MIXED-DOMAIN REDUCED ORDER IIR APPROXIMATION FOR FIR FILTER

Carrson C. Fung and Chi-Wah Kok

Dept. of Electrical and Electronic Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
email: c.fung@ieee.org

ABSTRACT

This paper proposed a mixed time and frequency domain approximation algorithm that approximates a high order FIR filter using an IIR filter with lower order. The high order FIR filter is first shortened by an eigenfilter based time domain equalizer. The equalization residue is approximated in the spectral domain using, for example, the minimax criteria, such that the two approximation filters will join together to form the resulting IIR filter. The mixed time and frequency design criteria is based on the observation that both time and frequency domain property are important and inherently inefficient when used alone. Design examples have shown that the proposed design method can obtain very good approximation in both the impulse response and spectral response of the given filter which cannot be obtained by other design techniques.

1. INTRODUCTION

Digital FIR filters have been used extensively due to its ease of design and implementation. Numerous design techniques, both in time and frequency domain, have been reported [1]-[5] and their behaviors are well understood. On the other hand, despite having less computational complexity when compared to that of FIR filter with similar design specification, digital IIR filters are generally less used. Due to their recursive nature, it is harder to design IIR filters that meet design specifications than their FIR counterparts. In addition, linear phase cannot be achieved in IIR filters unless they are non-causal. Greater care also has to be taken during implementation since any quantization error incurred in the process will be fed back in the recursive loop, amplifying the error as time persists. However, their computational efficiency has attracted a flurry of studies in the area.

The Prony method [5] and least squares inverse algorithm [6] are classified as time-domain based direct methods that treats the IIR design problem as an approximation problem by truncating an infinitely long rational function, i.e. the IIR filter impulse response, to approximate a polynomial with finite terms, i.e. the FIR filter impulse response. In [5], the problem of minimizing the error between the desired FIR filter and the IIR filter led to a set of over-determined linear equations where the solution is given by the pseudoinverse. [6] approached the problem by constraining the numerator of the IIR filter to be one of the spectral factor of the FIR filter and minimizing the error between the other spectral factor and the inverse of the denominator by approximating their product using a zero order constant unit polynomial.

This guarantees the resulting IIR filter to be stable. However, finding the minimum error requires $(M+1)^2$ multiplications, with M being the order of the denominator, due to the use of the Levinson algorithm. [12] also derived the solution to the design problem in the time domain which did not treat the problem as an approximation problem. Instead, the method in [12] minimizes a cost function of the error in the difference equation and solved the problem using an eigenfilter formulation to obtain a global minimum solution for the error. Model reduction techniques [7][8] based on state-space formulation that uses the same idea as the truncation methods above have also been applied to IIR filter design [9][10]. However, minimizing the truncation error will lead to an L_2 approximation problem which places no provision on the frequency response, resulting in poor spectral response. As a result, large differences in the spectral domain between the given FIR filter and the designed filter may occur. On the other hand, minimizing the spectral difference between the given FIR filter and that of the designed filter places no provision on the time-domain truncation problem and thus will result in awkward impulse response and undesired phase response. Therefore, time domain or frequency domain design methods alone are considered to be undesirable. We propose to tackle the IIR design problem in both the time and frequency domain. We called this the mixed-domain technique. We first approximate the desired FIR impulse response by designing a shortening filter in the time-domain under the L_2-norm criteria. The shortening filter turns out to be the denominator of the IIR filter. The problem can be formulated into a Rayleigh quotient which makes finding the global optimal easier because any optimal solution is the global one. The resulting coefficients are called eigenfilter [11]. A consequence of the shortening is an increase in oscillation around the band edge (Gibb's phenomenon) in the frequency domain. To mitigate such an effect, we propose to design the numerator of the IIR filter to minimize the maximum spectral difference between the shortened FIR filter and the numerator response of the resulting IIR filter, which is the L_∞-norm criteria. The resulting IIR filter is able to achieve lower stopband attenuation than the eigenfilter-based method in [12]. The derivation will be outlined in Section 2, followed by design examples and comparisons with the eigenfilter-based formulation [12] in Section 3. The paper is concluded in Section 4.

2. METHODOLOGY

The digital IIR filter design problem can be treated as an approximation problem that approximates a finite termed polynomial, i.e. the FIR filter impulse response $h_d(n)$, with a truncated rational function, i.e. the IIR filter impulse response $\frac{b(n)}{a(n)}$. This is called Padé approximation [13], which

is analogous to the Taylor series approximation of approximating a polynomial with another one of finite terms. The approximation error is then attributed to the remaining terms in the partial fraction $\frac{b(n)}{a(n)}$. The terms that are omitted are called the tail of the partial fraction. Define

$$\varepsilon^2(z) \triangleq \left| \frac{B(z)}{A(z)} - H_d(z) \right|^2$$

to be the squared error of the frequency response between the desired FIR filter, $H_d(z)$, and the IIR filter to be designed, $\frac{B(z)}{A(z)}$. For simplicity, an equivalent formulation

$$E^2(z) = A^2(z)\varepsilon^2(z) = |B(z) - A(z)H_d(z)|^2$$

known as the total squared error is considered. For the IIR filter to be feasible, $E^2(z)$ has to be minimized while the order of $b[n]$ and $a[n]$, the inverse z-transform of $B(z)$ and $A(z)$, respectively, has to be less than the order of $h_d[n]$, the inverse z-transform of $H_d(z)$. We can treat the IIR filter design problem as designing two FIR filters with coefficient $\mathbf{a} = [a[0], a[1], \ldots, a[n_a-1]]^\dagger$ and $\mathbf{b} = [b[0], b[1], \ldots, b[n_b-1]]^\dagger$, where $\dagger$ is the conjugate transpose, and n_a and n_b are the length of $\mathbf{a}$ and $\mathbf{b}$, respectively. Let N be the length of $h_d[n]$, choosing $n_a < N$, we can design $\mathbf{a}$ such that the resulting length of $H_d(z)A(z)$ is smaller than N so that $B(z)$ is designed to approximate $H_d(z)A(z)$ with $n_b < N$. Then $a[n]$ can be viewed as a shortening filter to shorten the length of $h_d[n]$ in order to approximate $b[n]$ as close as possible. As will be seen in the sequel, $\mathbf{a}$ is formulated as an orthonormal vector, thus the overall energy of the shortened response will be conserved after the shortening, resulting in an increase in amplitude of the shortened response in a small window. The combined effect of the truncation and increased amplitude will magnify the Gibb's phenomenon in the frequency domain. To find the IIR filter that approximates a shortened FIR filter and simultaneously mitigates the Gibb's phenomenon, we propose to design the coefficient $b[n]$ in the minimax sense so as to minimize the maximum peak amplitude in the frequency domain that is incurred during the shortening process. This will lower the stopband attenuation of the IIR filter in the frequency domain.

Before formulating the objective function, we must first define the following terms. Let $L_{\text{eff}} \triangleq N + n_a - 1$ be the length of the convolution result between $h_d[n]$ and $a[n]$. Then the shortening can be formulated by using the convolution matrix of $h_d[n]$, $\mathbf{H}_d \in \mathbb{R}_{L_{\text{eff}} \times n_a}$, the window function $\mathbf{W} \in \mathbb{R}_{L_{\text{eff}} \times L_{\text{eff}}}$ and its complement $\overline{\mathbf{W}} \in \mathbb{R}_{L_{\text{eff}} \times L_{\text{eff}}}$. They are defined as:

$$\mathbf{H}_d \triangleq \begin{bmatrix} h_d[0] & 0 & \cdots & 0 \\ h_d[1] & h_d[0] & \ddots & \vdots \\ \vdots & h_d[1] & \ddots & 0 \\ h_d[N-1] & \vdots & \ddots & h_d[0] \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \vdots & \ddots & h_d[N-1] \end{bmatrix},$$

$$\mathbf{W} \triangleq \begin{bmatrix} \mathbf{0}_\Delta & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{L_{\text{eff}}-p-\Delta} \end{bmatrix} \quad \text{and} \quad \overline{\mathbf{W}} \triangleq \mathbf{I}_{L_{\text{eff}}} - \mathbf{W},$$

where $\mathbf{I}_{L_{\text{eff}}}$ is an identity matrix of size L_{eff}. In the design examples below, the width of the window, p, is determined by the number of samples in the filtering result which have values that are within 10% to 100% of the peak value magnitude. However, other criteria can be used. The delay, Δ, has to be estimated since we do not know what combination of n_a and n_b will give us the optimum approximation result. To shorten the tail, we minimize the tail energy under the L_2-norm criteria. To do that, let $\mathbf{H}_1 = \mathbf{W}\mathbf{H}_d$ and $\mathbf{H}_2 = \chi\left(\mathbf{H}_d; \overline{\mathbf{W}}\right)$, where the $\chi(\mathbf{A}; \overline{\mathbf{W}})$ operator takes the matrix $\mathbf{A}$ as an input argument and the window matrix $\overline{\mathbf{W}}$ as a parameter such that it concatenates the 1st row of $\mathbf{A}$ to the Δth row and the $(\Delta+p+1)^{\text{th}}$ row to the last row of $\mathbf{A}$ into one matrix:

$$\mathbf{H}_d = \begin{bmatrix} \mathbf{H}_{2t} \\ \mathbf{H}_1 \\ \mathbf{H}_{2b} \end{bmatrix} \quad \Rightarrow \mathbf{H}_2 = \chi[\mathbf{H}d; \overline{\mathbf{W}}] = \begin{bmatrix} \mathbf{H}_{2t} \\ \mathbf{H}_{2b} \end{bmatrix}.$$

The cost function for designing the shortening filter can be stated as:

$$J_2 = \|\mathbf{H}_2\mathbf{a}\|_2. \tag{1}$$

Without loss of generality, we assumed that $\mathbf{a}^\dagger\mathbf{a} = 1$ to avoid trivial solution,

$$\begin{aligned} J_2 &= \frac{\mathbf{v}^\dagger\mathbf{H}_2^\dagger\mathbf{H}_2\mathbf{v}}{\mathbf{v}^\dagger\mathbf{v}} \\ &= \frac{\mathbf{v}^\dagger\mathbf{Q}\mathbf{v}}{\mathbf{v}^\dagger\mathbf{v}}, \end{aligned}$$

where $\mathbf{Q} \triangleq \mathbf{H}_2^\dagger\mathbf{H}_2$ is a Hermitian matrix. Since J_2 is a Rayleigh quotient, the minimum value is the minimum eigenvalue of $\mathbf{Q}$ and the corresponding eigenvector will render such a value for J_2. Therefore, to minimize J_2, the shortening filter coefficient $\mathbf{a}$ is chosen to be the eigenvector that corresponds to the minimum eigenvalue of $\mathbf{Q}$. Since $\mathbf{Q}$ is a Hermitian matrix, $\mathbf{a}$ is an orthonormal vector.

Since the shortening will contribute to an increase in oscillation in the band edge of the resulting filter in the frequency domain, and the possibility of imperfect shortening, we cannot simply use the shortening result as $\mathbf{b}$. Instead, we need to minimize the maximum of the ripple in order to minimize the effect the oscillation has on the frequency response. Thus, a mixed-domain design is needed. The cost function for the minimax optimization problem can be stated as:

$$J_\infty = \|\mathbf{W}_{M \times n_b}\mathbf{b} - \mathbf{W}_{M \times L_{\text{eff}}}\mathbf{H}_d\mathbf{a}\|_\infty, \tag{2}$$

where $\mathbf{W}_{M \times P}$ is a $M \times P$ DFT matrix. Before minimizing J_∞, we need to account for the fact that the boundary of the shortening filter may not monotonically decrease to zero. In order to minimize the boundary effect of the shortened filter, a Hamming window is used to extract the shortened filter impulse response that will be approximated by $\mathbf{b}$ in (2). The size of the Hamming window is large enough such that it extracts the entire unshortened portion of $\mathbf{H}_d\mathbf{a}$. In the design examples, the size is chosen to be twice as long as the length of $\mathbf{a}$. The Remez exchange algorithm in MATLAB is then used to minimize J_∞ using n_b and the frequency response of $\mathbf{H}_d\mathbf{a}$, sampled on a dense grid, as input parameters.

3. RESULTS

We present design examples using our method. The desired FIR filter response was generated by using the frequency sampling technique. The ideal amplitude response is:

$$|H_i(e^{j\omega})| = \begin{cases} 1 & , 0 \leq \omega \leq \omega_p \\ \frac{(\delta_s-1)\omega - \delta_s\omega_p + \omega_s}{\omega_s - \omega_p} & , \omega_p < \omega < \omega_s \\ \delta_s & , \omega_s \leq \omega \leq \pi \end{cases},$$

where ω_p is the passband band edge, ω_s is the stopband band edge, and δ_s is the stopband ripple size. ω_p, ω_s and δ_s are equal to 0.3, 0.6 and 1×10^{-4} (-80 dB), respectively, during the simulation. A first-order spline function is used for the transition function. The actual desired FIR filter impulse response, $h_d[n]$, is obtained by performing inverse DFT on $|H_i(e^{j\omega})|$ with the number of DFT points equals to the length of the desired FIR filter, N. Using $h_d[n]$, comparisons are made between our mixed-domain technique with the eigenfilter-based technique in [12].

Figure 1 shows the original and shortened impulse response $h_d[n]$ for length $N = 61$, $\omega_p = 0.3$, and $\omega_s = 0.6$. Figure 2 and 3 compares the frequency magnitude response between the mixed-domain designed IIR filter, the eigenfilter-based IIR filter [12] and the desired FIR filter when the $N = 61$ and 513, respectively, with $\omega_p = 0.3$, and $\omega_s = 0.6$. Table 1 shows the stopband attenuation of the IIR filter designed with the mixed-domain technique (labeled as *md*) and the IIR filter designed using the eigenfilter-based technique (labeled as *eig*). The stopband attenuation is measured from 0 dB to the 1st sidelobe. As shown in the table, our design was able to achieve higher stopband attenuation than that of the eigenfilter-based technique given the same order for the IIR filter. This is due to the minimization of the maximum amplitude in the frequency domain. Our design was able to approximate the passband of the desired FIR filter very well. However, the eigenfilter-based design suffers from occasional anomalies near the stopband band edge in the $N = 61, n_b = 9, n_a = 9$ and $N = 513, n_b = 17, n_a = 17$ cases. In the latter, a spike appears to account for the sudden drop in the transition band. The eigenfilter-based technique also suffers from higher computational complexity compared to our method since it requires performing the eigen-decomposition on a matrix with a high number of dimensions derived partly from the entire convolution matrix of $h_d[n]$. On the other hand, the eigen-decomposition performed for the mixed-domain design is on a smaller matrix which is derived from only the tail portion of the convolution result. For example, when $N = 513, n_b = 21, n_a = 21$, the eigenfilter-based method needs to carry out an eigen-decomposition on a 42×42 matrix, while our method only performs the decomposition on a 21×21 matrix. In fact, the dimension of matrix is equal to n_a so that we can trade-off between the shortening performance and the complexity. The frequency response of the resulting filter is shown to be very sensitive to the delay Δ and can be obtained by performing an exhaustive search. However, further investigation on an more efficient way of acquiring Δ is needed in order to further reduce the complexity. Figure 4 shows the phase response of mixed-domain designed IIR filter with $N = 513$, $w_p = 0.3$, $w_s = 0.6$ for both $n_b = n_a = 17$ and 21. It shows that our design cannot only approximate the amplitude response of the desired FIR response, but can also achieve near linear phase
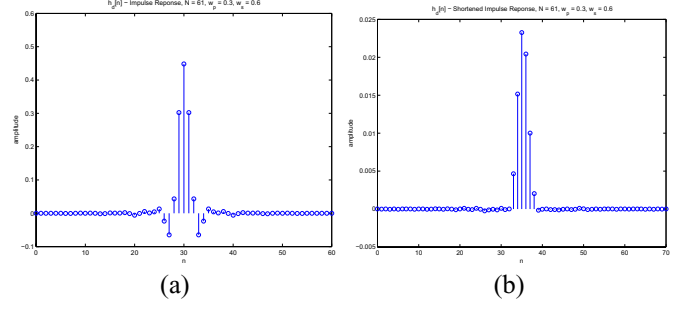
in the passband as well.

Figure 1: Impulse Response of Desired FIR filter, $N = 61$, $\omega_p = .3$, $\omega_s = 0.6$. a) Original. b) Shortened.
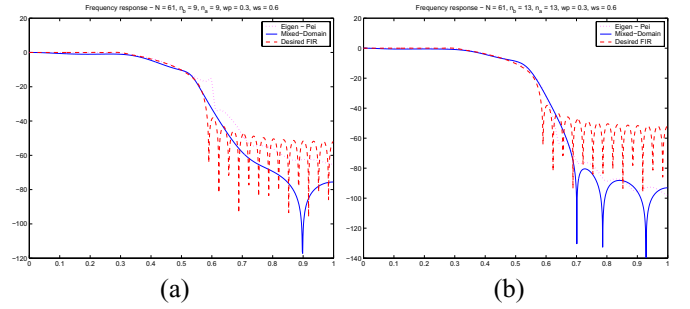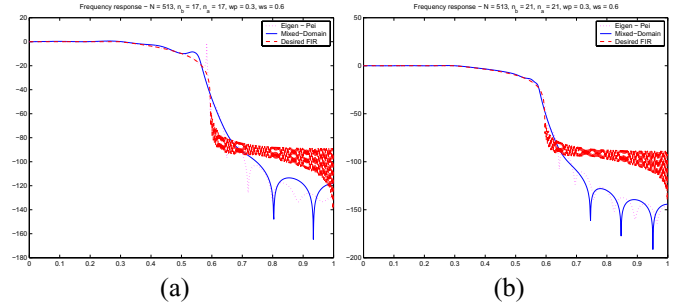

Figure 2: Frequency Response of IIR filters (Mixed-Domain and Eigenfilter ([12])), and Desired FIR filter, $N = 61$, $\omega_p = .3$, $\omega_s = 0.6$. a) $n_b = 9$, $n_a = 9$. b) $n_b = 13$, $n_a = 13$.

Figure 3: Frequency Response of IIR filters (Mixed-Domain and Eigenfilter ([12])), and Desired FIR filter, $N = 513$, $\omega_p = .3$, $\omega_s = 0.6$. a) $n_b = 17$, $n_a = 17$. b) $n_b = 21$, $n_a = 21$.

Table 1: Stopband Attenuation in dB for the IIR Filter using Mixed-Domain Method (*md*), and Eigenfilter-based Method ([12]) (*eig*), $\omega_p = 0.3$, $\omega_s = 0.6$

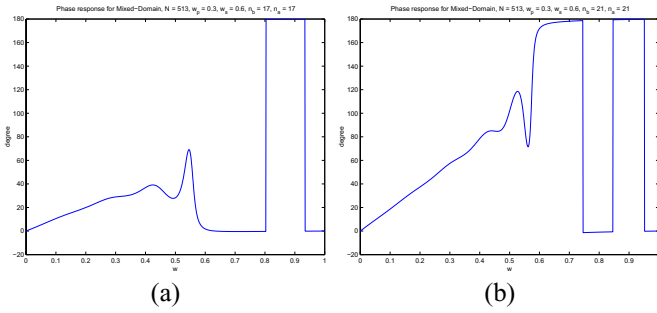| $N = 61$			$N = 513$		
	md	eig		md	eig
$n_b = 9, n_a = 9$	70	34	$n_b = 17, n_a = 17$	100	52
$n_b = 13, n_a = 13$	80.5	62	$n_b = 21, n_a = 21$	128	67

Figure 4: Phase Response of Mixed-Domain Designed IIR Filter filters , $N = 513$, $\omega_p = .3$, $\omega_s = 0.6$. a) $n_b = 17$, $n_a = 17$. b) $n_b = 21$, $n_a = 21$.

4. CONCLUSION

We have formulated the digital IIR filter design problem into an approximation problem by truncating or shortening the tail of the rational function (the FIR impulse response). Since a large peak error in the frequency domain is incurred from the shortening, by minimizing the maximum error in the frequency domain, a higher stopband attenuation can be achieved. The minimization in the frequency domain provided us with the design freedom to approximate a desired FIR filter as close as possible with reduction in the computational complexity. The shortening is formulated as an eigenfilter problem while the minimax optimization is done using the famous Remez exchange algorithm to offset the shortening effect. We have shown that our design achieves lower stopband attenuation than the eigenfilter-based method proposed by [12]. Our design can also achieve near linear phase in the passband.

REFERENCES

[1] T.W. Parks and C.S. Burrus, "Digital Filter Design", John Wiley & Sons, 1987.

[2] J.G. Proakis and D.G. Manolakis, "Digital Signal Processing: Principles, Algorithms, and Applications", 3rd ed., Prentice Hall, 1996.

[3] K. Steiglitz, "Computer-Aided Design of Recursive Digital Filters", *IEEE Trans. on Audio and Electroacoustics*, vol. 18(2), p. 123-129, Jun. 1970.

[4] A.G. Deczky, "Equripple and Minimax (Chebyshev) Approximations for Recursive Digital Filters", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-22(2), p. 98-111, Apr. 1974.

[5] C.S. Burrus and T.W. Parks, "Time Domain Design of Recursive Digital Filters", *IEEE Trans. on Audio and Electroacoustics*, Vol. 18(2), p. 137-141, Jun. 1970.

[6] A. Bester and E. Zeheb, "Reduced Order IIR Approximation to FIR Digital Filters", *IEEE Trans. on Signal Processing*, vol. 39(11), p. 2540-2544, Nov. 1991.

[7] B.C. Moore, "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction", *IEEE Trans. on Automatic Control*, vol. AC-26(1), p. 17-32, Feb. 1981.

[8] L. Pernebo and L.M. Silverman, "Model Reduction via Balanced State Space Representation", *IEEE Trans. on Automatic Control*, vol. AC-27(2), p. 382-387, Apr. 1982.

[9] B. Beliczynski, I. Kale, and G.D. Cain, "Approximation of FIR and IIR Digital Filters: An Algorithm Based on Balanced Model Reduction", *IEEE Trans. on Signal Processing*, vol. 44(3), p. 532-542, Mar. 1992.

[10] S. Holford and P. Agathoklis, "The Use of Model Reduction Techniques for Designing IIR Filters with Linear Phase in the Passband", *IEEE Trans. on Signal Processing*, vol. 44(10), p. 2396-2404, Oct. 1996.

[11] P.P. Vaidyanathan and T.Q. Nguyen, "Eigenfilters: A New Approach to Least-Squares FIR Filter Design and Applications Including Nyquist Filters", *IEEE Trans. on Circuits and Systems*, vol. CAS-34(1), Jan. 1987.

[12] S.-C. Pei and J.-J. Shyu, "Design of 1-D and 2-D IIR Eigenfilters", *IEEE Trans. on Signal Processing*, vol. 42(4), p. 962-966, Apr. 1994.

[13] E.W. Cheney, "Introduction to Approximation Theory", 2nd ed., Chelsea Publishing Co., 1982.