Efficient Low Delay Filter Banks

Gerald Schuller and Mark J. T. Smith
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

This paper treats the problem of designing efficient FIR analysis-synthesis filter banks with system delays that can be pre-specified. A framework is introduced that is comprised of the cascade of several distinctive matrices with invertibility properties. The explicit form of the matrices guarantees computational efficiency and exact reconstruction, and allows for control over the system delay.

1 Introduction

The design of uniform-band analysis/synthesis filter banks for subband coding has been studied for many years [1], [2], [3], [4]. Yet design formulations that allow for a complete range of control are few and far between. Practical systems often dictate that the filter banks satisfy constraints on the general time-domain and frequency-domain specifications for the individual filters, constraints on the overall reconstruction quality, constraints on the computational efficiency, and sometimes limitations on the maximum overall system delay [6], [7]. In this paper, the basic components of a framework are presented in which the above-mentioned design flexibility is provided. A more complete treatment of this formulation can be found in [5].

In a conventional N-band filter bank with input $x(n)$ and analysis filters $h_k(n)$, the analysis outputs, $y_k(n)$, are obtained by filtering the input with $h_k(n)$ and decimating each band by a factor of N. For reconstruction, the N subband signals, $y_k(n)$, are upsampled by N, filtered with synthesis filters and summed. The same operations are performed implicitly in the approach introduced here. However, signals are represented as N dimensional vectors and the filters are represented by matrices.

For an N-band analysis/synthesis filter bank, the input is represented by an N-dimensional vector $\mathbf{x}(n)$ composed of the downsampled input components

$$\mathbf{x}(n) = [x(nN), x(nN+1), \ldots, x(nN+N-1)]$$

where n may be viewed as the index of the downsampled sequences $x(nN+m), \quad m = 0, 1, \ldots, N-1$. Taking the z-transform of each element we obtain the vector

$$\mathbf{X} = [X_0(z), \ldots, X_{N-1}(z)].$$

For every block of N input samples, N output samples are produced. These outputs are $y_k(n)$ where $k = 0, 1, \ldots, N-1$ and are also expressed as the vector $\mathbf{y}(n)$ with corresponding z-transform vector $\mathbf{Y}$. The analysis filters $h_k(n)$ that convert $\mathbf{x}(n)$ into $\mathbf{y}(n)$ are represented as an analysis polyphase filter matrix $\mathbf{P_a}$. These filters are represented explicitly as having a length that is an integer multiple of the block length N. In particular, the length is represented by LN where L is a positive integer. Filters with arbitrary lengths can be accommodated implicitly in the formulation by restricting an appropriate number of coefficients at the end to be zero.

In matrix form, the analysis section can be completely described by the equation

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{P_a}$$

where the elements of $\mathbf{P_a}$ are polynomials in z. $\mathbf{P_a}$ contains the analysis impulse response. Similarly the synthesis can be described by the equation

$$\mathbf{X} = \mathbf{Y} \cdot \mathbf{P_s}$$

where $\mathbf{P_s}$ is the synthesis polyphase matrix. When $\mathbf{P_s}$ is the matrix inverse of $\mathbf{P_a}$, we have exact reconstruction.

2 Modulated Filter Banks

For many modulated filter banks $\mathbf{P_a}$ and $\mathbf{P_s}$ can be decomposed into a transform matrix $\mathbf{T}$ with real or complex entries, and a “filter” matrix $\mathbf{F_a}$ or $\mathbf{F_s}$

$$\mathbf{P_a} = \mathbf{F_a} \cdot \mathbf{T} \;\;,\; \mathbf{P_s} = \mathbf{T}^{-1} \cdot \mathbf{F_s}.$$

The filter matrices, which are sparse with polynomial entries, can be further decomposed, resulting in efficient realizations.

Three important points are appropriate to mention at this time. First, the analysis-synthesis systems considered here are cosine-modulated type II and IV filter banks with filter vectors of the form

$$h_k(n) = h(n)\cos\left(\frac{\pi}{N}k(n + 0.5 + n_0)\right) \qquad (1)$$

for the type II and

$$h_k(n) = h(n)\cos\left(\frac{\pi}{N}(k + 0.5)(n + 0.5 + n_0)\right) \qquad (2)$$

($k = 0..N-1$) for the type IV, with a time offset n_0. The cosine modulated constraint guarantees fast implementation. Second, in general the inverse of $\mathbf{P_a}$ will be IIR since matrix inversion involves division by determinants and the matrix elements are z-domain polynomials. Third, in general the synthesis polyphase matrix will be anti-causal, i.e. it will contain terms in z (advances) in addition to terms in z^{-1} (delays). This problem is handled by introducing a constant delay z^{-1} to make the system causal. The interest here is in having computationally efficient filter banks where both analysis and synthesis filters are FIR. Consequently, a structural constraint is imposed on the matrices to guarantee that this is so. In particular, the polyphase matrices are represented as a cascade of several distinct coefficient matrices, delay matrices, and a transform matrix given by

$$\mathbf{F_a} = \left(\prod_{i=1}^{m} \mathbf{C}_i \cdot \mathbf{D}_i^2\right) \cdot \mathbf{F} \cdot \mathbf{D} \cdot \left(\prod_{i=1}^{n} \mathbf{G}_i\right) \qquad (3)$$

for the analysis filters and

$$\mathbf{F_s} = \left(\prod_{i=0}^{n-1} \mathbf{G}_{n-i}^{-1}\right) \cdot \mathbf{D}^{-1} \cdot z^{-1} \cdot \mathbf{F}^{-1} \cdot \left(\prod_{i=0}^{m-1} \mathbf{D}_{m-i}^{-2} \cdot z^{-2} \cdot \mathbf{C}_{m-i}^{-1}\right) \qquad (4)$$

for the synthesis filters. This form results in $n_0 = -N/2$. In the remaining portion of this section, we will examine the matrices in equations (3) and (4) with respect to computational efficiency, invertibility, and system delay.

Each of the component matrices in the equations above is sparse and has a special form. First, consider the matrix $\mathbf{T}$. This matrix is the DCT kernel. Both the forward and inverse DCTs are essentially the same and can be implemented with high efficiency. The matrix $\mathbf{D}$ is a delay matrix. It is defined as

$$\mathbf{D} = \left[\begin{array}{cccccc} z^{-1} & & & & & \\ & \ddots & & & & \\ & & z^{-1} & & & 0 \\ 0 & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{array}\right]. \qquad (5)$$

Its purpose is to collect all the delay elements that could contribute to IIR synthesis filters together so they can be handled separately. This matrix requires no arithmetic and has a simple causal inverse

$$\mathbf{D^{-1}} \cdot z^{-1} = \left[\begin{array}{cccccc} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & z^{-1} & & \\ & & & & \ddots & \\ & & & & & z^{-1} \end{array}\right]. \qquad (6)$$

Note that the causal inverse is $\mathbf{D^{-1}}z^{-1}$ and not $\mathbf{D^{-1}}$. The z^{-1} term is included with the inverse ensures that all terms in the matrix are causal.

The matrices $\mathbf{C}_i$ and $\mathbf{F}$ are coefficient matrices. That is they only contain coefficients as elements, i.e. no delays or polynomials. Consequently their inversion is a simple matrix inversion. These matrices have a well-structured form $\mathbf{F} =$

$$\left[\begin{array}{cccccc} & & d_0 & d_N & & \\ & \cdot^{\cdot^{\cdot}} & & & \ddots & \\ d_{N/2-1} & & & & & d_{N+N/2-1} \\ d_{N/2} & & & & & d_{N+N/2} \\ & \ddots & & & \cdot^{\cdot^{\cdot}} & \\ & & d_{N-1} & d_{2N-1} & & \end{array}\right] \qquad (7)$$

and

$$\mathbf{C}_i = \left[\begin{array}{cccccc} c_0^i & & & & & 1 \\ & \ddots & & & \cdot^{\cdot^{\cdot}} & \\ & & c_{N/2-1}^i & 1 & & \\ & & 1 & c_{N/2}^i & & \\ & \cdot^{\cdot^{\cdot}} & & & \ddots & \\ 1 & & & & & c_{N-1}^i \end{array}\right] \qquad (8)$$

where $d_0 \ldots d_{2N-1}$ and $c_0^i \ldots c_{N-1}^i$ are the (coefficient) elements of $\mathbf{F}$ and $\mathbf{C}_i$ respectively. For convenience, we call matrix $\mathbf{F}$ a diamond matrix due to its structure and $\mathbf{C}_i$ we call a bi-diagonal matrix. These matrices are very convenient because the inverse of a diamond matrix is a diamond matrix and the inverse of a bi-diagonal matrix is bi-diagonal. Moreover, the coefficients of the inverses can be computed analytically. For $\mathbf{F}$, the inverse is $\mathbf{F^{-1}} =$

$$\begin{bmatrix} & d'_{N/2-1} & d'_{N/2} & \\ & \iddots & \ddots & \\ d'_0 & & & d'_{N-1} \\ d'_N & & & d'_{2N-1} \\ & \ddots & \iddots & \\ & d'_{N+N/2-1} & d'_{N+N/2} & \end{bmatrix}$$

with

$$d'_j = \frac{d_{2N-1-j}}{d_j d_{2N-1-j} - d_{N+j} d_{N-1-j}}$$

$$d'_{N+j} = \frac{-d_{N-1-j}}{d_j d_{2N-1-j} - d_{N+j} d_{N-1-j}}$$

where $j = 0 \dots N-1$. For the $\mathbf{C}_i$, the inverse $\mathbf{C_i^{-1}} =$

$$\begin{bmatrix} c'^i_0 & & & & & c'^i_N \\ & \ddots & & & \iddots & \\ & & c'^i_{N/2-1} & c'^i_{N+N/2-1} & & \\ & & c'^i_{N+N/2} & c'^i_{N/2} & & \\ & \iddots & & & \ddots & \\ c'^i_{2N-1} & & & & & c'^i_{N-1} \end{bmatrix}$$

where

$$c'^i_j = \frac{c^i_{N-1-i}}{c^i_j c^i_{N-1-j} - 1} \qquad c'^i_N + j = \frac{-1}{c^i_j c^i_{N-1-j} - 1}$$

and $j = 0, \dots, N-1$.

The matrices $\mathbf{G}_i$ are also coefficient matrices, but special ones that allow for the control of the system delay. They are defined as

$$\mathbf{G}_i = \begin{bmatrix} g^i_0 z^{-1} & & & & & 1 \\ & \ddots & & & \iddots & \\ & & g^i_{N/2-1} z^{-1} & 1 & & \\ & & 1 & 0 & & \\ & \iddots & & & \ddots & \\ 1 & & & & & 0 \end{bmatrix}$$

with inverse $\mathbf{G}_i^{-1} =$

$$\begin{bmatrix} 0 & & & & & 1 \\ & \ddots & & & \iddots & \\ & & 0 & 1 & & \\ & & 1 & -g^i_{N/2-1} z^{-1} & & \\ & \iddots & & & \ddots & \\ 1 & & & & & -g^i_0 z^{-1} \end{bmatrix}$$

For minimum delay (without any delay matrices $\mathbf{D}$) the following decomposition can be used

$$\mathbf{F_a} = \mathbf{E}_0 \cdot \prod_{i=1}^{n-1} \mathbf{E}_i \tag{9}$$

and

$$\mathbf{F_s} = (\prod_{i=0}^{n-2} \mathbf{E}_{n-1-i}^{-1}) \cdot \mathbf{E}_0^{-1}$$

with $\mathbf{E}_i =$

$$\begin{bmatrix} 0 & & & & & e^i_N \\ & \ddots & & & \iddots & \\ & & 0 & e^i_{N+N/2-1} & & \\ & & e^i_{N+N/2} & e^i_{N/2} z^{-1} & & \\ & \iddots & & & \ddots & \\ e^i_{2N-1} & & & & & e^i_{N-1} z^{-1} \end{bmatrix} \tag{10}$$

where e_j can be real or complex. The inverse is $\mathbf{E}_i^{-1} =$

$$\begin{bmatrix} \hat{e}^i_0 z^{-1} & & & & & \hat{e}_N \\ & \ddots & & & \iddots & \\ & & \hat{e}^i_{N/2-1} z^{-1} & \hat{e}^i_{N+N/2-1} & & \\ & & \hat{e}^i_{N+N/2} & 0 & & \\ & \iddots & & & \ddots & \\ \hat{e}^i_{2N-1} & & & & & 0 \end{bmatrix}$$

with

$$\hat{e}^i_j = \frac{e^i_{N-1-j}}{-e^i_{N+j} e^i_{2N-1-j}}, \quad j = 0 ... N/2-1, \quad \text{and}$$

$$\hat{e}^i_{N+j} = \frac{e^i_{N+j}}{e^i_{N+j} e^i_{2N-1-j}}, \quad j = 0 ... N-1.$$

For the $\mathbf{E}_i$ matrices with $i > 0$ the anti-diagonal is 1, i.e., $e^i_{N+j} = 1$ for $j = 0..N-1$ and $i > 0$. This results in systems with time offsets of $n_0 = 0$ or $n_0 = N$.

The quintessential element in analyzing the system delay is the matrix $\mathbf{D}$. Recall from equation (6) that the inverse of $\mathbf{D}$ introduces an advance which must be offset with a delay of z^{-1} to keep the syste causal. Moreover, since these delays are effectivel at the lower sampling rate, each z^{-1} actually corr sponds to an N sample delay of the input. Each the $\mathbf{C}_i$ and $\mathbf{F}$ matrices has a delay term associate with it that contributes to the overall system dela The matrices $\mathbf{G}_i$ and $\mathbf{E}_i$, however, do not. They onl increase the filter length. Thus owing to equation (3 the filter length is $2Nm + 2N + nN$ and the syste delay is $2Nm + 2N - 1$. For minimum delay systen the filter length is $mN + 0.5N$ and the delay is $N -$ By selectively choosing values for m and n in equatic (3), or using equation (9), the overall system delay ca be pre-specified.

3 Filter Bank Design

The matrices shown in equations (3) and (4) completely characterize the filter bank. They lead to an efficient implementation, similar in structure to the implementation introduced by Malvar [2], which is very efficient. Because the analysis and synthesis matrices are all inverses, exact reconstruction is guaranteed. Moreover, the overall system delay is pre-determined by the matrices as well. The only part remaining is to impose frequency domain (and perhaps time domain) constraints on the system. In other words, the coefficients of the constituent matrices must be determined. This may be done by iterative optimization, where the matrix elements are the parameters being optimized. In this approach, an error function is created that represents the appropriate characteristics, such as stopband attenuation, transition width, perhaps a tapered impulse response or low ripple step response. Whatever the time- or frequency-domain characteristics are, the baseband filters in equation (1) and (2) can be designed by iterative optimization of the matrix coefficients. Optimization can be performed using standard library optimization algorithms. Since the synthesis involves inverting matrices, optimization should be done subject to the constraint that the synthesis matrices are invertible and reasonably well conditioned.

As an example, we show in Figure 1 a low delay filter bank designed using the formulation discussed in this paper. It shows the magnitude response of two 8-band analysis filters. The solid line corresponds to a system with delay 8 and filter length 8. The dashed line corresponds to a system with delay 8 and filter length 12. Note that for a fixed system delay, it is possible to improve the quality of the magnitude response beyond that of optimal 8-tap filters.

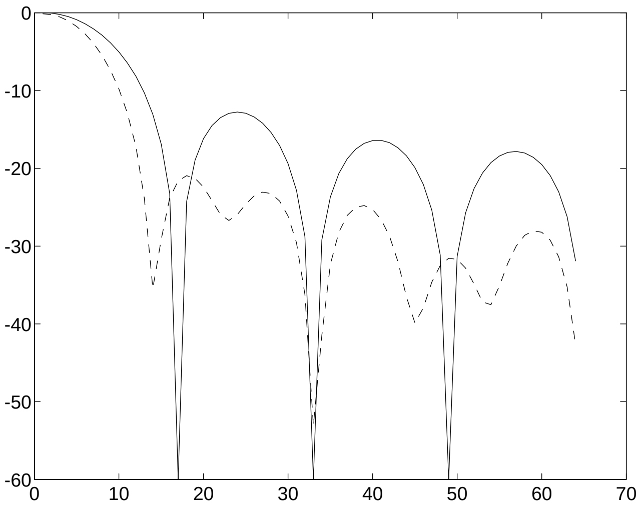

Figure 1: Magnitude response of two 8-band analysis filters. The solid line corresponds to a system with delay 8 and filter length 8. The dashed line corresponds to a system with delay 8 and filter length 12.

References

[1] J.P.Princen, A.B.Bradley, "Analysis/Synthesis Filter Bank Design Based on TDAC", *Trans. ASSP*, Oct., 1986, pp.1153-1161.

[2] H.S. Malvar, "Lapped Trans for Efficient T/S Coding", Trans. ASSP, June 1990, pp.969

[3] P.P. Vaidyanathan, "Theory and Des of M-chan Max Dec QMFs ... Trans. ASSP, Apr, 87

[4] T. Ramstad and J. Tanem, "Cos-Mod A/S Filter Banks..," ICASSP 1991

[5] G. Schuller and M. J. T. Smith, "A General Formulation for Mod PR..." NJIT 94 Symp on Appl of Subbands and Wavelets.

[6] K. Nayebi, T. Barnwell, and M. Smith, " Low Delay FIR F Banks..." Trans. SP, Jan 94.

[7] K. Nayebi, T.P. Barnwell,III, M.J.T. Smith, "On the Design of FIR A/S FB with High Comp Eff", Trans. SP, April 1994.

[8] T. Vaupel, "Transform Coding with Multiple Overlapping Blocks and Time Domain Aliasing Cancellation," Frequenz, 11-12, 1990.