow the designer to trade off the aliasing error against other errors, by removing the constraint that aliasing must be completely eliminated.

Short filters obtained by Johnston have very poor performance. The reconstruction error, as measured by the ripple in the overall frequency response, is quite large for these filters. Moreover, these filters offer very little frequency band separation. Short filters are necessary where time localization is important, such as in image processing. The large reconstruction error and stopband error of these filters make them useless for such applications.

There are several problems associated with using the Hooks and Jeaves optimization routine. The starting point of the search and the initial step size used are critical to the success of this routine. A starting point that is too good may not minimize well, because it will get trapped in a local minimum. If the step size is initially set large enough to avoid these local minima, it is usually so large that the search will not find any better minima and terminate. Like all other direct-search algorithms, this one does not guarantee termination at the global minimum. Manual intervention in form of different starting points and increments is essential, and so is careful observation of the rum-time output results.

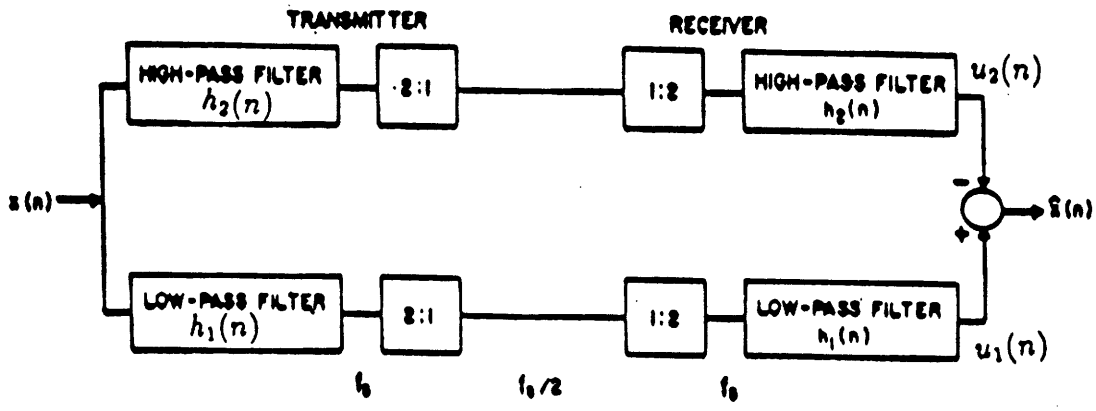

Figure 2.2: Two-channel QMF bank, $\hat{x}(n) = g(n)x(n)$

2.3 Jain and Crochiere's Work

Jain and Crochiere formulated the QMF design problem in the time domain, and showed that it results in an optimization problem requiring minimization of a quadratic multinomial [8]. They suggested an iterative solution which involves computation of eigenvectors of a matrix with dimensionality equal to one half of the number of filter taps in each iteration. In each iteration, the optimal synthesis filters for a given pair of analysis filters are found. The authors further claim that this technique converges to the solution independent of the initial guess. As in Johnston's technique, the user can specify the stopband edge frequency,the relative weights of the reconstruction error and the stopband error, and the number of filter taps. A weighted sum of the reconstruction error and the stopband error is then minimized using the iterative technique mentioned above.

Basic Assumptions and Definitions

Figure 2.2 shows the QMF bank used by Jain and Crochiere [8]. To obtain the aliasing cancellation property in the QMF bank, filters h_1 and h_2 are respectively symmetrical

and antisymmetrical with N,even, number of taps. The symmetry and antisymmetry properties imply that

$$h_1(n) = h_1(N-1-n) \quad n = 0, \ldots, \tfrac{N}{2} - 1 \tag{2.11}$$

$$h_2(n) = -h_2(N-1-n) \quad n = 0, \ldots, \tfrac{N}{2} - 1 \tag{2.12}$$

The aliasing cancellation further requires that the filters in Fig. 2.2 satisfy the following condition

$$h_2(n) = (-1)^n h_1(n) \quad n = 0, \ldots, N-1, \tag{2.13}$$

which is the mirror image relationship of the filters. With the above constraints, the aliasing cancellation property of the QMF bank is guaranteed, and the QMF bank becomes an LTI system (characterized by an impulse response).

Time Domain Analysis of the QMF Bank

To carry out the design, it is necessary to find the input-output relation of the QMF bank in the time domain. The reconstructed signal $\hat{x}(n)$ is obtained as the difference $u_1(n) - u_2(n)$ which can be written as

$$\hat{x}(n) = \tilde{x}(n) + \bar{x}(n), \tag{2.14}$$

where

$$\tilde{x}(n) = (-1)^n \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} h(i)h(k)x(n-k-i)[(-1)^k - (-1)^i] \tag{2.15}$$

$$\bar{x}(n) = \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} h(i)h(k)x(n-k-i)[1 - (-1)^{i+k}]. \tag{2.16}$$

Note that for simplicity of notation we have let $h(n) = h_1(n)$. Now, we observe that $\tilde{x}(n) \equiv 0$. This is so because whenever $i = k$, the quantity within the summation sign

is zero, and for $i \neq k$ the quantity within the summation sign is exactly the negative of that obtained for the transposed indexes. On the other hand, it can be shown that

$$\bar{x}(n) = \sum_{m=0}^{2N-2} \sum_{k=0}^{N-1} h(k)h(m-k)[1-(-1)^m]x(n-m) \tag{2.17}$$

$$= \sum_{m=0}^{2N-2} g(m)x(n-m), \tag{2.18}$$

where

$$g(m) = [1-(-1)^m] \sum_{k=0}^{N-1} h(k)h(m-k). \tag{2.19}$$

Then, it is clear that $g(m)$ is the impulse response of the QMF bank. Note that $\tilde{x}(n) \equiv 0$ is the time domain requirement for alias-free reconstruction.

Optimization Criteria

This design algorithm minimizes a weighted sum of reconstruction error and the stop-band error. In absence of aliasing, a necessary and sufficient condition on for perfect reconstruction is that $g(n)$ be a delta function . From (2.19) we see that $g(N-1)$ equals the energy of the filter taps (and thus cannot be be zero). Therefore, we require that $g(n) = \delta[n-N+1]$. This condition can be restated in two parts:

$$E_r = \sum_{k=0, k\neq N-1}^{2N-2} g^2(k) \tag{2.20}$$

$$g(N-1) = 1 \tag{2.21}$$

Another consideration in the QMF design is the out-of-band rejection of the the analysis filters. In subband coders, filtering is intended to split the input signal into various frequency bands for independent processing. E_{sb}, the stopband energy error, is defined as

$$E_{sb} = \int_{w_{sb}}^{\pi} \left|H(e^{jw})\right|^2 dw, \tag{2.22}$$

where $H(e^{jw})$ is the Fourier transform of $h(n)$. Substituting the definition of the Fourier transform in (2.22) we get

$$E_{sb} = \frac{1}{\pi} \sum_{i=0}^{M-1} \sum_{k=0}^{M-1} 2h(i)h(k). \int_{w_{sb}}^{\pi} [\cos(i-k)w + \cos(i+k-N+1)w]dw \qquad (2.23)$$

This is a quadratic form and can be more elegantly written using vector matrix notation as

$$E_{sb} = 2\vec{w}^T F \vec{w}, \qquad (2.24)$$

where $\vec{w}(n) = h(2n) \quad n = 0, \ldots, n = M-1$, and the i,k entry of the $M \times M$ matrix F is

$$f(i,k) = \frac{1}{\pi} \left\{ \begin{array}{ll} \pi - w_{sb} & i = k \\ \frac{\sin(i-k)w_{sb}}{(i-k)} & i \neq k \end{array} \right\} + \frac{1}{\pi} \left\{ \begin{array}{ll} \pi - w_{sb} & i+k = N-1 \\ \frac{\sin(i+k-N+1)w_{sb}}{(i+k-N+1)} & i+k \neq N-1 \end{array} \right\}. \qquad (2.25)$$

Note that for a specific stopband frequency w_{sb}, the matrix F can be pre-computed and used throughout the iteration process. The design problem can now be posed as: minimize E,

$$E = E_r + \alpha E_{sb}, \qquad (2.26)$$

subject to the constraint that

$$\vec{w}^T \vec{w} = \frac{1}{2}, \qquad (2.27)$$

where E_r is the reconstruction error, and E_{sb} is the stopband error. The user can specify the stopband edge w_{sb}, the weight-factor α, and of course the number of filter taps N.

Optimization Technique

Having chosen the objective function to be minimized, Jain and Crochiere developed an iterative algorithm to carry out the optimization. Closed-form solutions for all four optimal filters cannot be derived; however, the optimal synthesis filters for a given pair

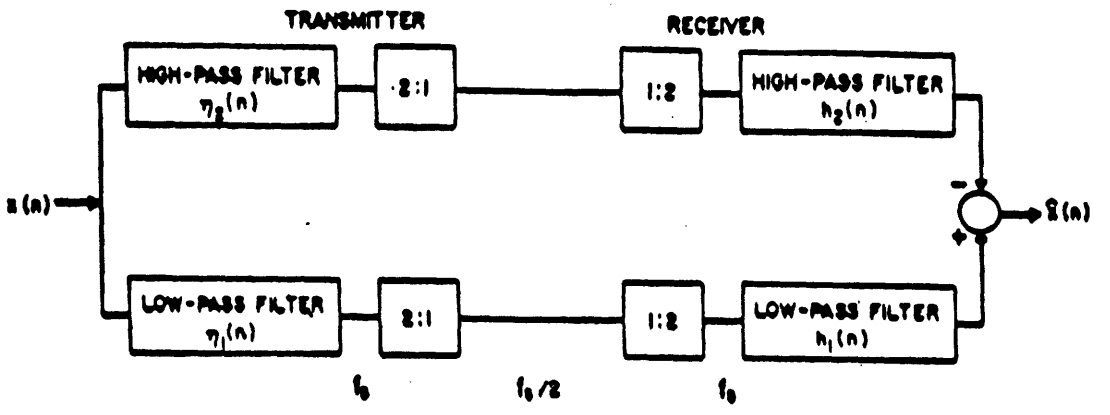

Figure 2.3: Two-channel QMF bank used by Jain and Crochiere

of analysis filters can be analytically found and vice versa. Therefore, their algorithm is an iterative procedure that first finds the optimal synthesis filters for a given pair of analysis filters and then replaces the old analysis filters by these newly found synthesis filters and proceeds to find another set of optimal synthesis filters. This procedure is repeated until the error is below a certain threshold.

The Optimum Synthesis Filters

To develop an iterative algorithm, it is reasonable to approximate the system of Fig. 2.2 by that of Fig. 2.3. Notice that in Fig. 2.2 the analysis and synthesis filters are the same, but in Fig. 2.3 they are not. In Fig. 2.3, we have used the filters $\eta_1(n) = \eta(n)$ and $\eta_2(n) = (-1)^n\eta(n)$ in the analysis bank. Similarly in the synthesis bank, $h_1(n) = h(n)$ and $h_2(n) = (-1)^n h(n)$. It is very important to keep in mind that $h(n)$ is yet to be determined.

Next, we need to express E_r and E_{sb} in terms of the unknown filters taps $h(n)$ $n = 0, \ldots, N-1$. Ignoring aliasing, the impulse response of the filter bank of Fig. 2.3 is

$$g(n) = \begin{cases} 0 & \text{if } n \text{ is even} \\ \sum_{k=0}^{N-1} \eta(n-k)h(k) & \text{if } n \text{ is odd .} \end{cases} \tag{2.28}$$

Indeed, it can be shown that

$$g(2n-1) = \sum_{k=0}^{M-1} a(n,k)\vec{w}(k), \tag{2.29}$$

where $M = \frac{N}{2}$, and

$$a(n,k) = u(n-1-k)\gamma(n-k) + u(n+k-M)\beta(n+k-M). \tag{2.30}$$

Here, u(.) denotes the unit step sequence, and the following M-length sequences are derived from the fixed analysis filters:

$$\gamma(n) = \eta(2n) \quad n = 0,\ldots,M-1 \tag{2.31}$$

$$\beta(n) = \eta(2n+1) \quad n = 0,\ldots,M-1 \tag{2.32}$$

$$\vec{w}(n) = h(2n) \quad n = 0,\ldots,M-1. \tag{2.33}$$

As expected from the setup of Fig. 2.3, the impulse response $g(n)$ is linear in the unknowns $\vec{w}(n) \quad n = 0,\ldots,M-1$. We can now compute E_r,the reconstruction error, as

$$E_r = \sum_{k=0,k\neq N-1}^{2N-1} g^2(k) \tag{2.34}$$

$$= 2\sum_{k=1}^{M-1} g^2(2k-1) \tag{2.35}$$

$$= 2\sum_{n=0}^{M-1}\sum_{l=0}^{M-1}\sum_{k=0}^{M-1} \vec{w}(i)a(n,k)\vec{w}(k) \tag{2.36}$$

$$= 2\vec{w}^T A\vec{w}, \tag{2.37}$$

where the i,k entry of the $M \times M$ matrix A is $\sum_{n=0}^{M-1} a(n,i)a(n,k)$. Recall that $a(n,k)$ was defined in (2.30), and $M = \frac{N}{2}$. On the other hand, since

$$g(N-1) = \sum_{n=0}^{N-1} h^2(n) \tag{2.38}$$

$$= 2\vec{w}^T\vec{w}, \tag{2.39}$$

the constraint of (2.21) can be rewritten as

$$\vec{w}^T\vec{w} = \frac{1}{2}. \tag{2.40}$$

Next, we need to calculate E_{sb}, the stopband error, in terms of $h(n)$. This was already done in (2.12);that is,

$$E_{sb} = 2\vec{w}^T F\vec{w}, \tag{2.41}$$

where the i,k entry of the $M \times M$ matrix F is

$$f(i,k) = \frac{1}{\pi}\left\{ \begin{array}{ll} \pi - w_{sb} & i = k \\ \frac{\sin(i-k)w_{sb}}{(i-k)} & i \neq k \end{array} \right\} + \frac{1}{\pi}\left\{ \begin{array}{ll} \pi - w_{sb} & i+k = N-1 \\ \frac{\sin(i+k-N+1)w_{sb}}{(i+k-N+1)} & i+k \neq N-1 \end{array} \right\}. \tag{2.42}$$

The problem of finding the optimum synthesis filters can now be posed (using the approximate system of Fig. 2.3) as follows: find $\vec{w}(n)$ $n = 0,\ldots,M-1$ which minimizes

$$\begin{aligned} E &= E_r + \alpha E_{sb} && (2.43)\\ &= 2\vec{w}^T A\vec{w} + 2\vec{w}^T \alpha F\vec{w} && (2.44)\\ &= 2\vec{w}^T (A + \alpha F)\vec{w} && (2.45) \end{aligned}$$

subject to the constraint

$$\vec{w}^T\vec{w} = \frac{1}{2}. \tag{2.46}$$

The solution to this type of problem, by using a Lagrange multiplier, is well known. This solution is given by the eigenvector of $(A + \alpha F)$ corresponding to its minimum eigenvalue λ_{min}, where the length is constrained to be $\frac{1}{\sqrt{2}}$. Also it can be shown that the minimum value of $(E_r + \alpha E_{sb})$ equals λ_{min}.

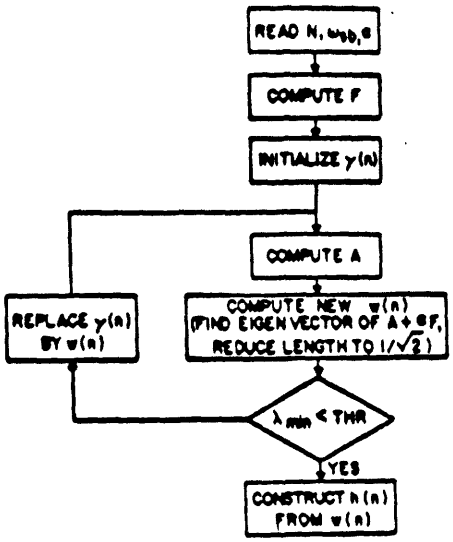

Figure 2.4: Iterative solution of the filter taps optimization.

Iteration Cycle

The iteration process is depicted in Fig. 2.4 [8]. Recall that the following M-length sequences are derived from the fixed analysis filters:

$$\gamma(n) = \eta(2n) \quad n = 0, \ldots, M-1 \tag{2.47}$$

$$\beta(n) = \eta(2n+1) \quad n = 0, \ldots, M-1. \tag{2.48}$$

Starting with a given pair of analysis filters $\eta(n)$ and $(-1)^n\eta(n)$, the optimization procedure aims at finding a pair of synthesis filters $h(n)$ and $(-1)^n h(n)$ which minimizes the error measure E, but no constraints are imposed for aliasing cancellation. In the next iteration, the old analysis filters are replaced by these newly found synthesis filters, and again an optimum pair of synthesis filters is found. Continuing this procedure, we hope that in the limit of convergence the analysis and synthesis filters are equal, and therefore aliasing is automatically canceled.

2.4 Critique of Jain and Crochiere's Work

There are two problems with the Jain and Crochiere's approach.Recall that aliasing and non-flat frequency response are the two sources of distortion in the QMF bank. Jain and Crochiere's algorithm only minimizes the reconstruction error in each iteration. The implicit assumption made is that minimizing the reconstruction error would result in filters which guarantee alias-free reconstruction. This assumption is not obviously valid, and no justification for it is given by Jain and Crochiere. A more intelligent algorithm would introduce another error term corresponding to aliasing and would minimize it in each iteration as well.

To circumvent the above problem with aliasing cancellation, Jain and Crochiere simply make the final analysis filters the same as the synthesis filters obtained in the last iteration. But, the resulting QMF bank is then not guaranteed to have the same distortion as the QMF bank obtained in the last iteration. Moreover, such elimination of aliasing reduces the degrees of freedom in choosing the filters by a factor of four. But it was shown in chapter 1 that a reduction by a factor of two is sufficient to achieve alias-free reconstruction. A more flexible algorithm would allow the designer to trade off the aliasing error against other errors, by removing the constraint that aliasing must be completely eliminated.

Chapter 3

Jointly-Optimal Analysis and Synthesis QMF Design

The design of jointly-optimal FIR analysis and synthesis QMF filters is the main focus of this chapter, with the distortion measure to be minimized being a weighted mean-square difference between the input and the reconstructed output. Unlike other design techniques, this algorithm assumes no relationship between the four filters, therefore it has more degrees of freedom in choosing the filter coefficients. The input signal to the QMF bank is a discrete-time signal, and the channel is assumed to be noiseless. All filters are FIR with even number of taps. Furthermore, the two lowpass filters are symmetric, and the two highpass filters are antisymmetric.

Closed-form solutions for the optimal filters cannot be derived; however, the optimal analysis filters for a given pair of synthesis filters can be analytically found as a solution to a system of linear equations whose coefficients are easily determined. The computation of the optimal analysis filters for a given pair of synthesis filters is similarly possible. The joint optimization algorithm is an iterative procedure that alternates between computing a pair of optimal analysis filters and optimal synthesis filters, until the error reduction is negligible.

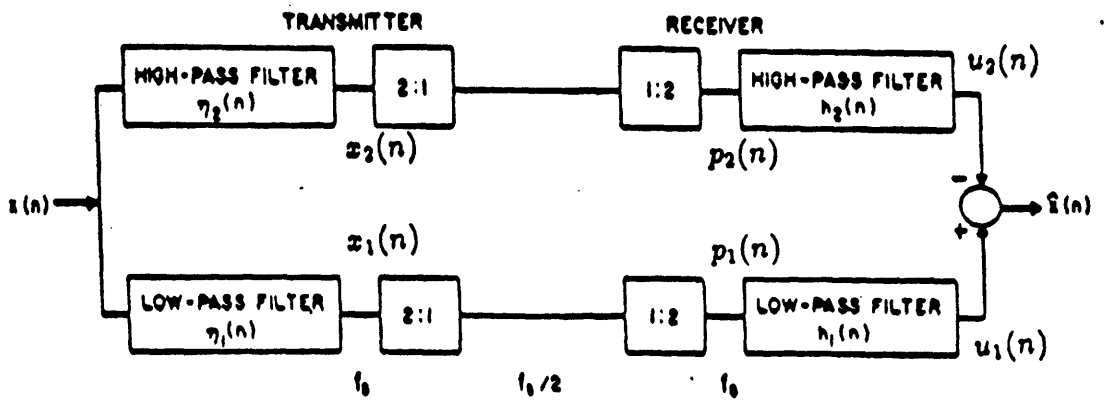

Figure 3.1: Two-channel QMF bank used for the jointly-optimal design.

The optimization problem is formulated as a minimization of a scalar function of filter coefficients $\eta_1(n)$, $\eta_2(n)$, $h_1(n)$, and $h_2(n)$, but it is virtually impossible to analyze such issues as convexity and convergence since the error is a quartic form. However, if we fix the analysis (synthesis) filter coefficients, then the error is a quadratic form on the synthesis (analysis) filter coefficients, which is much easier to study. Thus, we shall follow the following route toward derivation of the jointly-optimal filters: obtaining first independent solutions for analysis and synthesis filters, and then combining them in an iterative procedure that computes the jointly-optimal filters.

This chapter consists of four sections. The analysis of the QMF bank in the time domain is discussed in the first section. Derivation of the optimal analysis filters for a given pair of synthesis filters is presented in the second section. Derivation of the optimal synthesis filters for a given pair of analysis filters is discussed in the third section. Finally, the joint optimization of the analysis and synthesis filters is presented in the fourth section, combining the results of the previous sections.

3.1 Time Domain Analysis of the QMF Bank

We shall consider the system depicted in Fig. 3.1. The filters η_1, η_2, h_1, and h_2 are FIR with N,even, number of taps. The symmetry and antisymmetry properties imply the followings:

$$\eta_1(n) = \eta_1(N-1-n) \;\; n = 0, \ldots, N \tag{3.1}$$

$$h_1(n) = h_1(N-1-n) \;\; n = 0, \ldots, N \tag{3.2}$$

$$\eta_2(n) = -\eta_2(N-1-n) \;\; n = 0, \ldots, N \tag{3.3}$$

$$h_2(n) = -h_2(N-1-n) \;\; n = 0, \ldots, N. \tag{3.4}$$

Deriving an input-output relation is the first step in analyzing the QMF bank. From the lower branch of the QMF bank in Fig. 3.1, we can write

$$x_1(n) = \sum_{i=0}^{N-1} \eta_1(i) x(n-i) \tag{3.5}$$

$$p_1(n) = 0.5 \sum_{i=0}^{N-1} \eta_1(i) x(n-i)[1 + (-1)^n] \tag{3.6}$$

$$u_1(n) = 0.5 \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} \eta_1(i) h_1(k) x(n-i-k)[1 - (-1)^{n-k}]. \tag{3.7}$$

Similarly, from the upper branch we obtain an expression for $u_2(n)$ except that n_1 is replaced by n_2, and h_1 is replaced by h_2 in (3.7) ,that is,

$$u_2(n) = 0.5 \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} \eta_2(i) h_2(k) x(n-i-k)[1 - (-1)^{n-k}]. \tag{3.8}$$

Then, the reconstructed signal $\hat{x}(n)$ is obtained as the difference $2[u_1(n) - u_2(n)]$ which can be written as a sum of two terms:

$$\hat{x}(n) = \bar{x}(n) + \tilde{x}(n), \tag{3.9}$$

where

$$\bar{x}(n) = \sum_{m=0}^{2N-2} \left\{ \sum_{k=0}^{N-1} h_1(k)\eta_1(m-k) - \sum_{k=0}^{N-1} h_2(k)\eta_2(m-k) \right\} x(n-m) \tag{3.10}$$

$$\tilde{x}(n) = -(-1)^n \sum_{m=0}^{2N-2} \left\{ \sum_{k=0}^{N-1} (-1)^k h_1(k)n_1(m-k) - \sum_{k=0}^{N-1} (-1)^k h_2(k)\eta_2(m-k) \right\} x(n-m). \tag{3.11}$$

Much insight can be gained by making the following substitutions in (3.10) and (3.11) :

$$h_1'(k) = (-1)^k h_1(k) \quad k = 0, \ldots, N-1 \tag{3.12}$$

$$h_2'(k) = (-1)^k h_2(k) \quad k = 0, \ldots, N-1 \tag{3.13}$$

$$g_1(m) = \sum_{k=0}^{N-1} h_1(k)\eta_1(m-k) \quad m = 0, \ldots, 2N-2 \tag{3.14}$$

$$g_1'(m) = \sum_{k=0}^{N-1} h_1'(k)\eta_1(m-k) \quad m = 0, \ldots, 2N-2 \tag{3.15}$$

$$g_2(m) = \sum_{k=0}^{N-1} h_2(k)\eta_2(m-k) \quad m = 0, \ldots, 2N-2 \tag{3.16}$$

$$g_2'(m) = \sum_{k=0}^{N-1} h_2'(k)\eta_2(m-k) \quad m = 0, \ldots, 2N-2. \tag{3.17}$$

By substituting (3.12)-(3.17) in (3.10) and (3.11), we can rewrite the expressions for $\hat{x}(n)$ and $\bar{x}(n)$ in the following way:

$$\bar{x}(n) = \sum_{m=0}^{2N-2} [g_1(m) - g_2(m)]x(n-m) \tag{3.18}$$

$$\tilde{x}(n) = (-1)^n \sum_{m=0}^{2N-2} [g_1'(m) - g_2'(m)]x(n-m). \tag{3.19}$$

By substituting (3.18) and (3.19) in (3.9), we can also rewrite the reconstructed output as

$$\hat{x}(n) = \sum_{m=0}^{2N-2} [g_1(m) - g_2(m)]x(n-m) - (-1)^n \sum_{m=0}^{2N-2} [g_1'(m) - g_2'(m)]x(n-m). \tag{3.20}$$

Expression (3.20) is the desired input-output relation which expresses the output signal , $\hat{x}(n)$, in terms of the input signal, $x(n)$, and the filter coefficients.

The second term in (3.20), containing $(-1)^n$, is the aliased component of the reconstructed signal. If this term was equal to zero for all n, the reconstructed signal would be given by $\sum_{m=0}^{2N-2}[g_1(m) - g_2(m)]x(n-m)$. In this case, $\hat{x}(n)$ is simply the result of convolving the input signal with an impulse response equal to $[g_1(m) - g_2(m)]$. Being an output of an LTI system, the reconstructed signal $\hat{x}(n)$ is obviously alias-free in this case.

3.2 Errors to be Minimized

There are four error terms that are minimized by the optimization technique presented in this chapter. Assuming that aliasing is eliminated, the first error term, e_1, measures the flatness of the frequency response of the QMF bank. The second error term, e_2, measures the degree of aliasing in the reconstructed output. The third error term, e_3, corresponds to the frequency band separation of the analysis filters; in other words, it measures the energy compaction of the filters in the frequency domain. The last error term, e_4, corresponds to the flatness of the frequency response of the individual filters in their passbands.

As mentioned before, it is desirable to make the first term in (3.20) equal to a unit impulse at $n = N-1$; that is,

$$[g_1(n) - g_2(n)] = \begin{cases} 1 & n = N-1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.21}$$

This would in turn result in a flat frequency response in the absence of aliasing. We can guarantee that $g_1(N-1) - g_2(N-1) = 1$ is always satisfied, by repharsing this optimization problem as a constrained optimization. In other words, the problem is now to minimize the reconstruction error under the constraint that $g_1(N-1) - g_2(N-1) = 1$.

The expression for e_1 in this case is

$$e_1 = \sum_{n=0, n\neq N-1}^{2N-2} [g_1(n) - g_2(n)]^2. \tag{3.22}$$

As mentioned before, aliasing is eliminated whenever the second term in (3.20) equals zero. Therefore, it is reasonable to set e_2 equal to the energy of this term, that is,

$$e_2 = \sum_{n=0}^{2N-2} [g_1'(n) - g_2'(n)]^2. \tag{3.23}$$

The third error term, e_3, corresponds to the frequency band separation of the filters. More specifically, it is desirable to make $\eta_1(n)$ a lowpass filter and $\eta_2(n)$ a highpass filter with each having as little stopband energy as possible. Therefore, a reasonable expression for this error term is:

$$e_3 = \frac{1}{\pi}\int_{w_{sb1}}^{\pi} \left|N_1(e^{jw})\right|^2 dw + \frac{1}{\pi}\int_{0}^{w_{sb2}} \left|N_2(e^{jw})\right|^2 dw, \tag{3.24}$$

where $N_1(e^{jw})$ and $N_2(e^{jw})$ are the Fourier transforms of $\eta_1(n)$ and $\eta_2(n)$ respectively. The designer can specify the lowpass stopband edge w_{sb1}, and the highpass stopband edge w_{sb2}. Note, if the analysis filters are designed so that they have good frequency band separations, the corresponding synthesis filters will also have good band separation as long as the reconstruction error is small. Hence, this error term is only considered for the analysis filters and not for the synthesis filters.

The last error term, e_4, corresponds to the passband ripple of the synthesis filters. As in any other filter design problem, it is desirable to have filters that have as little passband distortion as possible. For example, one would like the magnitude of the frequency response of $h_1(n)$ to equal one in the frequency range $0 \leq w \leq w_{sb1}$. Therefore, a reasonable expression for e_4 is :

$$e_4 = \frac{1}{\pi}\int_{0}^{w_{sb1}} [|H_1(e^{jw})| - 1]^2 dw + \frac{1}{\pi}\int_{w_{sb2}}^{\pi} [|H_2(e^{jw})| - 1]^2 dw. \tag{3.25}$$

Note, if the above synthesis filters are designed so that they have little passband ripple, the corresponding analysis filters will also have little passband ripple as long as the reconstruction error is small. Hence, this error term is only considered for the synthesis filters and not for the analysis filters.

3.3 Optimum Synthesis Filters

This section is devoted to the design of the optimum synthesis filters for a given pair of analysis filters. The optimum synthesis filters are obtained as the solution of a system of linear equations whose coefficients are determined from the known analysis filters.

The dimension of the optimization problem can be halved due to the symmetry of the filters, since only $\frac{N}{2}$ coefficients are needed to completely specify an Nth-order, symmetric, FIR filter. For convenience, we define $M = \frac{N}{2}$ and the following M-length sequences which are to be determined as to minimize the errors defined in the last section :

$$w_1(n) = h_1(2n) \;\; n = 0, \ldots, M-1 \tag{3.26}$$

$$w_2(n) = h_2(2n) \;\; n = 0, \ldots, M-1 \tag{3.27}$$

$$v_1(n) = h_1(2n+1) \;\; n = 0, \ldots, M-1 \tag{3.28}$$

$$v_2(n) = h_2(2n+1) \;\; n = 0, \ldots, M-1. \tag{3.29}$$

We will also make use of the following M-length sequences which are derived from the known analysis filters and are fixed throughout this section:

$$\beta_1(n) = \eta_1(2n) \;\; n = 0, \ldots, M-1 \tag{3.30}$$

$$\beta_2(n) = \eta_2(2n) \;\; n = 0, \ldots, M-1 \tag{3.31}$$

$$\gamma_1(n) = \eta_1(2n+1) \;\; n = 0, \ldots, M-1 \tag{3.32}$$

$$\gamma_2(n) = \eta_2(2n+1) \;\; n = 0, \ldots, M-1. \tag{3.33}$$

Note that $v_1(n) = w_1(M-1-n)$, $v_2(n) = -w_2(M-1-n)$, $\gamma_1(n) = \beta_1(M-1-n)$, and $\gamma_2(n) = -\beta_2(M-1-n)$. These relations hold because every filter is either symmetric or antisymmetric. Note that each of these $\frac{N}{2}$- length sequences contains all the necessary information to completely determine the N-length filter corresponding to it.

To carry out the optimization over the $\frac{N}{2}$-length vectors $\vec{w}_1$ and $\vec{w}_2$, it is necessary to express the relevant errors in terms of these two vectors. Since these errors are already expressed in terms of $g_1(n)$, $g_1'(n)$, $g_2(n)$ and $g_2'(n)$ we need to rewrite $g_1(n)$, $g_1'(n)$, $g_2(n)$ and $g_2'(n)$ in terms of the these two vectors. It is easier to derive two separate expressions for each of these sequences depending on whether index n is even or odd. The following expressions can be easily derived by simple substitution of (3.26)-(3.33) in (3.12)-(3.17):

$$\begin{gathered} g_1(2n-1) = C_1(n,k) * \vec{w}_1(k), \\ \text{where} \\ C_1(n,k) = \gamma_1(n-1-k)u(n-1-k) + \beta_1(n+k-M)u(k-M+n) \\ n = 1,\ldots,N-1 \text{ and } k = 0,\ldots,M-1. \end{gathered} \tag{3.34}$$

$$\begin{gathered} g_2(2n) = B_1(n,k) * \vec{w}_1(k), \\ \text{where} \\ B_1(n,k) = \gamma_1(n-m+k)u(k-M+n) + \beta(n-k)u(n-k) \\ n = 0,\ldots,N-1 \text{ and } k = 0,\ldots,M-1. \end{gathered} \tag{3.35}$$

$$\begin{gathered} g_1'(2n-1) = D_1(n,k) * \vec{w}_1(k), \\ \text{where} \\ D_1(n,k) = \gamma_1(n-1-k)u(n-1-k) - \beta(n+k-M)u(k-M+n) \\ n = 1,\ldots,N-1 \text{ and } k = 0,\ldots,M-1. \end{gathered} \tag{3.36}$$

$$
\begin{gathered}
g_1'(2n) = E_1(n,k) * \vec{w}_1(k), \\
\text{where,} \\
E_1(n,k) = -\gamma_1(n-M+k)u(k-M+n+1) + \beta(n-k)u(n-k) \\
n = 0,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.
\end{gathered}
\tag{3.37}
$$

$$
\begin{gathered}
g_2(2n-1) = C_2(n,k) * \vec{w}_2(k) \\
\text{where,} \\
C_2(n,k) = \gamma_2(n-k-1)u(n-1-k) + \beta_2(n-M+k)u(k-M+n) \\
n = 1,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.
\end{gathered}
\tag{3.38}
$$

$$
\begin{gathered}
g_2(2n) = B_2(n,k) * \vec{w}_2(k) \\
\text{where,} \\
B_2(n,k) = -\gamma_2(n-M+k)u(k-M+n+1) + \beta_2(n-k)u(n-k) \\
n = 0,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.
\end{gathered}
\tag{3.39}
$$

$$
\begin{gathered}
g_2'(2n-1) = D_2(n,k) * \vec{w}_2(k) \\
\text{where,} \\
D_2(n,k) = \gamma_2(n-k-1)u(k-M+n+1) + \beta_2(n-k)u(n-k) \\
n = 1,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.
\end{gathered}
\tag{3.40}
$$

$$
\begin{gathered}
g_2'(2n) = E_2(n,k) * \vec{w}_2(k) \\
\text{where,} \\
E_2(n,k) = \gamma_2(n-M+k)u(k-M+n+1) + \beta_2(n-k)u(n-k) \\
n = 0,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.
\end{gathered}
\tag{3.41}
$$

Having obtained the above expressions for $g_1(n)$, $g_1'(n)$, $g_2(n)$, and $g_2'(n)$, we can easily express the relevant errors in terms of the unknown vectors $\vec{w}_1$ and $\vec{w}_2$. The three

error terms used in finding the optimum synthesis filters are the ones which correspond to perfect reconstruction, e_1, aliasing, e_2, and passband ripple, e_4.

The first error term, corresponding to the flatness of the frequency response in the absence of aliasing, can be broken into an even and odd part as:

$$e_1 = \sum_{n=1}^{N-1}(g_1(2n-1) - g_2(2n-1) - \delta[2n-1])^2 + \sum_{n=0}^{N-1}(g_1(2n) - g_2(2n) - \delta[2n])^2. \quad (3.42)$$

Letting $\vec{\delta}_o(n) = \delta[2n-1] \;\; n = 1, \ldots, N-1$ and $\vec{\delta}_e(n) = \delta(2n) \;\; n = 0, \ldots, N-1$, we can rewrite (3.42) in matrix format as :

$$e_1 = (C_1\vec{w}_1 - C_2\vec{w}_2 - \vec{\delta}_o)^T(C_1\vec{w}_1 - C_2\vec{w}_2 - \vec{\delta}_o) + (B_1\vec{w}_1 - B_2\vec{w}_2 - \vec{\delta}_e)^T(B_1\vec{w}_1 - B_2\vec{w}_2 - \vec{\delta}_e), \quad (3.43)$$

where matrices C_1, C_2, B_1, and B_2 are known and defined in (3.34), (3.38), (3.36), and (3.39) respectively.

The second error term, corresponding to aliasing, can be similarly expressed in matrix format as :

$$e_2 = (D_1\vec{w}_1 - D_2\vec{w}_2)^T(D_1\vec{w}_1 - D_2\vec{w}_2) + (E_1\vec{w}_1 - E_2\vec{w}_2)^T(E_1\vec{w}_1 - E_2\vec{w}_2), \quad (3.44)$$

where D_1, D_2, E_1, and E_2 are defined in (3.36), (3.40), (3.37), and (3.41) respectively.

The last error term to be considered here is e_4, corresponding to the passband ripple of the synthesis filters. With a bit of algebra, we can write this error term in matrix format as

$$e_4 = 2\vec{w}_1^T G_1 \vec{w}_1 + 2\vec{w}_1^T \vec{q}_1 + 2\vec{w}_2 G_2 \vec{w}_2 + 2\vec{w}_2^T \vec{q}_2, \quad (3.45)$$

where the i, k entries of the $M \times M$ matrices G_1 and G_2 are :

$$g_1(i,k) = \frac{1}{\pi}\left\{ \begin{array}{ll} w_{sb1} & i = k \\ -\frac{\sin(i-k)w_{sb1}}{(i-k)} & i \neq k \end{array} \right\} + \frac{1}{\pi}\left\{ \begin{array}{ll} w_{sb1} & i+k = N-1 \\ -\frac{\sin(i+k-N+1)w_{sb1}}{(i+k-N+1)} & i+k \neq N-1 \end{array} \right\} \quad (3.46)$$

$$g_2(i,k) = \frac{1}{\pi}\left\{ \begin{array}{ll} \pi - w_{sb2} & i = k \\ -\frac{\sin(i-k)w_{sb2}}{(i-k)} & i \neq k \end{array} \right\} + \frac{1}{\pi}\left\{ \begin{array}{ll} w_{sb2} - \pi & i+k = N-1 \\ -\frac{\sin(i+k-N+1)w_{sb2}}{(i+k-N+1)} & i+k \neq N-1 \end{array} \right\}. \tag{3.47}$$

The elements of the $M \times 1$ vectors $\vec{q}_1$ and $\vec{q}_2$ are

$$\vec{q}_1(i) = -\frac{2}{\pi}\frac{\sin(i - \frac{N-1}{2})w_{sb1}}{(i - \frac{N-1}{2})} \qquad \vec{q}_2(i) = \frac{2}{\pi}\frac{\cos(\frac{N-1}{2} - i)w_{sb2}}{(\frac{N-1}{2} - i)}. \tag{3.48}$$

In short, the problem of finding the optimum pair of synthesis filters for a given pair of analysis filters reduces to finding $\vec{w}_1$ and $\vec{w}_2$ which minimize

$$e = e_1 + e_2 + e_4 \tag{3.49}$$

$$\begin{aligned} = {} & (C_1\vec{w}_1 - C_2\vec{w}_2 - \vec{\delta}_o)^T(C_1\vec{w}_1 - C_2\vec{w}_2 - \vec{\delta}_o) + (B_1\vec{w}_1 - B_2\vec{w}_2 - \vec{\delta}_e)^T(B_1\vec{w}_1 - B_2\vec{w}_2 - \vec{\delta}_e) + \\ & (D_1\vec{w}_1 - D_2\vec{w}_2)^T(D_1\vec{w}_1 - D_2\vec{w}_2) + (E_1\vec{w}_1 - E_2\vec{w}_2)^T(E_1\vec{w}_1 - E_2\vec{w}_2) + \\ & \vec{w}_2^T G_2\vec{w}_2 + \vec{w}_2^T\vec{q}_2 + \vec{w}_1^T G_1\vec{w}_1 + \vec{w}^T\vec{q}_1, \end{aligned} \tag{3.50}$$

subject to the constraint that

$$(\vec{w}_1^T\vec{\beta}_1 + \vec{w}_2^T\vec{\beta}_2) = \frac{1}{2}. \tag{3.51}$$

By introducing a Lagrange multiplier, λ, we can easily transform this constrained optimization problem into an equivalent unconstrained optimization. More specifically, the problem can be restated as finding $\vec{w}_1$, $\vec{w}_2$, and λ which minimize

$$\hat{e} = e + \lambda[\vec{w}_1^T\vec{\beta}_1 + \vec{w}_2^T\vec{\beta}_2 - \frac{1}{2}], \tag{3.52}$$

where e is defined in (3.49). Due to the quadratic nature of each term involved in $\hat{e}$, the resulting optimization problem is a least-square one. With the mild assumption that each matrix involved has linearly independent columns, the existence of a unique solution which minimizes $\hat{e}$ is guaranteed [12].

The necessary conditions for $\vec{w}_1$, $\vec{w}_2$, and λ to minimize $\hat{e}$ are: $\frac{\partial \hat{e}}{\partial \vec{w}_1} = \vec{0}$, $\frac{\partial \hat{e}}{\partial \vec{w}_2} = \vec{0}$, and $\frac{\partial \hat{e}}{\partial \lambda} = 0$. These conditions lead respectively to the following set of simultaneous linear equations:

$$\begin{aligned}
\vec{0} &= \lambda\vec{\beta}_1 + (\underbrace{B_1^T B_1 + C_1^T C_1 + D_1^T D_1 + E_1^T E_1 + G_1}_{A})\vec{w}_1 - \\
&\quad (\underbrace{B_1^T B_2 + C_1^T C_2 + D_1^T D_2 + E_1^T E_2}_{B})\vec{w}_2 - (\underbrace{C_1^T\vec{\delta}_e + B_1^T\vec{\delta}_o - \vec{q}_1}_{\vec{c}}). \qquad (3.53)\\
\vec{0} &= \lambda\vec{\beta}_2 - (\underbrace{B_2^T B_1 + C_2^T C_1 + D_2^T D_1 + E_2^T E_1}_{D})\vec{w}_1 + \\
&\quad (\underbrace{B_2^T B_2 + C_2^T C_2 + D_2^T D_2 + E_2^T E_2 + G_2}_{E})\vec{w}_2 + (\underbrace{C_2^T\vec{\delta}_e + B_2^T\vec{\delta}_o + \vec{q}_2}_{\vec{f}}). \qquad (3.54)\\
\frac{1}{2} &= \vec{w}_1^T\vec{\beta}_1 + \vec{w}_2^T\vec{\beta}_2. \qquad (3.55)
\end{aligned}$$

With a bit of algebra, it can be easily shown that the solutions to the above set of equations are

$$\begin{aligned}
\lambda &= \frac{\frac{1}{2} - (\vec{\beta}_1^T\vec{k} + \vec{\beta}_2^T\vec{m})}{\vec{\beta}_1^T\vec{l} + \vec{\beta}_2^T\vec{n}} \qquad (3.56)\\
\vec{w}_1 &= \vec{k} + \lambda\vec{l} \qquad (3.57)\\
\vec{w}_2 &= \vec{m} + \lambda\vec{n}, \qquad (3.58)\\
& \qquad (3.59)
\end{aligned}$$

where

$$\begin{aligned}
\vec{k} &= (A - BE^{-1}D)^{-1}(BE^{-1}\vec{f} - \vec{c}) \qquad (3.60)\\
\vec{l} &= (A - BE^{-1}D)^{-1}(BE^{-1}\vec{\beta}_2 - \vec{\beta}_1) \qquad (3.61)\\
\vec{m} &= (B - AD^{-1}E)^{-1}(AD^{-1}\vec{f} - \vec{c}) \qquad (3.62)\\
\vec{n} &= (B - AD^{-1}D)^{-1}(AD^{-1}\vec{\beta}_2 - \vec{\beta}_1). \qquad (3.63)
\end{aligned}$$

Note that (3.57)-(3.58) determine $\vec{w}_1$ and $\vec{w}_2$ which in turn completely specify the optimum synthesis filters, $h_1(n)$ and $h_2(n)$, according to the following relations:

$$h_1(n) = \begin{cases} \vec{w}_1(\frac{n}{2}) & \text{if } n \text{ is even} \\ \vec{w}_1(M - \frac{n+1}{2}) & \text{if } n \text{ is odd} \end{cases} \tag{3.64}$$

$$h_2(n) = \begin{cases} \vec{w}_2(\frac{n}{2}) & \text{if } n \text{ is even} \\ -\vec{w}_2(M - \frac{n+1}{2}) & \text{if } n \text{ is odd.} \end{cases} \tag{3.65}$$

This concludes the design of optimum synthesis filters for a given pair of analysis filters. The important observation from this section is that it is possible to obtain optimal synthesis filters for a given pair of analysis filters as the solution to a system of linear equations whose coefficients can be easily determined from the given analysis filters.

3.4 Optimum Analysis Filters

In this section we assume that the synthesis filters are fixed. Similar to the last section, the optimum analysis filters are obtained as the solution to a system of linear equations whose coefficients are determined from the known synthesis filters. The only conceptual difference between this section and the previous one is that the objective function to be minimized here has an additional term that corresponds to the frequency band separation of the analysis filters.

For convenience of notation and reduction of the optimization order, we define $M = \frac{N}{2}$ and the following M-length sequences which are to be determined as to minimize the error measure:

$$w_1(n) = \eta_1(2n) \quad n = 0, \ldots, M-1 \tag{3.66}$$

$$w_2(n) = \eta_2(2n) \quad n = 0, \ldots, M-1 \tag{3.67}$$

$$v_1(n) = \eta_1(2n+1) \quad n = 0, \ldots, M-1 \tag{3.68}$$

$$v_2(n) = \eta_2(2n+1) \quad n = 0, \ldots, M-1. \tag{3.69}$$

We will also make use of the following M-length sequences which are derived from the known synthesis filters and are fixed throughout this section:

$$\beta_1(n) = h_1(2n) \;\; n = 0, \ldots, M-1 \tag{3.70}$$

$$\beta_2(n) = h_2(2n) \;\; n = 0, \ldots, M-1 \tag{3.71}$$

$$\gamma_1(n) = h_1(2n+1) \;\; n = 0, \ldots, M-1 \tag{3.72}$$

$$\gamma_2(n) = h_2(2n+1) \;\; n = 0, \ldots, M-1. \tag{3.73}$$

Note that $v_1(n) = w_1(M-1-n)$, $v_2(n) = -w_2(M-1-n)$, $\gamma_1(n) = \beta_1(M-1-n)$, and $\gamma_2(n) = -\beta_2(M-1-n)$. These relations hold because every filter is either symmetric or antisymmetric. Each of these $\frac{N}{2}$−length sequences contains all the necessary information to completely determine the N-tap filter corresponding to it.

To carry out the optimization over the $\frac{N}{2}$−length vectors $\vec{w}_1$ and $\vec{w}_2$, we need to express the relevant errors in terms of these two vectors. Since all these errors are already expressed in terms of $g_1(n)$, $g_1'(n)$, $g_2(n)$ and $g_2'(n)$, we need to express $g_1(n)$, $g_1'(n)$, $g_2(n)$ and $g_2'(n)$ in terms of these two vectors. It is easier to derive two separate expressions for each of these sequences depending on weather index n is even or odd. By simple substitution of variables, the following expressions can be derived from (3.14)-(3.17):

$$\begin{gathered} g_1(2n-1) = C_1(n,k) * \vec{w}_1(k), \\ \text{where} \\ C_1(n,k) = \gamma_1(n-1-k)u(n-1-k) + \beta_1(n+k-M)u(k-M+n) \\ n = 1, \ldots, N-1 \text{ and } k = 0, \ldots, M-1. \end{gathered} \tag{3.74}$$

$$\begin{gathered} g_2(2n) = B_1(n,k) * \vec{w}_1(k), \\ \text{where} \\ B_1(n,k) = \gamma_1(n-m+k)u(k-M+n) + \beta(n-k)u(n-k) \\ n = 0, \ldots, N-1 \text{ and } k = 0, \ldots, M-1. \end{gathered} \tag{3.75}$$

$$g_1'(2n-1) = D_1(n,k) * \vec{w}_1(k),$$
where
$$D_1(n,k) = -\gamma_1(n-1-k)u(n-1-k) + \beta(n+k-M)u(k-M+n) \tag{3.76}$$
$$n = 1,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.$$

$$g_1'(2n) = E_1(n,k) * \vec{w}_1(k),$$
where,
$$E_1(n,k) = -\gamma_1(n-M+k)u(k-M+n+1) + \beta(n-k)u(n-k) \tag{3.77}$$
$$n = 0,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.$$

$$g_2(2n-1) = C_2(n,k) * \vec{w}_2(k)$$
where,
$$C_2(n,k) = \gamma_2(n-k-1)u(n-1-k) - \beta_2(n-M+k)u(k-M+n) \tag{3.78}$$
$$n = 1,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.$$

$$g_2(2n) = B_2(n,k) * \vec{w}_2(k)$$
where,
$$B_2(n,k) = -\gamma_2(n-M+k)u(k-M+n+1) + \beta_2(n-k)u(n-k) \tag{3.79}$$
$$n = 0,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.$$

$$g_2'(2n-1) = D_2(n,k) * \vec{w}_2(k)$$
where,
$$D_2(n,k) = -\gamma_2(n-k-1)u(k-M+n+1) - \beta_2(n-k)u(n-k) \tag{3.80}$$
$$n = 1,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.$$

$$g_2'(2n) = E_2(n,k) * \vec{w}_2(k)$$
where,
$$E_2(n,k) = \gamma_2(n-M+k)u(k-M+n+1) + \beta_2(n-k)u(n-k) \tag{3.81}$$
$$n = 0,\ldots,N-1 \text{ and } k = 0,\ldots,M-1.$$

Now, we can easily express the relevant error terms as a function of the unknown vectors $\vec{w}_1$ and $\vec{w}_2$. The first error term,corresponding to the flatness of the frequency response in the absence of aliasing, can be written in matrix format as:

$$e_1 = (C_1\vec{w}_1 - C_2\vec{w}_2 - \vec{\delta}_o)^T(C_1\vec{w}_1 - C_2\vec{w}_2 - \vec{\delta}_o) + (B_1\vec{w}_1 - B_2\vec{w}_2 - \vec{\delta}_e)^T(B_1\vec{w}_1 - B_2\vec{w}_2 - \vec{\delta}_e), \tag{3.82}$$

where $\vec{\delta}_o(n) = \delta[2n-1] \quad n = 1,\ldots,N-1$, $\vec{\delta}_e(n) = \delta[2n] \quad n = 0,\ldots,N-1$. Furthermore, C_1, C_2, B_1, and B_2 are defined in (3.74), (3.78), (3.75), and (3.79) respectively.

The second error term, corresponding to the amount of aliasing present in the reconstructed output, can be similarly expressed in matrix format as:

$$e_2 = (D_1\vec{w}_1 - D_2\vec{w}_2)^T(D_1\vec{w}_1 - D_2\vec{w}_2) + (E_1\vec{w}_1 - E_2\vec{w}_2)^T(E_1\vec{w}_1 - E_2\vec{w}_2), \tag{3.83}$$

where D_1, D_2, E_1, and E_2 are defined in (3.76), (3.80), (3.77), and (3.81) respectively.

Finally, the third error term, representing the frequency band separation of $\eta_1(n)$ and $\eta_2(n)$, must be expressed in terms of $\vec{w}_1$ and $\vec{w}_2$. Since there was no error term similar to this one in the last section, we shall present in detail how this term is expressed as a function of $\vec{w}_1$ and $\vec{w}_2$. To be concise, we will concentrate on the part of e_3 which corresponds to $\eta_1(n)$. Recall from (3.24) that

$$e_3 = \frac{1}{\pi}\int_{w_{sb1}}^{\pi} \left|N_1(e^{jw})\right|^2 dw + \frac{1}{\pi}\int_0^{w_{sb2}} \left|N_2(e^{jw})\right|^2 dw. \tag{3.84}$$

Substituting the following definition of the Fourier transform for $\eta_1(n)$

$$\begin{aligned} N_1(e^{jw}) &= \sum_{n=0}^{N-1} \eta_1(n)e^{-jwn} \\ &= \sum_{n=0}^{M-1} 2\eta_1(n)e^{-j(\frac{N-1}{2})w}[\cos(n - \frac{N-1}{2})w], \end{aligned} \tag{3.85}$$

into the first part of e_3 and realizing that $|e^{-j(\frac{N-1}{2})w}| = 1$, we obtain the following:

$$\frac{1}{\pi}\int_{w_{sb1}}^{\pi} |N_1(e^{jw})|^2 dw = \frac{1}{\pi}\int_{w_{sb1}}^{\pi} 4(\sum_{n=0}^{M-1} \eta_1(n)[\cos(n - \frac{N-1}{2})w])^2 dw \tag{3.86}$$

$$= \frac{4}{\pi}\sum_{i=0}^{M-1}\sum_{k=0}^{M-1}\eta_1(i)\eta_1(k)\int_{w_{sb1}}^{\pi}\cos(i-\frac{N-1}{2})w\cos(k-\frac{N-1}{2})wdw.$$

By using the closed-form solution to the integral in (3.86), we can show that

$$\frac{1}{\pi}\int_{w_{sb1}}^{\pi}|N_1(e^{jw})|^2dw = 2\sum_{i=0}^{M-1}\sum_{k=0}^{M-1}\eta_1(i)F_1(i,k)\eta_1(k), \tag{3.87}$$

where the i,k entry of the $M \times M$ matrix F_1 is

$$F_1(i,k) = \frac{1}{\pi}\left\{\begin{array}{ll} \pi - w_{sb1} & i=k \\ -\frac{\sin(i-k)w_{sb1}}{(i-k)} & i\neq k \end{array}\right\} + \frac{1}{\pi}\left\{\begin{array}{ll} \pi - w_{sb1} & i+k=N-1 \\ -\frac{\sin(i+k-N+1)w_{sb1}}{(i+k-N+1)} & i+k\neq N-1 \end{array}\right\}. \tag{3.88}$$

Next, it is necessary to express (3.89) in terms of $\vec{w}_1$. Due to the long and unenlightening derivation of this expression, we only give the final results here:

$$\begin{aligned}\frac{1}{\pi}\int_{w_{sb1}}^{\pi}|N_1(e^{jw})|^2dw &= 2\sum_{i=0}^{M-1}\sum_{k=0}^{M-1}\eta_1(i)F_1(i,k)\eta_1(k) \\ &= 2\sum_{i=0}^{M-1}\sum_{k=0}^{M-1}\vec{w}_1(i)G_1(i,k)\vec{w}_1(k) \\ &= 2\vec{w}_1^T G_1\vec{w}_1, \end{aligned}\tag{3.89}$$

where the i,k entry of the $M \times M$ matrix G_1 is derived from the entries of matrix F_1 according to the following relations:

$$G_1(i,k) = \left\{\begin{array}{lll} F_1(2i,2k) & 0\leq k\leq \frac{M}{2}-1 & 0\leq i\leq \frac{M}{2}-1 \\ F_1(N-1-2i,2k) & 0\leq k\leq \frac{M}{2}-1 & \frac{M}{2}\leq i\leq M-1 \\ F_1(2i,N-1-2k) & \frac{M}{2}\leq k\leq M-1 & \frac{M}{2}\leq i\leq \frac{M}{2}-1 \\ F_1(N-1-2i,N-1-2k) & \frac{M}{2}\leq k\leq M-1 & \frac{M}{2}\leq i\leq M-1. \end{array}\right. \tag{3.90}$$

Similarly, the part of e_3 that corresponds to η_2 can be elegantly written in matrix format as:

$$\frac{1}{\pi}\int_{0}^{w_{sb2}}|N_2(e^{jw})|^2dw = 2\sum_{i=0}^{M-1}\sum_{k=0}^{M-1}\eta_2(i)F_2(i,k)\eta_2(k), \tag{3.91}$$

where the i,k entry of $M \times M$ matrix F_2 is

$$F_1(i,k) = \frac{1}{\pi}\left\{ \begin{array}{ll} w_{sb2} & i=k \\ \frac{\sin(i-k)w_{sb2}}{(i-k)} & i \neq k \end{array} \right\} + \frac{1}{\pi}\left\{ \begin{array}{ll} w_{sb2} & i+k=N-1 \\ -\frac{\sin(i+k-N+1)w_{sb2}}{(i+k-N+1)} & i+k \neq N-1 \end{array} \right\} \quad (3.92)$$

Similarly, we need to express the second part of e_3 in terms of $\vec{w}_2$. Due to the long and unenlightening derivation of this expression, we only give the final results here:

$$\begin{aligned} \frac{1}{\pi}\int_0^{w_{sb1}} |N_2(e^{jw})|^2 dw &= 2\sum_{i=0}^{M-1}\sum_{k=0}^{M-1} \eta_2(i)F_2(i,k)\eta_2(k) \\ &= 2\sum_{i=0}^{M-1}\sum_{k=0}^{M-1} \vec{w}_2(i)G_2(i,k)\vec{w}_2(k) \\ &= 2\vec{w}_2^T G_2 \vec{w}_2, \end{aligned} \quad (3.93)$$

where the i,k entry of the $M \times M$ matrix G_2 is derived from the entries of matrix F_2 according to the following relations:

$$G_2(i,k) = \left\{ \begin{array}{lll} F_2(2i,2k) & 0 \leq k \leq \frac{M}{2}-1 & 0 \leq i \leq \frac{M}{2}-1 \\ -F_2(N-1-2i,2k) & 0 \leq k \leq \frac{M}{2}-1 & \frac{M}{2} \leq i \leq M-1 \\ -F_2(2i,N-1-2k) & \frac{M}{2} \leq k \leq M-1 & \frac{M}{2} \leq i \leq \frac{M}{2}-1 \\ F_2(N-1-2i,N-1-2k) & \frac{M}{2} \leq k \leq M-1 & \frac{M}{2} \leq i \leq M-1. \end{array} \right. \quad (3.94)$$

Having obtained expressions for both parts of e_3 in matrix format, we can rewrite e_3 as the sum of these two expressions:

$$e_3 = 2\vec{w}_1^T G_1 \vec{w}_1 + 2\vec{w}_2^T G_2 \vec{w}_2. \quad (3.95)$$

An analogous derivation would easily show that constraint (3.21), expressed in vector format, is

$$(\vec{w}_1^T \vec{\beta}_1 + \vec{w}_2^T \vec{\beta}_2) = \frac{1}{2}. \quad (3.96)$$

In short, the problem of finding the optimum analysis filters for a given pair of synthesis filters reduces to finding $\vec{w}_1$ and $\vec{w}_2$ that minimize

$$\begin{aligned} e &= e_1 + e_2 + e_3 \qquad (3.97) \\ &= (C_1\vec{w}_1 - C_2\vec{w}_2 - \vec{\delta}_o)^T(C_1\vec{w}_1 - C_2\vec{w}_2 - \vec{\delta}_o) + (B_1\vec{w}_1 - B_2\vec{w}_2 - \vec{\delta}_e)^T(B_1\vec{w}_1 - B_2\vec{w}_2 - \vec{\delta}_e) + \\ &\quad (D_1\vec{w}_1 - D_2\vec{w}_2)^T(D_1\vec{w}_1 - D_2\vec{w}_2) + (E_1\vec{w}_1 - E_2\vec{w}_2)^T(E_1\vec{w}_1 - E_2\vec{w}_2) + \\ &\quad \vec{w}_2^T G_2\vec{w}_2 + \vec{w}_1^T G_1\vec{w}_1, \qquad (3.98) \end{aligned}$$

subject to the constraint that

$$(\vec{w}_1^T\vec{\beta}_1 + \vec{w}_2^T\vec{\beta}_2) = \frac{1}{2}. \qquad (3.99)$$

By introducing a Lagrange multiplier, λ, this constrained optimization problem is easily transformed into an equivalent unconstrained one. Specifically, the above problem can be restated as finding $\vec{w}_1$, $\vec{w}_2$, and λ which minimize

$$\hat{e} = e + \lambda(\vec{w}_1^T\vec{\beta}_1 + \vec{w}_2^T\vec{\beta}_2 - \frac{1}{2}). \qquad (3.100)$$

Due to the quadratic nature of each term involved in $\hat{e}$, the final optimization problem is a least-square one. With the mild assumption that each matrix involved has linearly independent columns, the existence of a unique solution which minimizes $\hat{e}$ is guaranteed [12]. Therefore, the necessary conditions for $\vec{w}_1$, $\vec{w}_2$, and λ to minimize $\hat{e}$ are: $\frac{\partial\hat{e}}{\partial\vec{w}_1} = \vec{0}$, $\frac{\partial\hat{e}}{\partial\vec{w}_2} = \vec{0}$, and $\frac{\partial\hat{e}}{\partial\lambda} = 0$. These conditions lead respectively to the following set of simultaneous linear equations:

$$\begin{aligned} \vec{0} &= \lambda\vec{\beta}_1 + (\underbrace{B_1^TB_1 + C_1^TC_1 + D_1^TD_1 + E_1^TE_1 + G_1}_{A})\vec{w}_1 - \\ &\quad (\underbrace{B_1^TB_2 + C_1^TC_2 + D_1^TD_2 + E_1^TE_2}_{B})\vec{w}_2 - (\underbrace{C_1^T\vec{\delta}_e + B_1^T\vec{\delta}_o}_{\vec{c}}). \qquad (3.101) \\ \vec{0} &= \lambda\vec{\beta}_2 - (\underbrace{B_2^TB_1 + C_2^TC_1 + D_2^TD_1 + E_2^TE_1}_{D})\vec{w}_1 + \end{aligned}$$

$$\underbrace{(B_2^T B_2 + C_2^T C_2 + D_2^T D_2 + E_2^T E_2 + G_2)}_{E}\vec{w}_2 + \underbrace{(C_2^T \vec{\delta}_e + B_2^T \vec{\delta}_o)}_{\vec{f}}. \quad (3.102)$$

$$\frac{1}{2} = \vec{w}_1^T \vec{\beta}_1 + \vec{w}_2^T \vec{\beta}_2. \quad (3.103)$$

With a bit of algebra, it can be easily shown that the solutions to the above set of equations are

$$\lambda = \frac{\frac{1}{2} - (\vec{\beta}_1^T \vec{k} + \vec{\beta}_2^T \vec{m})}{\vec{\beta}_1^T \vec{l} + \vec{\beta}_2^T \vec{n}} \quad (3.104)$$

$$\vec{w}_1 = \vec{k} + \lambda \vec{l} \quad (3.105)$$

$$\vec{w}_2 = \vec{m} + \lambda \vec{n}, \quad (3.106)$$

$$(3.107)$$

where

$$\vec{k} = (A - BE^{-1}D)^{-1}(BE^{-1}\vec{f} - \vec{c}) \quad (3.108)$$

$$\vec{l} = (A - BE^{-1}D)^{-1}(BE^{-1}\vec{\beta}_2 - \vec{\beta}_1) \quad (3.109)$$

$$\vec{m} = (B - AD^{-1}E)^{-1}(AD^{-1}\vec{f} - \vec{c}) \quad (3.110)$$

$$\vec{n} = (B - AD^{-1}D)^{-1}(AD^{-1}\vec{\beta}_2 - \vec{\beta}_1) \quad (3.111)$$

Note that, (3.105)-(3.106) determine $\vec{w}_1$ and $\vec{w}_2$ which in turn completely specify the analysis filters, $\eta_1(n)$ and $\eta_2(n)$, according to the following relations:

$$\eta_1(n) = \begin{cases} \vec{w}_1(\frac{n}{2}) & \text{if } n \text{ is even} \\ \vec{w}_1(M - \frac{n+1}{2}) & \text{if } n \text{ is odd} \end{cases} \quad (3.112)$$

$$\eta_2(n) = \begin{cases} \vec{w}_2(\frac{n}{2}) & \text{if } n \text{ is even} \\ -\vec{w}_2(M - \frac{n+1}{2}) & \text{if } n \text{ is odd.} \end{cases} \quad (3.113)$$

This concludes the design of the optimum analysis filters for a given pair of synthesis filters. The important observation from this section is that it is possible to obtain

optimum analysis filters for a given pair of synthesis filters as the solution to a system of linear equations whose coefficients can be easily determined from the given synthesis filters. The only conceptual difference between this section and the previous one is the presence of the frequency band separation error term.

3.5 Jointly-Optimal Solution

In the last two sections, we derived the optimal analysis filters for a given pair of synthesis filters and *vice versa*. The availability of these solutions suggest using them alternatively, until they converge to an optimal set of filters. More formally, the above procedure corresponds to the flow chart of Fig. 3.2.

The above algorithm is in the class of *coordinate descent* algorithms for minimization of functions of several variables [6],[7], since at each step it finds the unique global minimum of the error, with either the analysis or synthesis filters kept fixed. Therefore, the algorithm necessarily converges to a stationary point of the error [6], with a monotonic decrease in the error at each step. Unfortunately, there is no guarantee that the stationary point will be the global minimum; it could be a local minimum or a saddle point.

Our experiments show that convergence to the optimum QMF coefficients is stable, and the accuracy of the final solution is limited only by the accuracy of the matrix inversion routine. With rather arbitrary starting points for the analysis filters, we have always been able to find a solution with very small error. The aliasing error of the resulting QMF banks have always been extermely small, about half the reconstruction error of the alias-free QMF banks obtained by other design techniques.

The iterative algorithm discussed above has a rate of convergence typical of coordinate descent methods; i.e, a weakly linear convergence [7] that is somewhat slower than that of the steepest descent algorithms [6]. Faster convergence, in terms of the number of iterations, could be obtained by using steepest descent or Newton's method.

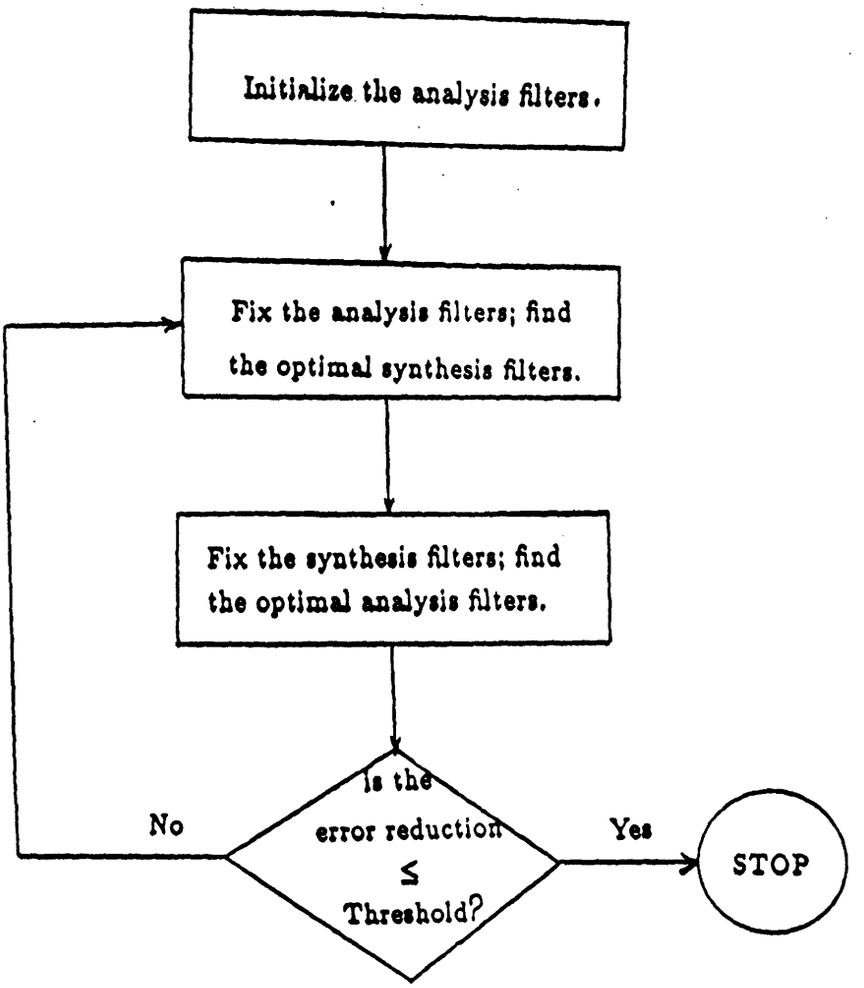

Figure 3.2: Iterative algorithm for jointly-optimal QMF design.

In either of these cases, additional information would have to be computed, namely the gradient of the error for the steepest descent method, and both the gradient and the Hessian for the Newton's method. For example, the Newton's method requires about 1200 operations per iteration with $N = 8$. Typically the coordinate descent algorithm would have converged before a handful of iterations of the Newton's method are completed,since much fewer computations are needed in each iteration of the coordinate descent. Another advantage of the coordinate descent method is that at the end of every iteration we have a partially-optimal solution, in the sense that at least the analysis filters are optimal for the current synthesis filters or *vice versa*.

Chapter 4

Design Examples

Three algorithms for designing quadrature mirror filters were presented in the last two chapters. To compare these algorithms, it is necessary to consider several different issues. Some of these issues are naturally related to the performance of the designed filters, while others are concerned with the efficiency of the design algorithms.

We used these algorithms to design four set of quadrature mirror filters. The number of filter coefficients, the stopband edge frequency, and the passband edge frequency were given to each algorithm and the resulting filters are shown in figures 4.3-4.6. For each set of specifications, the filters designed by our procedure are compared to those designed by either Johnston's or Crochiere's algorithm [9],[8]. For convenience of notation, these filters will be referred to by the following abbreviations:

JON12:	Johnston's 12-tap	stop band edge $=0.7\ \pi$
JON16:	Johnston's 16-tap	stop band edge $=0.7\ \pi$
JON24:	Johnston's 24-tap	stop band edge $=0.625\ \pi$
JAC32:	Jain and Crochiere's 32-tap	stop band edge $=0.6\ \pi$

Corresponding to each of the above QMF banks, there will be a QMF bank designed by our algorithm as well. The QMF banks designed by our algorithm will have the same specifications as the ones given in the above table.

In the last section of this chapter we will present a QMF bank whose lowpass analysis filter is fixed. Since such a QMF bank can not be designed by the other design techniques, no comparison is made in this case.

4.1 Performance Criterion

This section describes the various criterion we will be using to compare the QMF design algorithms. These criterion can be divided into two groups. The first group is related to the performance of the filters obtained by each algorithm. On the other hand, the second group is concerned with the efficiency of the design algorithms.

For a given number of filter coefficients and transition width, different algorithms produce filters with different passband and stopband characteristics. Like any other filter design algorithm, it is desirable to obtain quadrature mirror filters that have unity gain in their passbands. Another consideration is the amount of attenuation at the stopband edge, which is usually the minimum attenuation observed in the stopband. Finally, it is most desirable to have as little reconstruction error as possible. Due to the time-varying nature of the QMF banks, we need to consider the response of each QMF bank to two impulses ($\delta[n]$ and $\delta[n-1]$). We then use an average of these two impulse responses to measure the overall reconstruction error.

We also looked at several issues related to the efficiency of these algorithms. For each set of specifications, we considered the number of iterations each algorithm took to converge to its final result. Furthermore, we examined the complexity of each iteration by calculating the number of multiplications needed to perform each iteration. Finally, we considered the sensitivity of each algorithm to the choice of initial guess.

Example	Stopband Edge	Edge Attenuation	Max Reconstruction Error
JON12	0.7π	-28dB	0.043dB
JON16	$0.7\ \pi$	-35.5dB	0.0174dB
JON24	$0.625\ \pi$	-35.4dB	0.0174dB
JAC32	0.6π	-35dB	0.0174dB

Figure 4.1: Performance of previous design algorithms

Example	Stopband Edge	Edge Attenuation	Max Reconstruction Error
12-tap	0.7π	-26dB	0.05dB
16-tap	$0.7\ \pi$	-42dB	0.0174dB
24-tap	$0.625\ \pi$	-31.4dB	0.026dB
32-tap	0.6π	-37.0dB	0.0174dB

Figure 4.2: Performance of the Jointly-Optimal design algorithm.

4.2 Performance Results

Figures 4.1 and 4.2 summarize the performance of filters obtained by various algorithms. As far as stopband edge attenuation is concerned, our algorithm performs about the same as the previous algorithms. For $N = 12$ and $N = 24$, our filters have slightly less attenuation than the ones given by Johnston. On the other hand, with $N = 16$ and $N = 32$ our filters have slightly more attenuation than the ones given by Johnston and Crochiere.

The overall distortion of our QMF banks is about the same as the ones designed by the other algorithms. The last column in figures 4.1 and 4.2 represents the reconstruction error of the various QMF banks. For $N = 12$ and $N = 24$, our filters produce slightly more distortion Johnston's filters. For $N = 16$ and $N = 32$, our filters produce the same amount of distortion as Johnston's filters or Crochiere's filters [9],[8].

Figures 4.3-4.6 show the performance of various filters in more detail. Consider, for example, figures 4.3(a)-4.3(f). Figures 4.3(a) and 4.3(b) depict respectively the Fourier transform magnitude of the 12-tap analysis and synthesis filters obtained by our design algorithm. Figure 4.3(d) is a plot of the Fourier transform magnitude of the impulse response of our QMF bank. For ease of comparison, Figure 4.3(e) plots the frequency response of our analysis filters and JON12 analysis filters on the same graph. Figure 4.3(f) shows, on the same graph, the overall impulse responses of JON12 QMF bank and our 12-tap QMF bank.

To make a complete comparison, we shall examine the efficiency of each design algorithm based on the number of calculations it performs before converging to a final solution.

First, we compare the number of iterations each algorithm takes to converge to its final solution. This number naturally increases as the number of filter coefficients increases;therefore, we will base our comparisons on the number of iterations required by each algorithm for $N = 32$. Johnston's algorithm requires about 300 iterations;Jain

and Crochiere's algorithm requires about 30 iterations; and our algorithm requires about 40 iterations.

The next issue is the sensitivity of each algorithm to the choice of initial guess. Johnston's method is extremely dependent on this choice. Manual intervention in the form of different starting points and starting step sizes is essential. On the other hand, both Crochiere's method and our method converged to a solution regardless of the initial guess used. Using different starting points effectively increases the number of iterations needed for the algorithm to converge.

The last important issue is the complexity of the calculations required in each iteration. Johnston's technique is the simplest one, because it only needs to evaluate the error function at the neighboring points. Our method requires to invert four $\frac{N}{2} \times \frac{N}{2}$ matrices, with each inversion requiring about $(\frac{N}{2})^3$ multiplications. On the other hand, Jain and Crochiere's method requires finding all the eigenvalues of a $\frac{N}{2} \times \frac{N}{2}$ matrix. This requires about 30 $\times (\frac{N}{2})^3$ multiplications. Therefore, Jain and Crochiere's technique requires about eight times more computations per iteration than our technique.

4.3 QMF Bank with a Guassian Analysis Filter

Using the jointly-optimal algorithm, it is possible to design QMF banks in which the lowpass analysis filter is pre-described. A sharpened Guassian filter was used as the lowpass analysis filter in the QMF bank of figure 4.7(a). Figure 4.7(a) and 4.7(b) depict respectively the frequency responses of the analysis and the synthesis filters. Figure 4.7(c) is a plot of the Fourier transform magnitude of the impulse response of this QMF bank.

The frequency response of the lowpass synthesis filter has a large gain at about $w = \frac{\pi}{2}$, because the corresponding analysis filter has a large attenuation at the same frequency. At those frequencies which the highpass analysis filter has an attenuation in its passband, the corresponding synthesis filter has a gain and *vice versa*. Although

the individual filters do not have a flat frequency response in their passbands, each cascade of the analysis and synthesis filters has a flat response in its corresponding passband. Therefore, the non-flat frequency response of the filters in their passbands is not a problem.

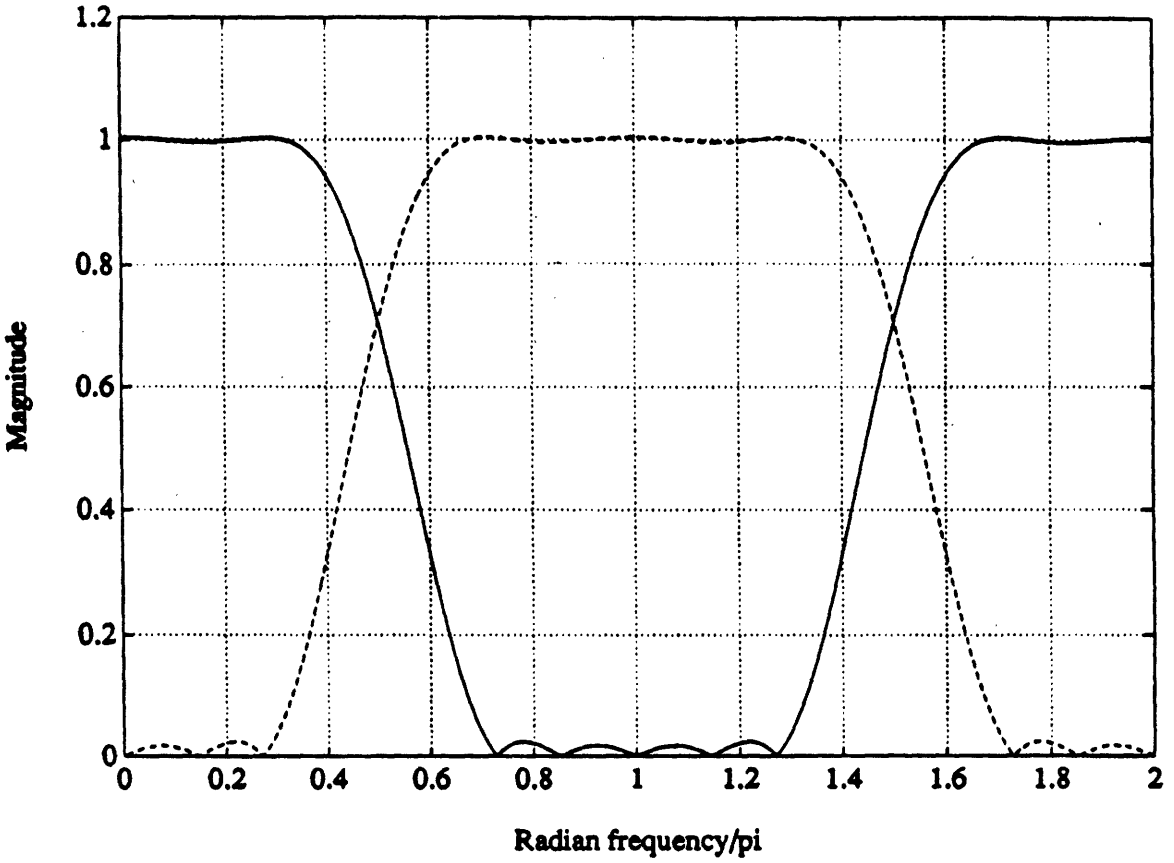

Figure 4.3(a): Our 12-tap analysis filters.

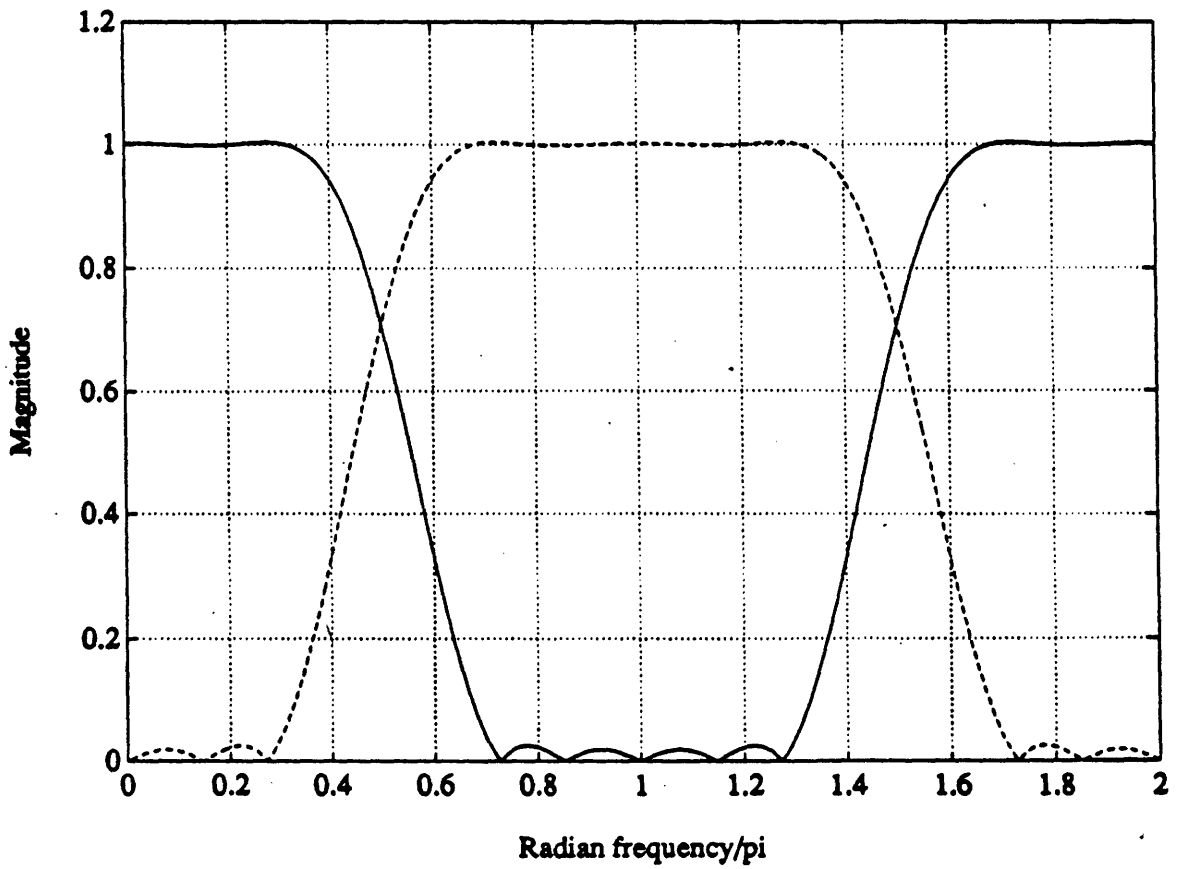

Figure 4.3(b): Our 12-tap synthesis filters.

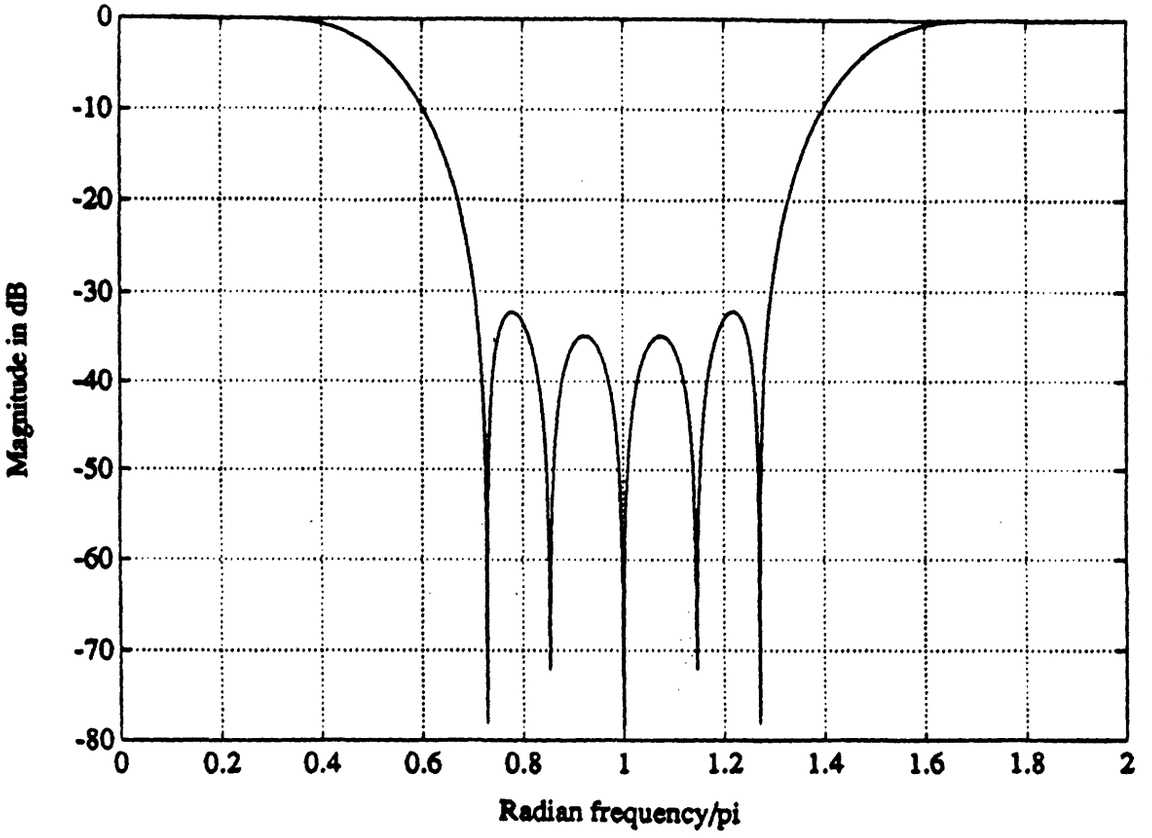

Figure 4.3(c): Our 12-tap, lowpass analysis filter.

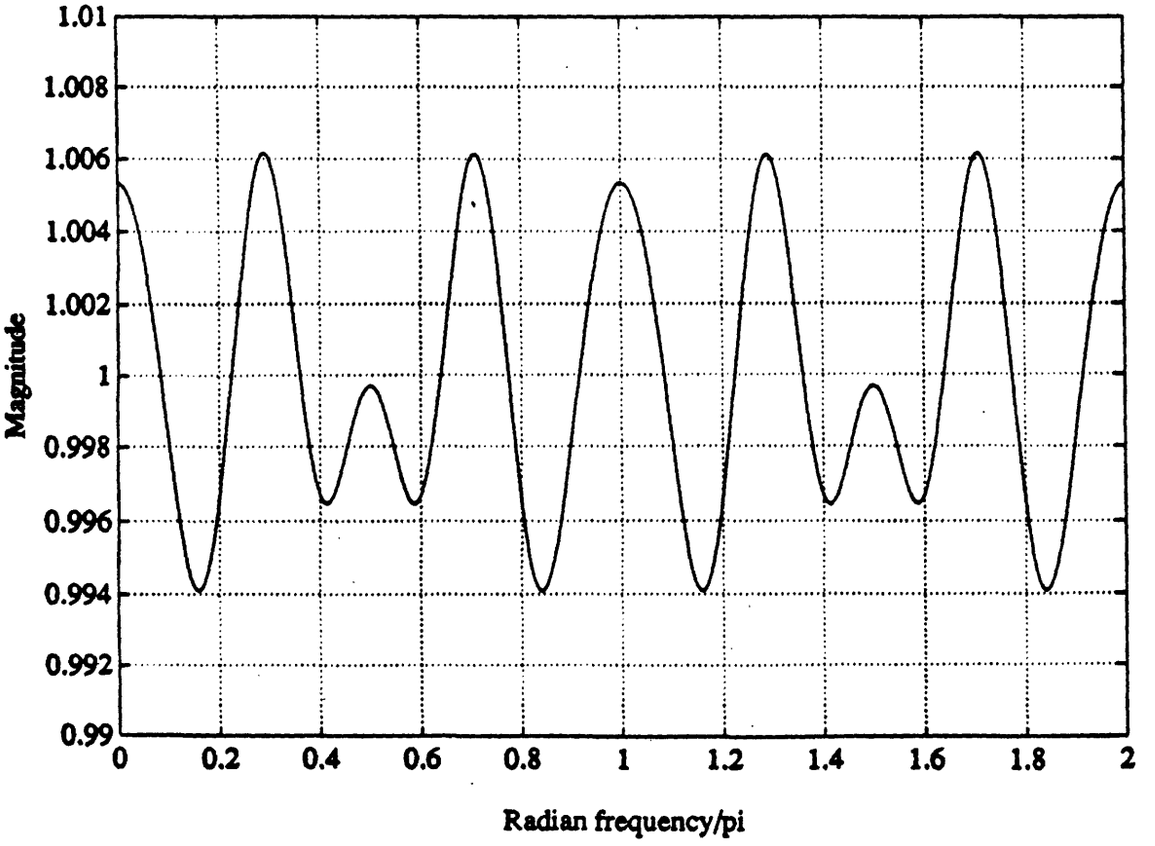

Figure 4.3(d): Overall response of our 12-tap QMF bank.

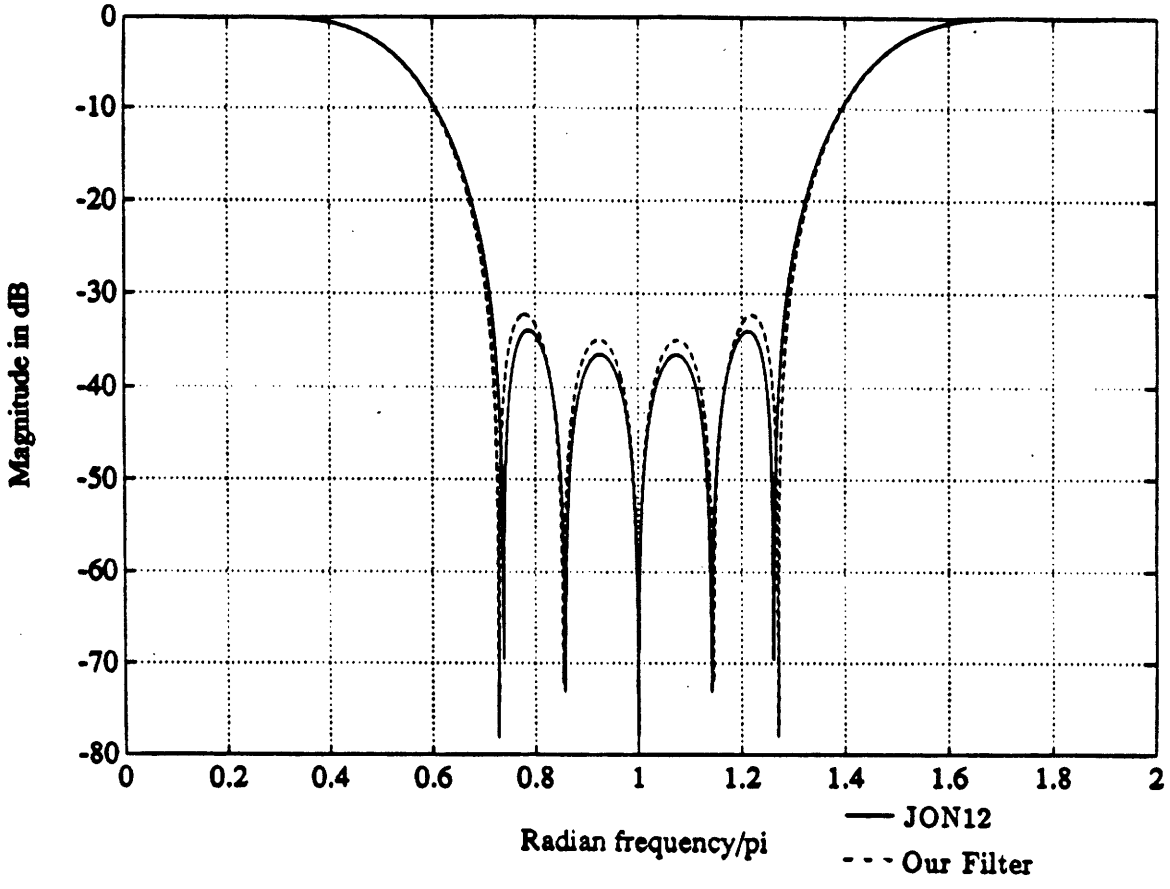

Figure 4.3(e): JON12 *vs* our 12-tap, lowpass analysis filter.

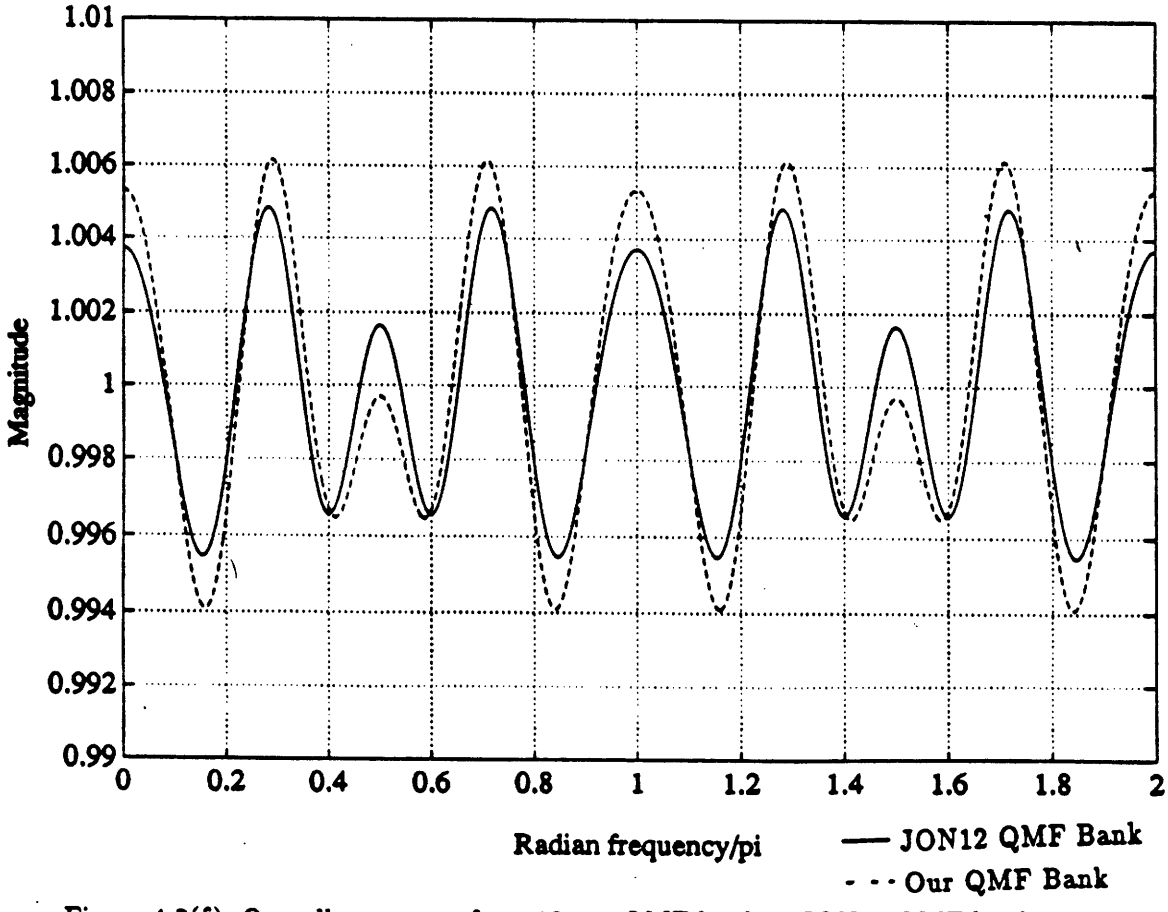

Figure 4.3(f): Overall response of our 12-tap QMF bank *vs* JON12 QMF bank.

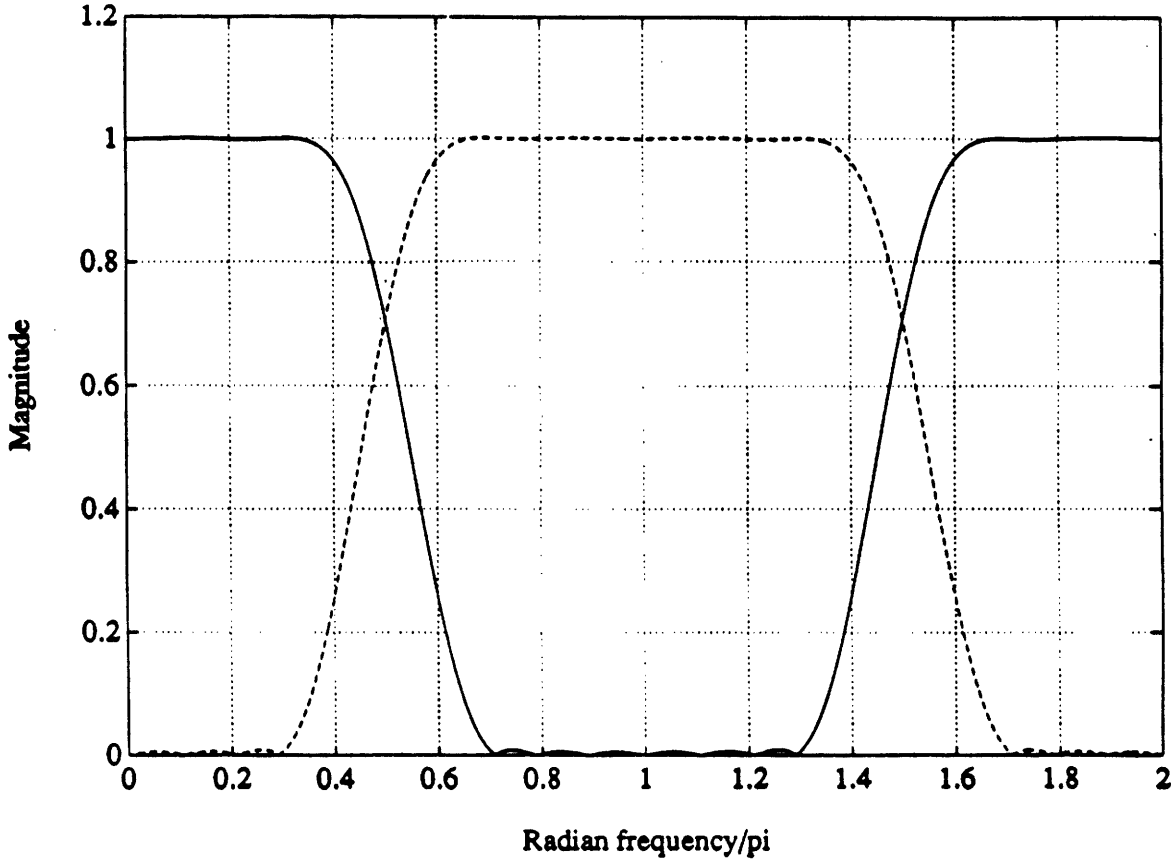

Figure 4.4(a): Our 16-tap analysis filters.

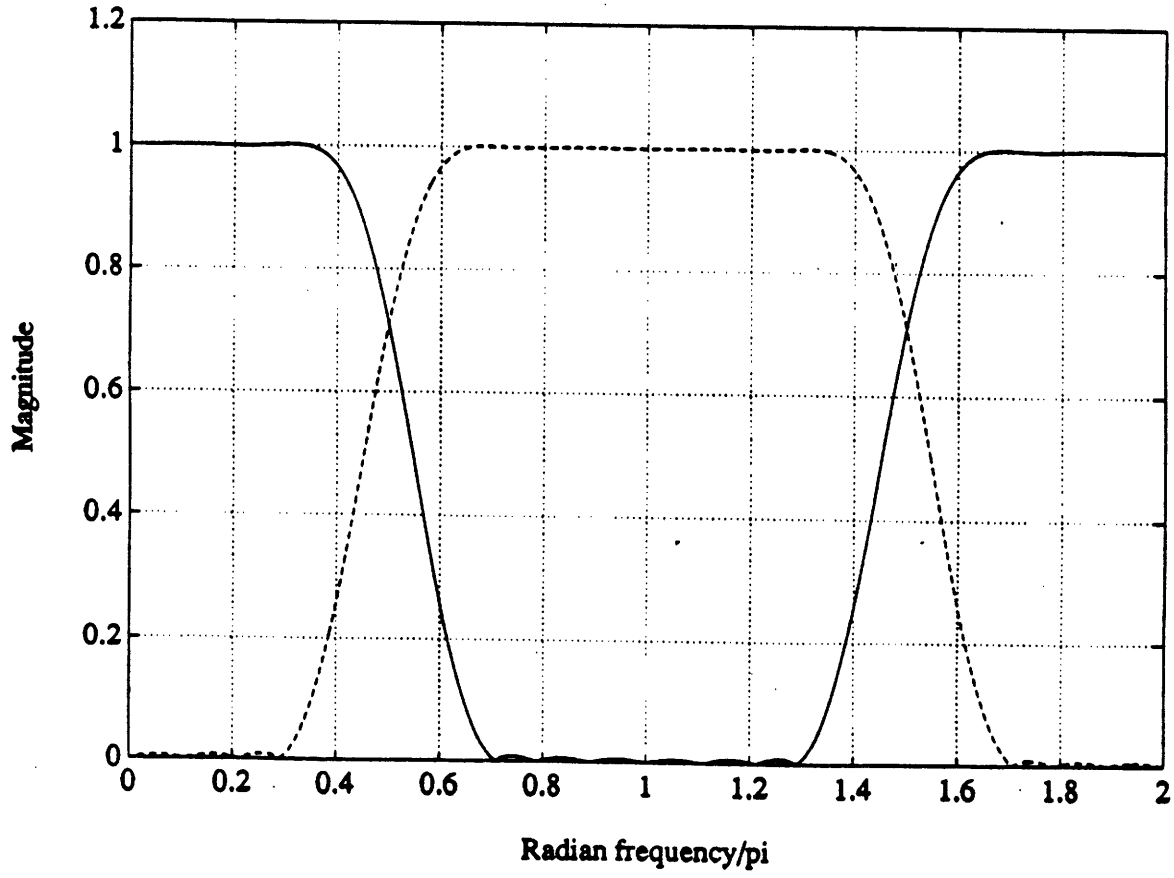

Figure 4.4(b): Our 16-tap synthesis filters.

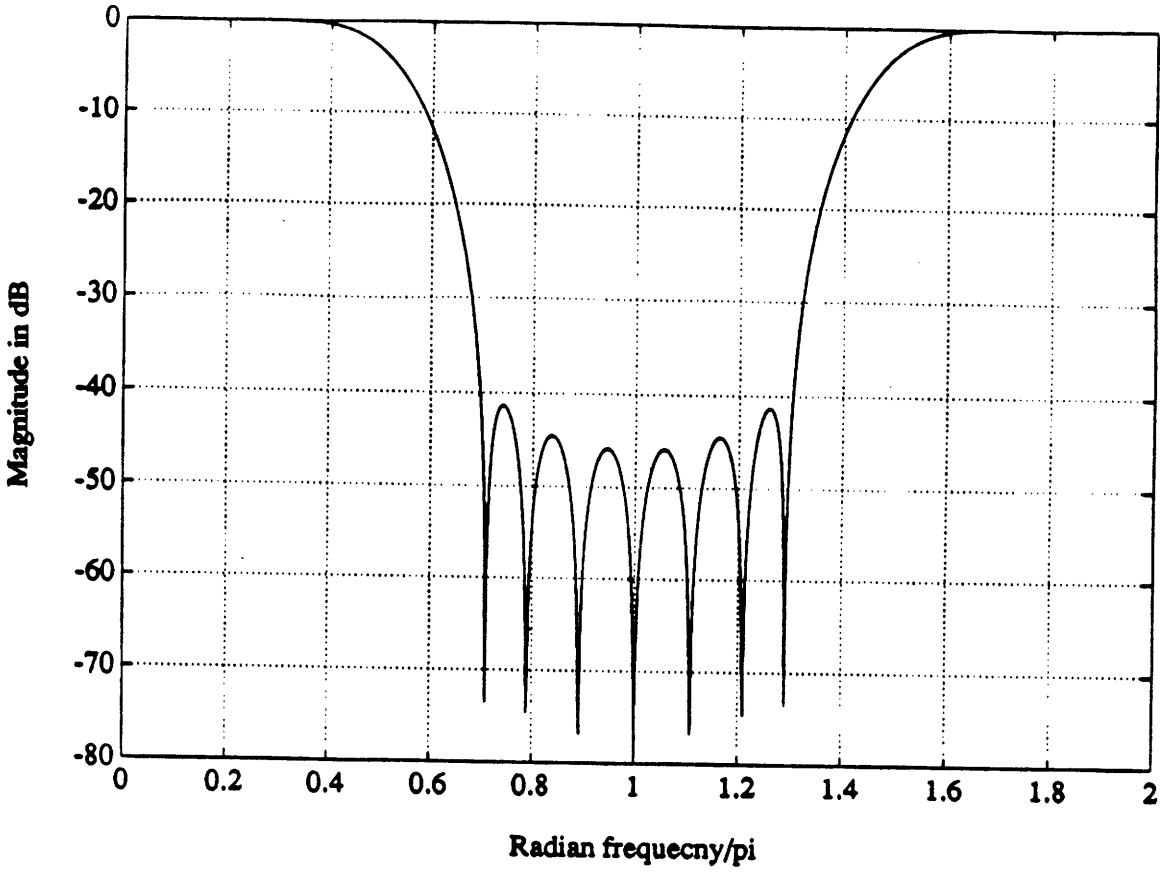

Figure 4.4(c): Our 16-tap, lowpass analysis filter.

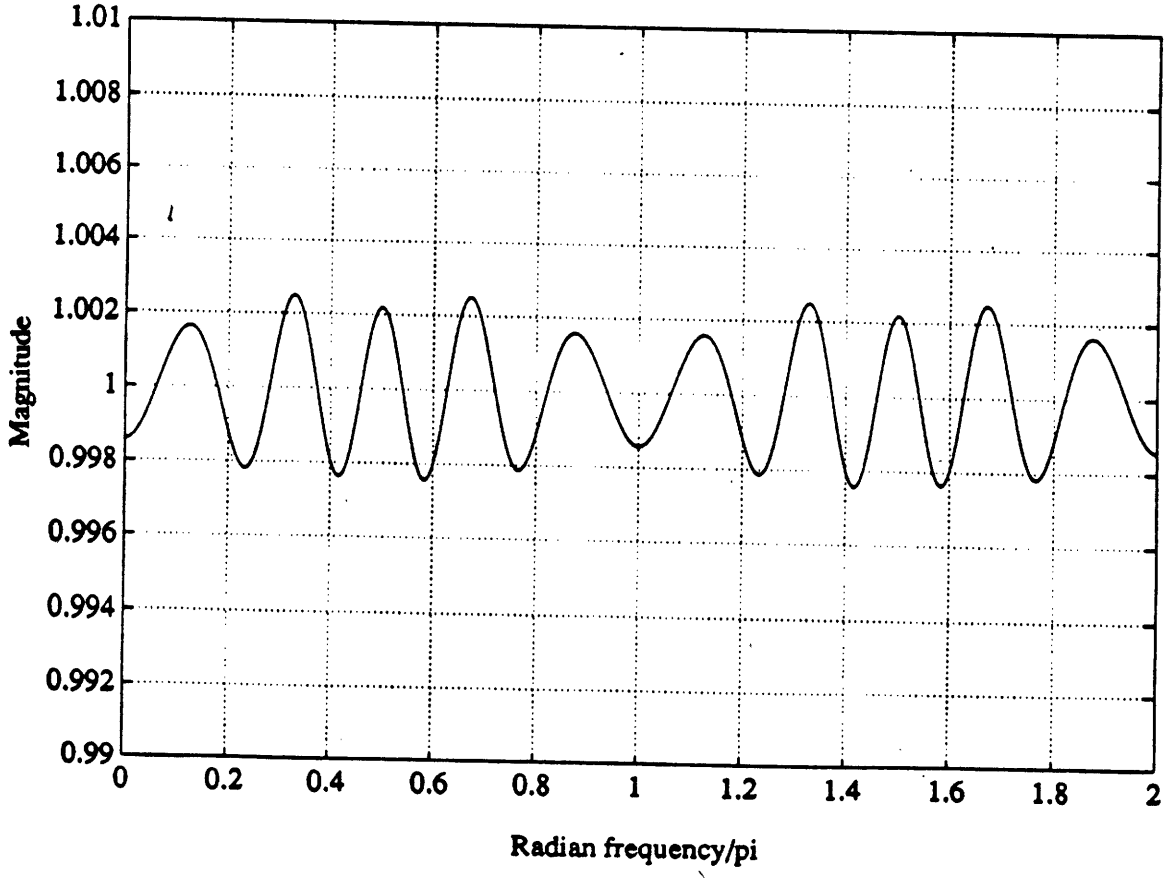

Figure 4.4(d): Overall response of our 16-tap QMF bank.

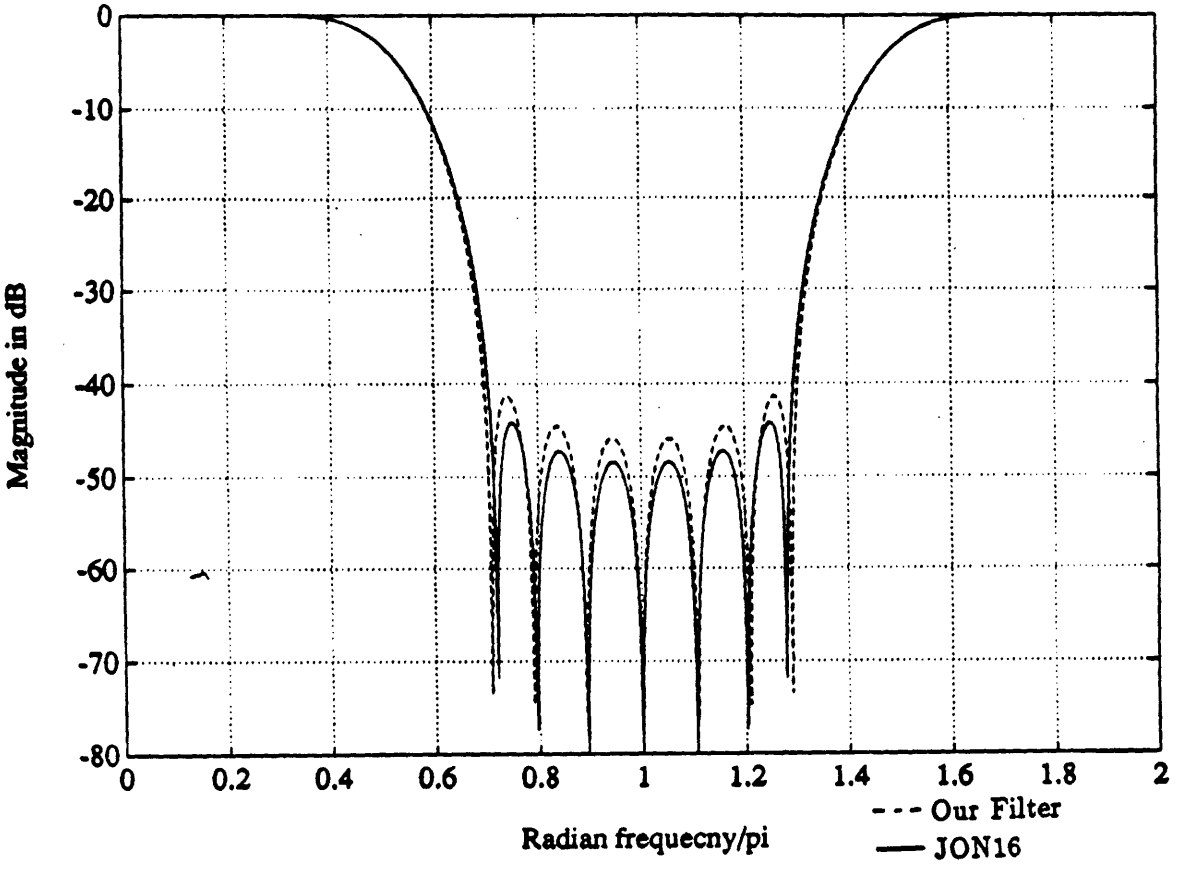

Figure 4.4(e): JON16 *vs* our 16-tap, lowpass analysis filter.

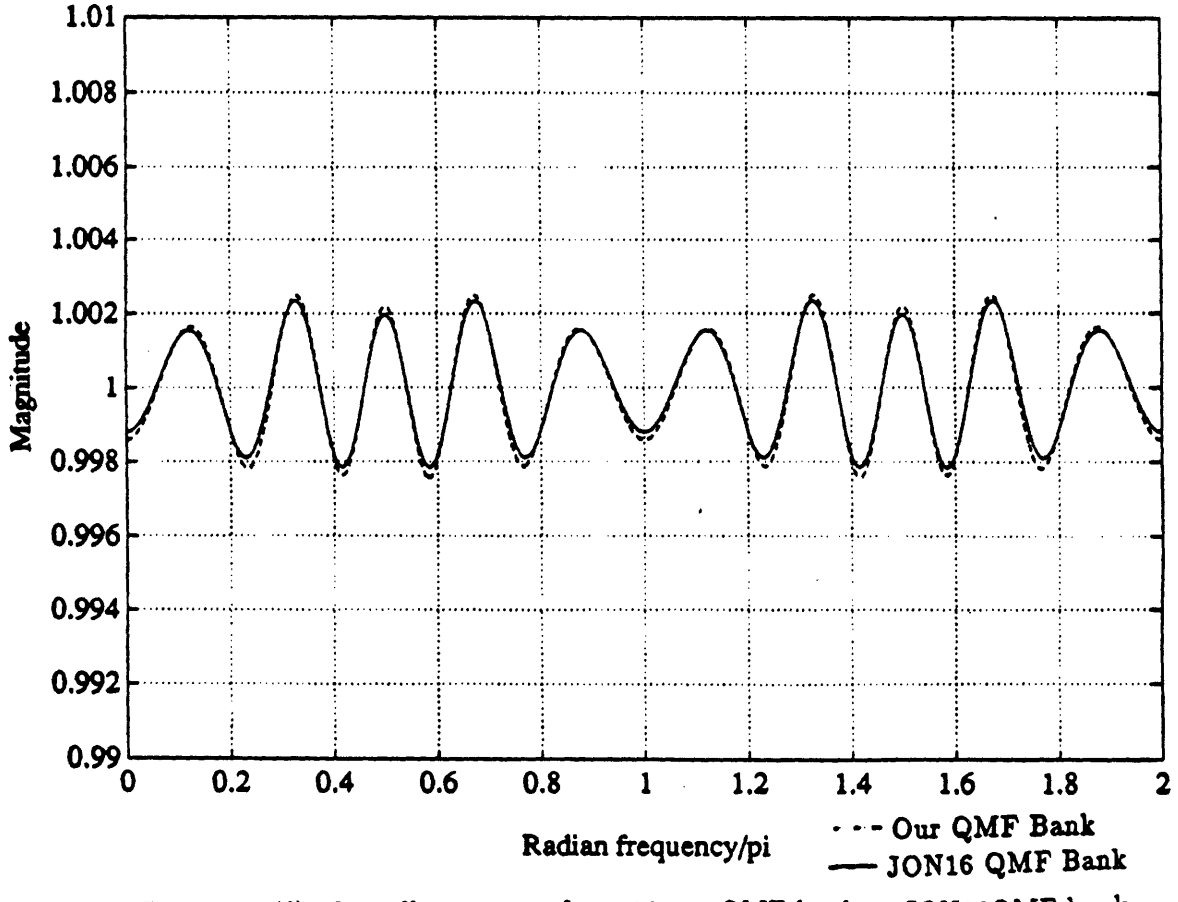

Figure 4.4(f): Overall response of our 16-tap QMF bank *vs* JON16QMF bank.

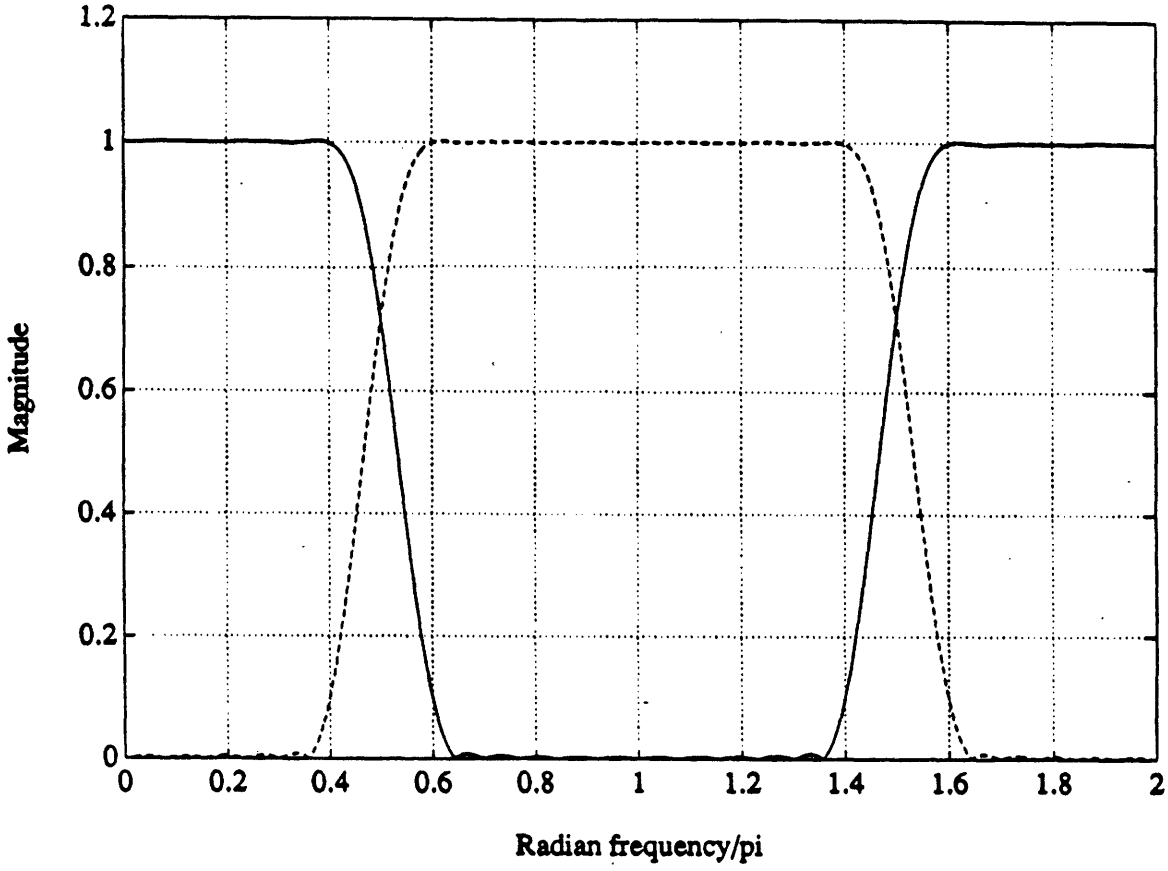

Figure 4.5(a): Our 24-tap analysis filters.

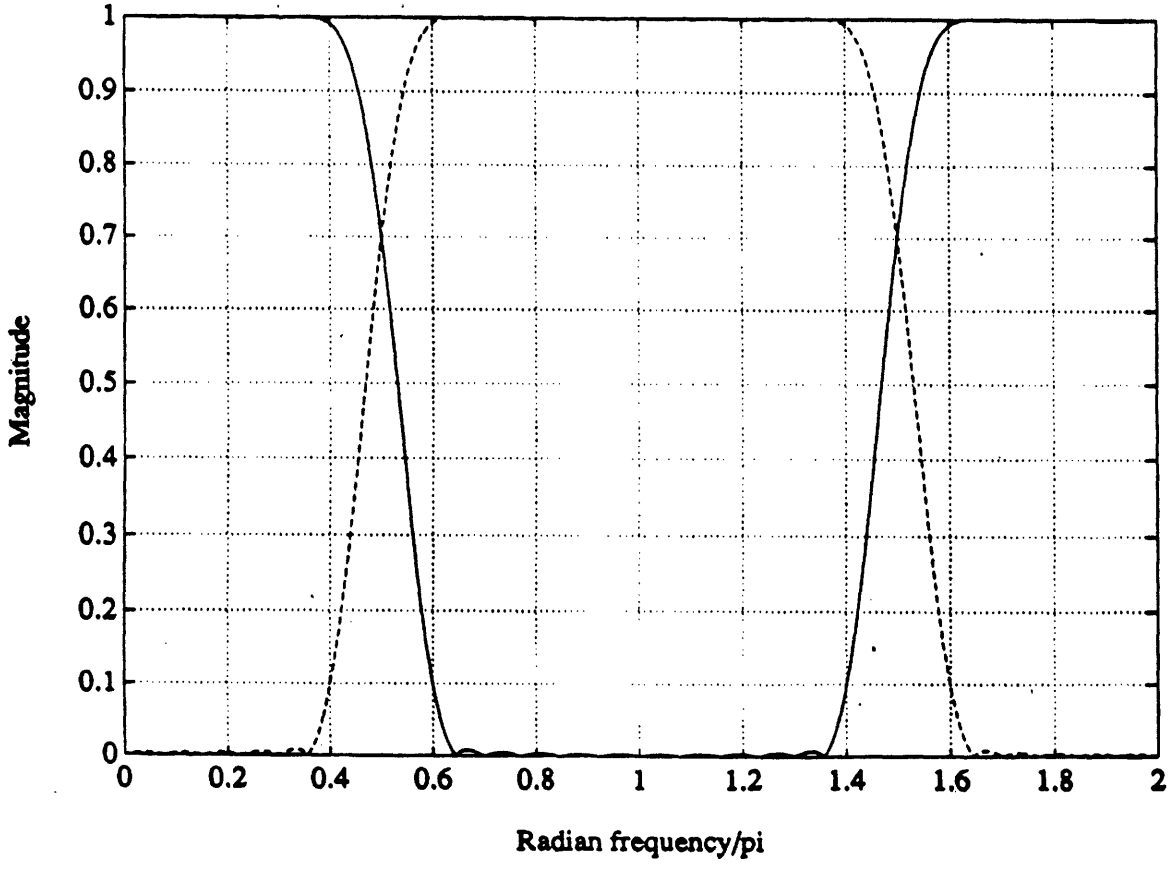

Figure 4.5(b): Our 24-tap synthesis filters.

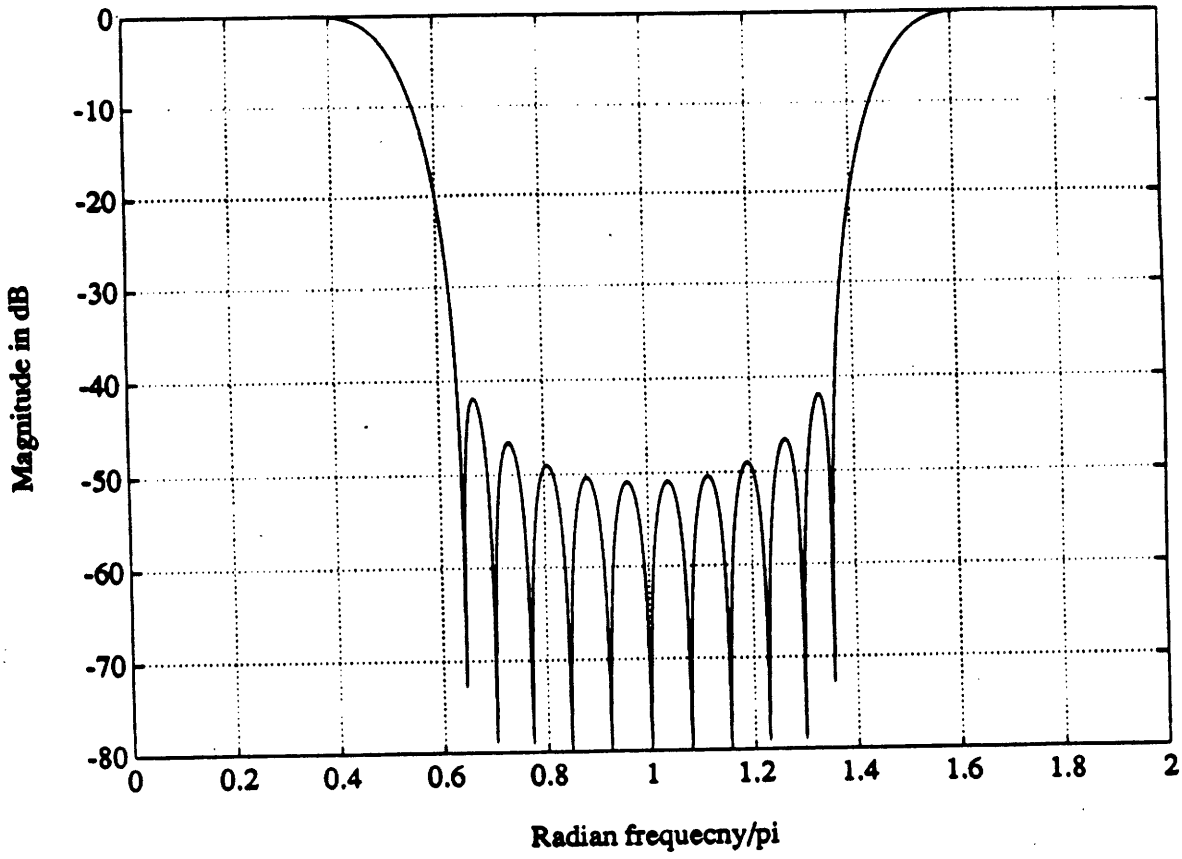

Figure 4.5(c): Our 24-tap, lowpass analysis filter.

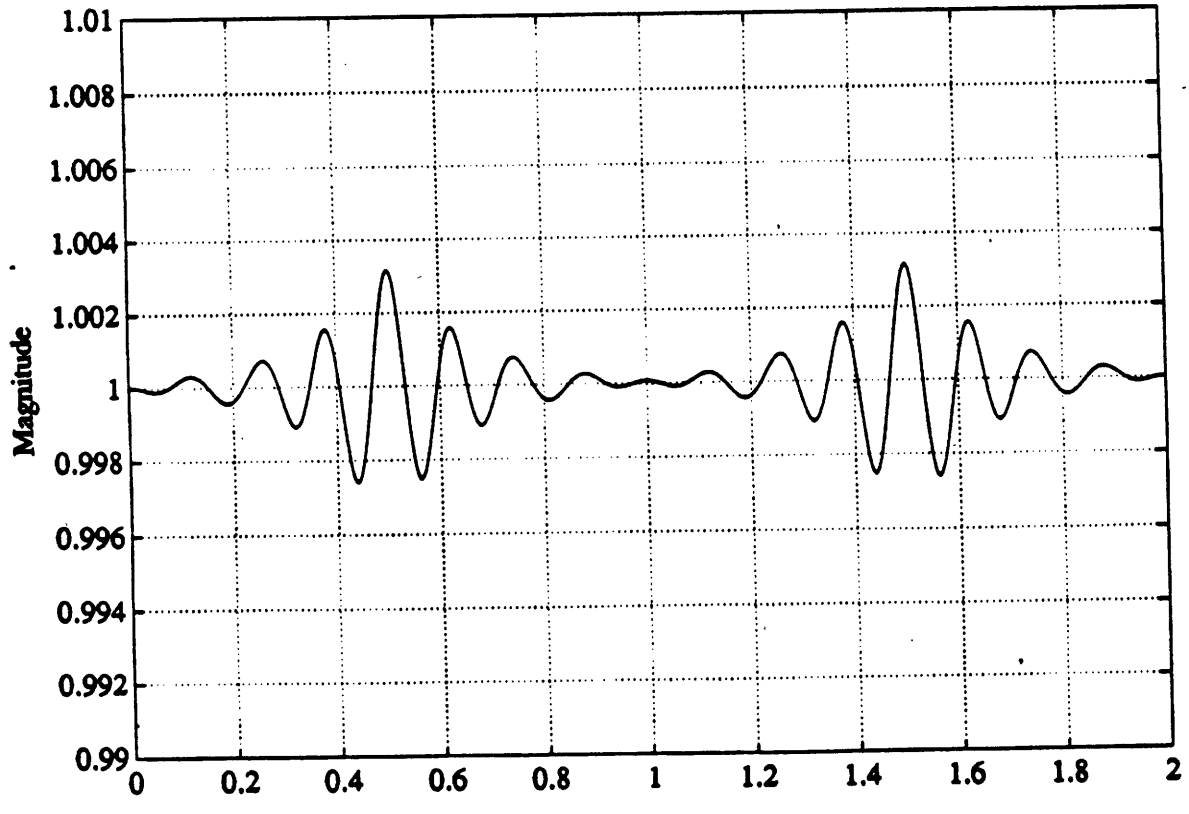

Figure 4.5(d): Overall response of our 24-tap QMF bank.

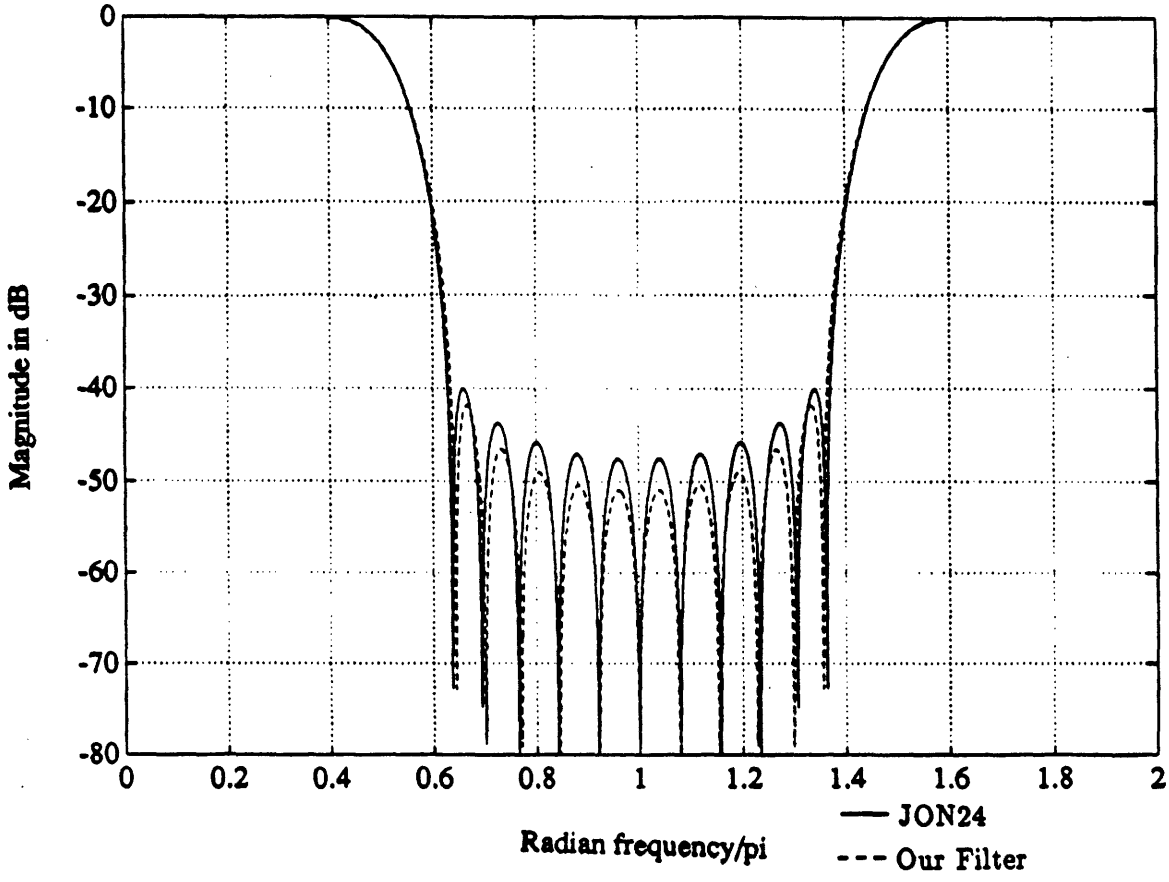

Figure 4.5(e): JON24 *vs* our 24-tap, lowpass analysis filter.

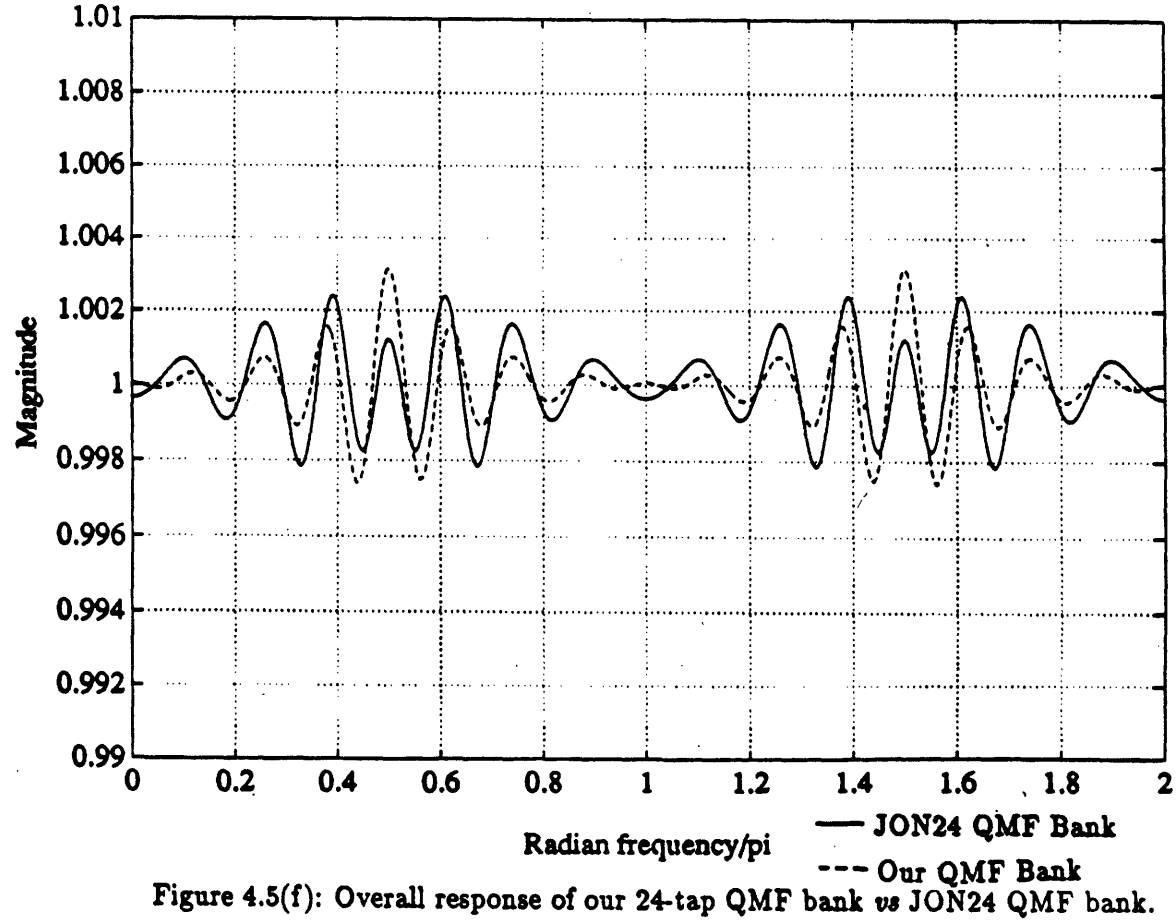

Figure 4.5(f): Overall response of our 24-tap QMF bank *vs* JON24 QMF bank.

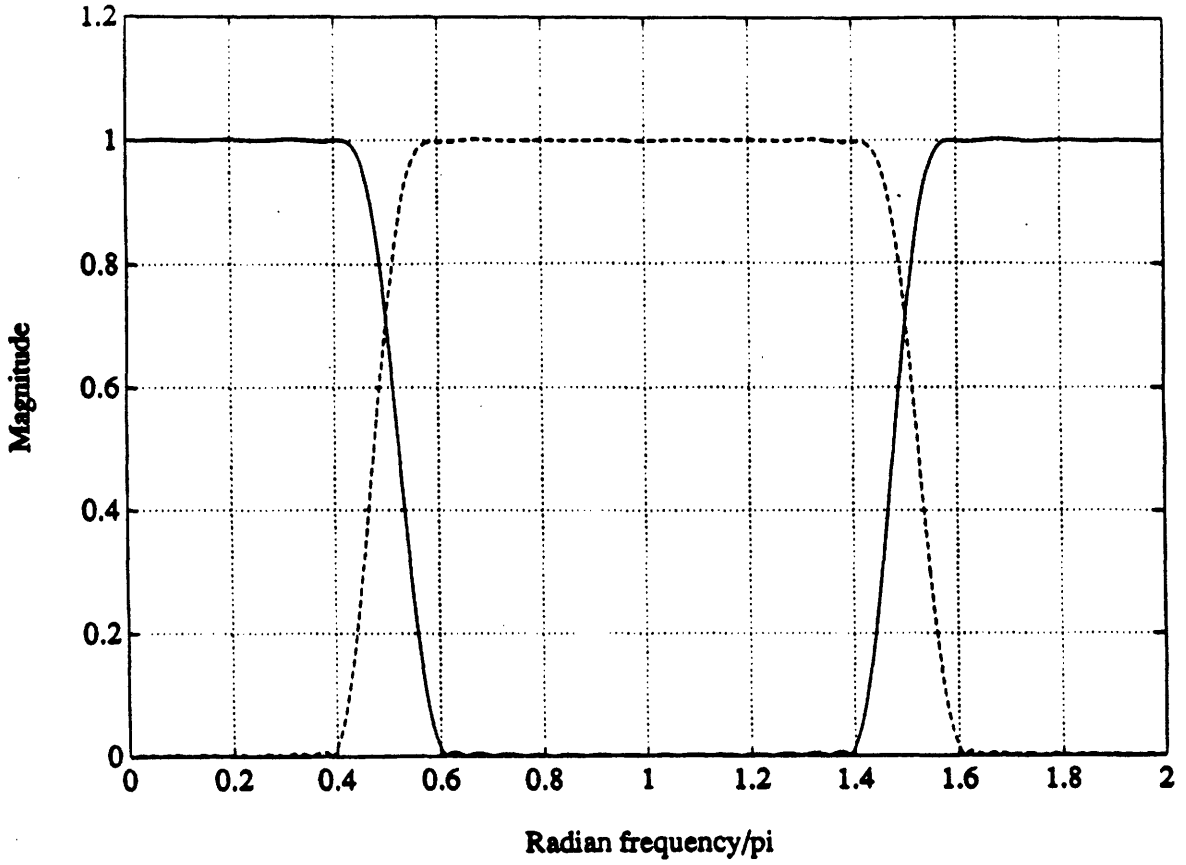

Figure 4.6(a): Our 32-tap analysis filters.

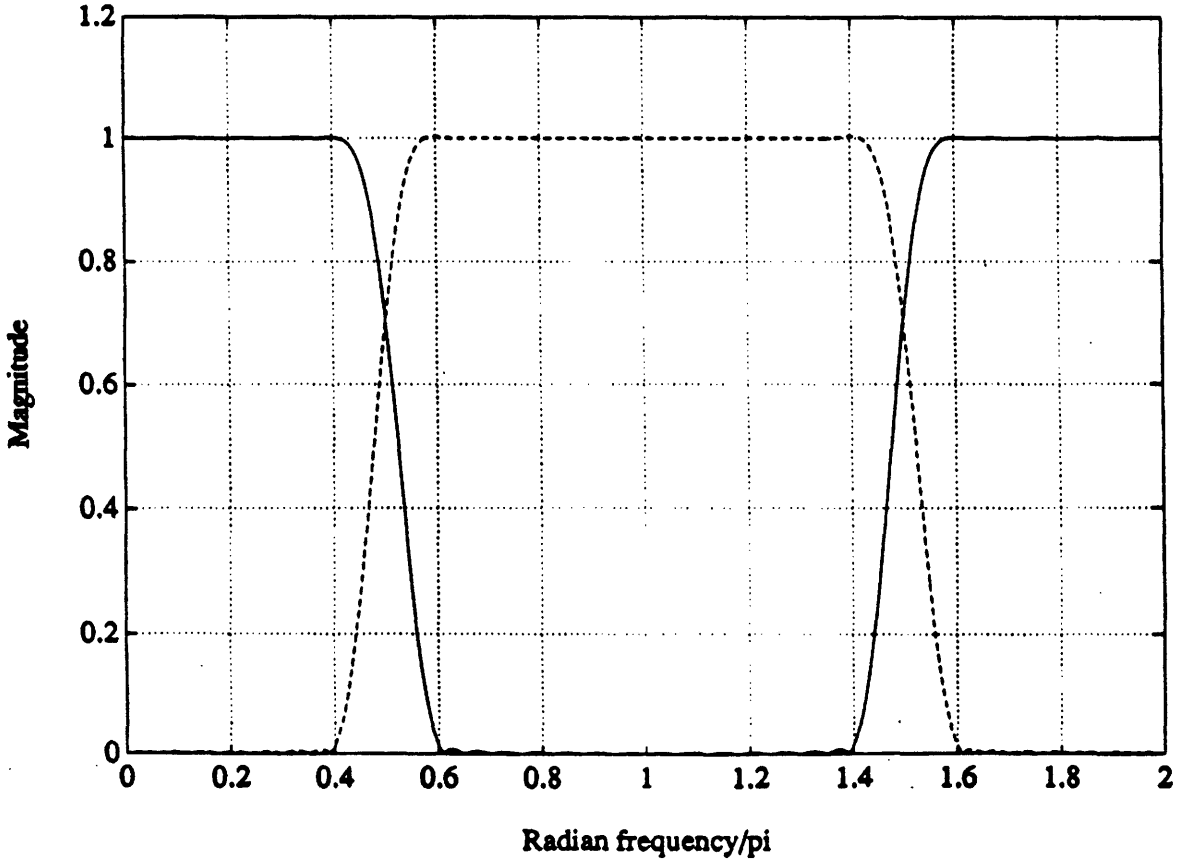

Figure 4.6(b): Our 32-tap synthesis filters.

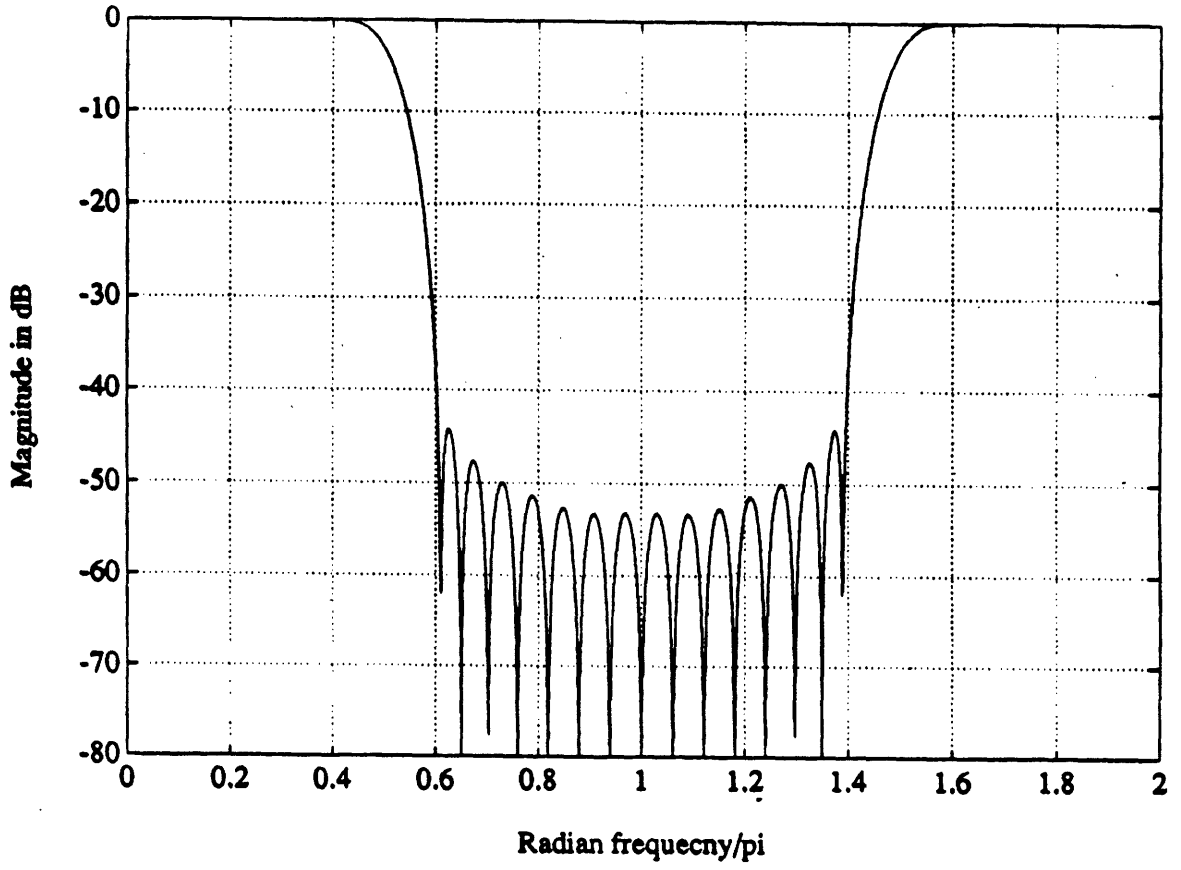

Figure 4.6(c): Our 32-tap, lowpass analysis filter.

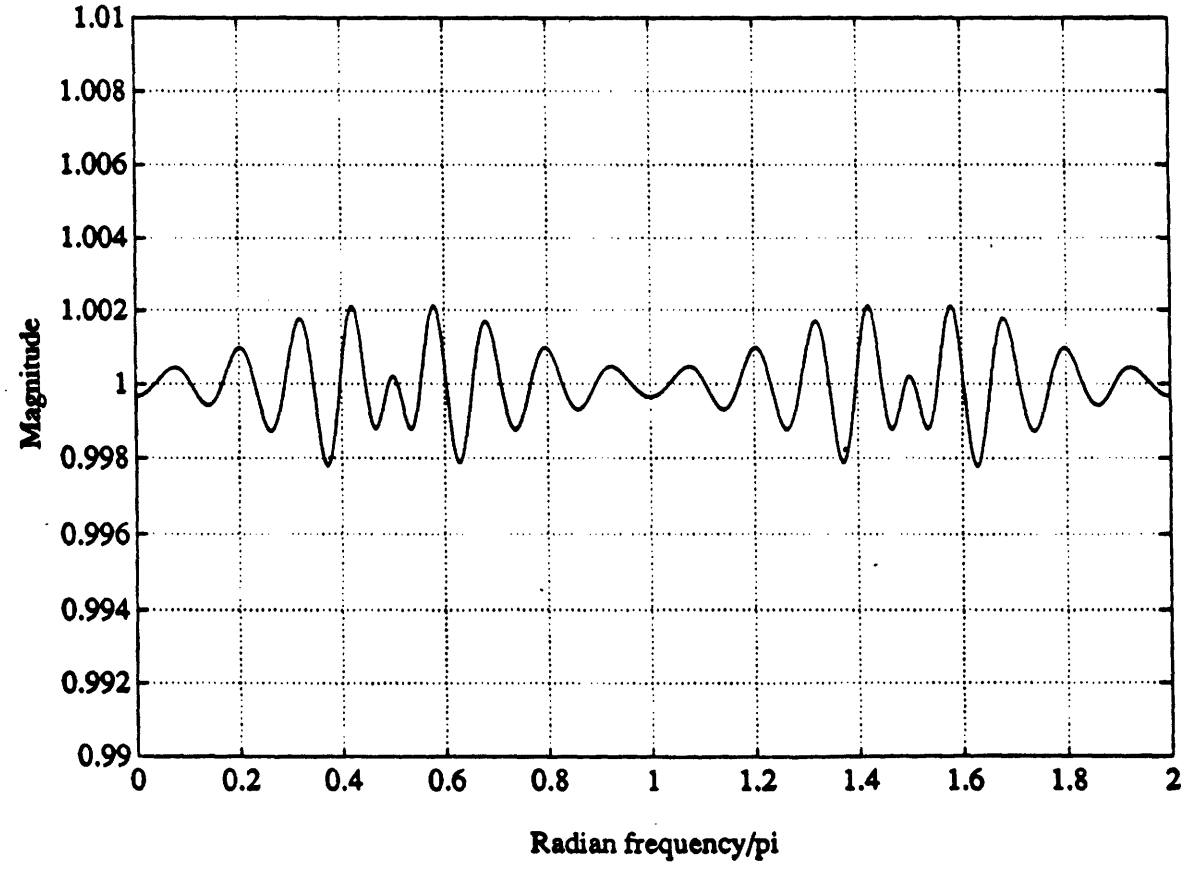

Figure 4.6(d): Overall response of our 32-tap QMF bank.

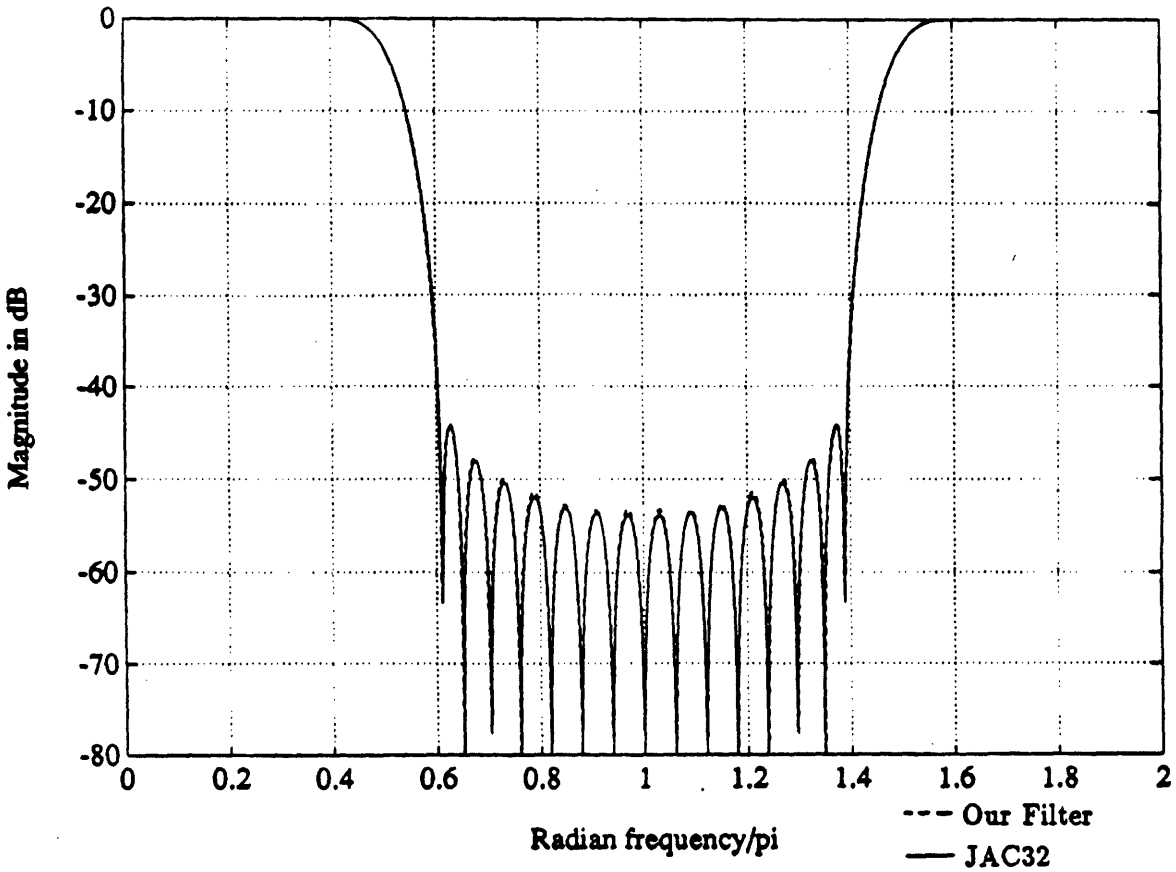

Figure 4.6(e): JAC32 *vs* our 32-tap, lowpass analysis filter.

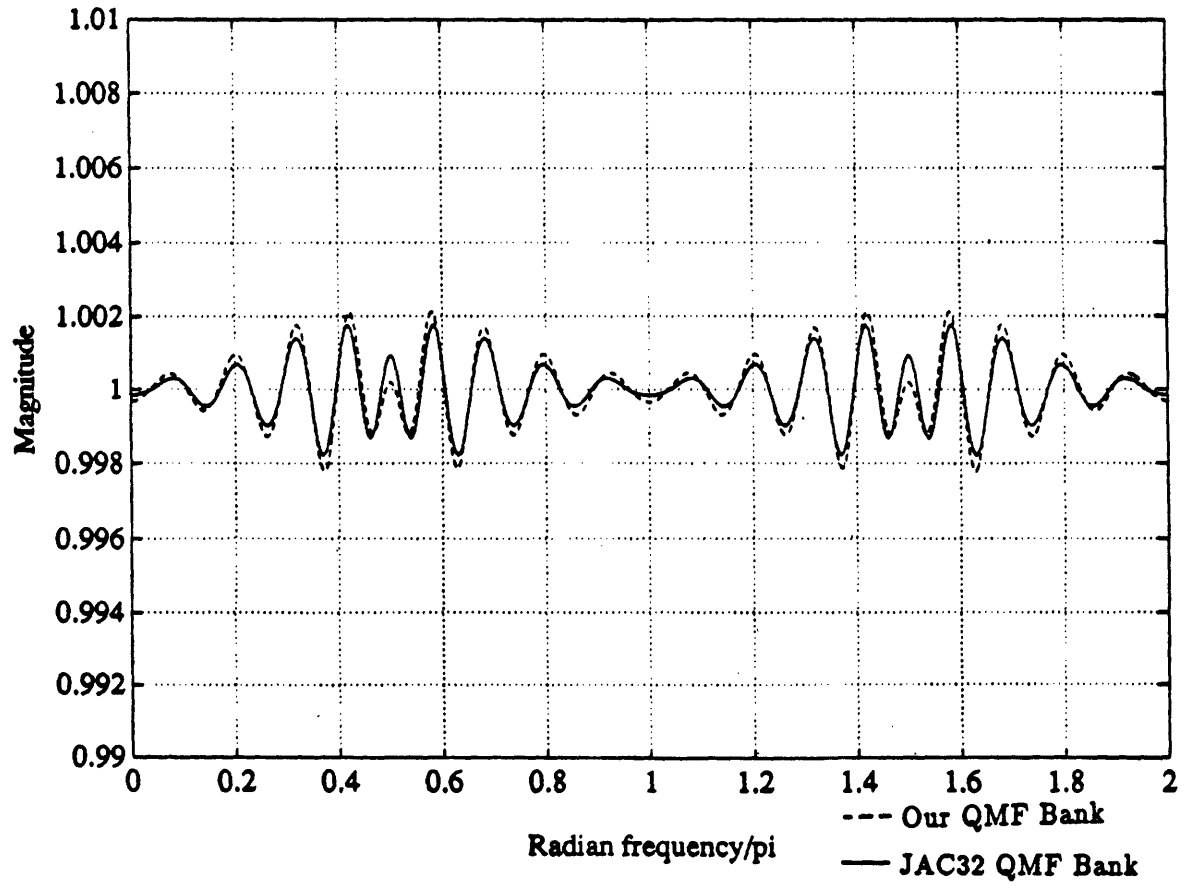

Figure 4.6(f): Overall response of our 32-tap QMF bank *vs* JAC32 QMF bank.

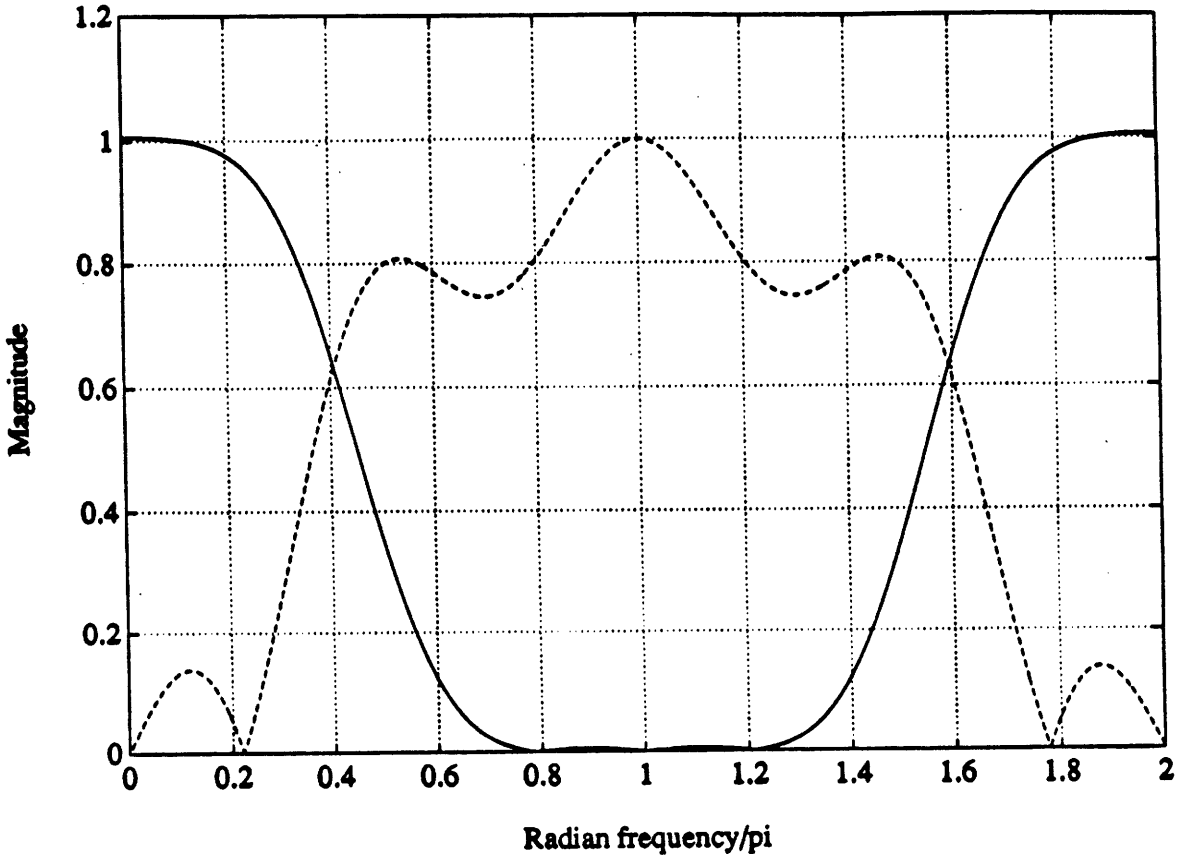

Figure 4.7(a): Analysis filters of the Guassian QMF bank.

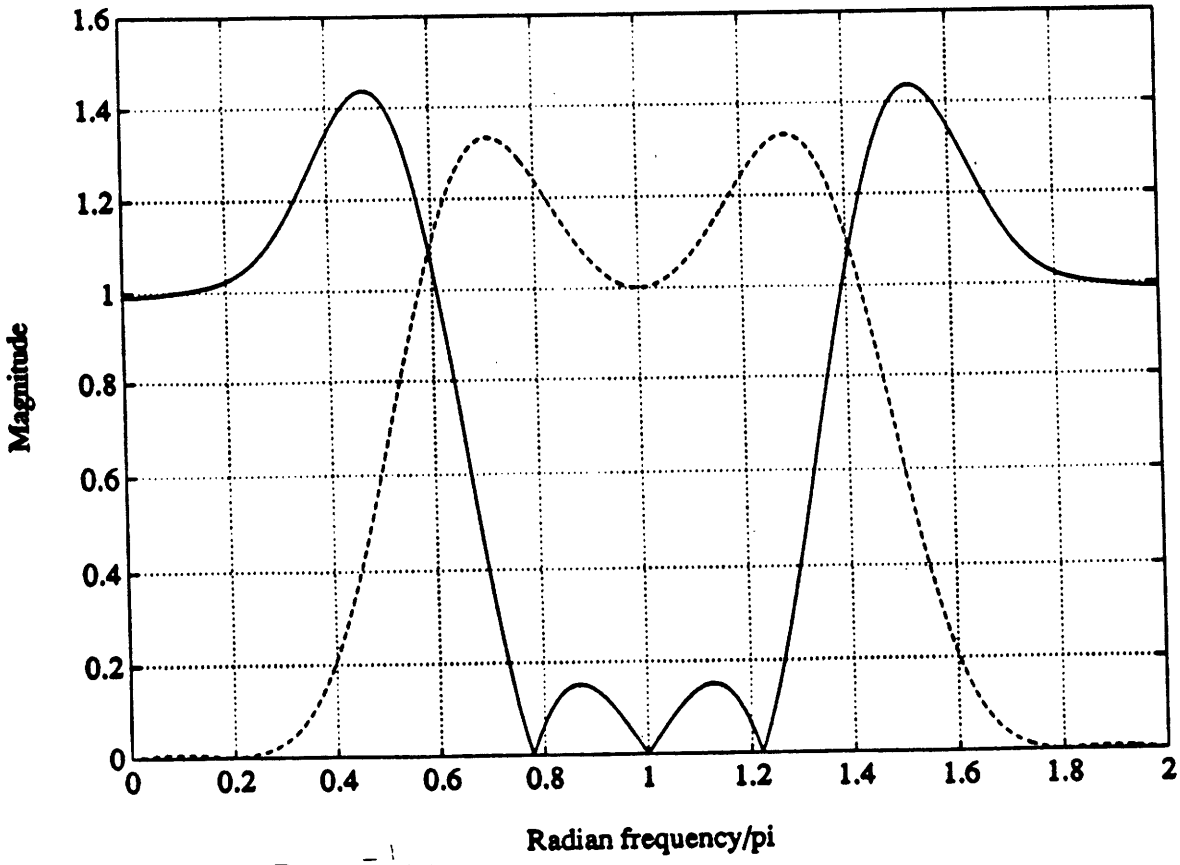

Figure 4.7(b): Synthesis filters of the Guassian QMF bank.

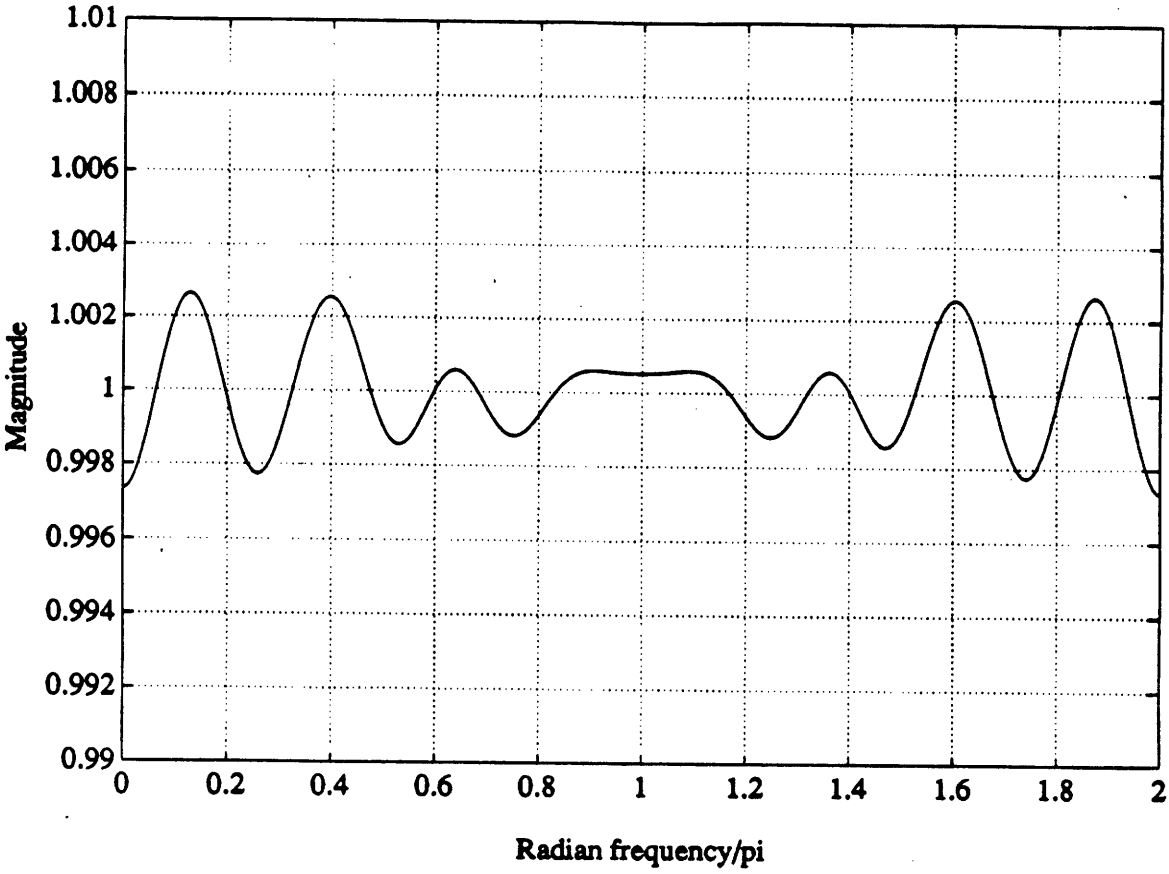

Figure 4.7(c):Overall response of the Guassian QMF bank.

Chapter 5

Conclusion

A new algorithm for designing two-band quadrature mirror filter banks was presented in this thesis. It was shown that this algorithm is computationally more efficient than the existing algorithms by a factor of eight. Moreover, this algorithm is not sensitive to the choice of initial guess. By using this algorithm, we were able to design two-band QMF banks in which the lowpass analysis filter was pre-described. Unfortunately, the frequency band separation of the resulting filters is not very good.

The ability to choose one of the filters in the QMF bank is extremely useful in certain applications. For example, in a compatible EDTV system where the NTSC picture is obtained by pre-filtering and subsampling the EDTV picture, it is essential to use pre-filters which do not cause ringing. However, the previous QMF design techniques provide no way of controlling this ringing. Using our algorithm, one can completely eliminate this ringing by an appropriately choosing the lowpass analysis filter.

The jointly-optimal analysis and synthesis QMF design algorithm is in the class of *coordinate descent* algorithms for minimization of a function of several variables. It was shown that the optimal analysis filters for a given pair of synthesis filters can be analytically computed as the solution to a system of linear equations. The computation of the optimum synthesis filters for a given pair of analysis filters is similarly possible.

The joint optimization algorithm is an iterative procedure which alternates between computing a pair of optimal analysis filters and optimal synthesis filters, until the error reduction is negligible.

The extension of this algorithm for designing M-band QMF banks is the most obvious area for future research. This extension should be quite easy, since the resulting optimization problem is a least-square one. In the M-band case, care must be taken as to how many of the filters can be fixed. It would also be useful to improve the frequency band separation of the filters, in the case that one of the filters is fixed.

Bibliography

[1] A. Croisier, D. Esteban, and Galand, "Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques," *International Conf. on Information Science and Systems*, Patras, Greece, 1976.

[2] D. Esteban and C. Galand, "Application of QMF's to split-band voice coding schemes," *Proc. of the IEEE Int. Conf. ASSP,* Hartford, Connecticut, pp. 191-195, May 1977.

[3] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*, Engelwood Cliffs. NJ: Prentice-Hall, 1983.

[4] A. Constantinides and R. Valenzuela, "An efficient and modular transmultiplexer design," *IEEE Trans. on Communication*, vol. 30, pp. 1629-1641, July 1982.

[5] P.P. Vaidyanathan, "QMF banks, M-Band Extensions, and Perfect- Reconstruction Techniques," *IEEE ASSP Magazine,* July 1987.

[6] D.G. Luenberger, *Linear and Non-linear Programming,* Engelwood Cliffs, NJ: Prentice-Hall, 1974.

[7] T. Abatzoglou and B. O'Donnel, "Minimization by coordinate descent," *J. Optimization Theory Appl.*, vol. 36, pp. 163-174, Feb 1982.

[8] V.K. Jain and R.E. Crochiere, "Quadrature mirror filters design in the time domain," *IEEE Trans. on ASSP,* vol. ASSP-32, pp. 353- 361, April 1984.

[9] J.D. Johnston, "A filter family designed for use in quadrature mirror filter banks," *Proc. IEEE Int. Conf. ASSP,* Apr. 1980, pp. 291- 294.

[10] G.R. Walsh, *Methods of optimization,* John Wiley and Sons, 1975.

[11] A.V. Oppenheim and R.W. Schafer, *Discrete-time signal processing,* Engelwood Cliffs, NJ: Prentice-Hall, 1989.

[12] G. Strang, *Intoduction to applied mathematics,* Wellesley MA: Wellesley-Cambridge Press, 1986.