An Algorithm for the Optimization of Pipelined Recursive Digital Filters

Juha Yli-Kaakinen* and Tapio Saramäki*

Abstract — **In very large-scale integration (VLSI) implementations of infinite-impulse response (IIR) filters, the maximal achievable sampling rate is limited by the ratio between the number of delay elements and the latency of the arithmetic operations in the critical recursive loop. One alternative to increase the maximal sampling rate is to use the so-called pipelined IIR filters, where some first coefficients in the filter denominator are forced to be zero, increasing the number of delays in the critical loop. Typically, pipelined filters are generated from conventional IIR filters with the aid of proper transformations. The purpose of this paper is twofold. First, an efficient optimization technique is proposed for improving the performance of the transformed pipelined filters. Second, the latency is decreased by optimizing the filter coefficients, with the aid of another algorithm, to have very simple finite-precision forms, thereby increasing the sampling rate even further. The advantages of the proposed algorithms are illustrated by means of an example taken from the literature.**

Index Terms — **Optimization, pipelined recursive digital filters, IIR filters, multiplierless design, common subexpression elimination, VLSI implementations, clustered look-ahead transformation.**

1 Introduction

In general, infinite-impulse response (IIR) digital filters require much lower orders to achieve the given magnitude specifications than their finite-impulse response (FIR) counterparts, especially in cases requiring narrow transition bands. However, the maximum achievable sampling rate in very large-scale integration (VLSI) implementations of the IIR filters is limited by the ratio between the number of delay elements and the latency of the arithmetic operations in the critical feedback loop [1]. Hence, there exist two techniques for increasing the maximal sampling rate. The first one is to increase the number of delay elements in the critical loop, whereas the second one is to decrease the latency to perform the arithmetic operations in this loop.

Up to now, two methods have been proposed for generating filters with several delay elements in their recursive loops, namely, transformation techniques [2, 3] and special filter design techniques with certain constraints on the transfer function. One attractive approach, belonging to the latter of these methods, is to use frequency-masking techniques [4–6]. In these techniques, the overall filter is constructed using several subfilters where, instead of an unit delay, a block delay larger than a unit delay is used as the basic delay element. This contribution concentrates on the former of these two methods. In these techniques, the original transfer function is transformed into a new transfer function by adding several cancelling pole-zero pairs. The number of cancelling pole-zero pairs depends upon the transformation technique employed.

This contribution shows how to optimize, with the aid of the two-step procedure described in [7], the magnitude response of the pipelined IIR filters obtained using the transformation techniques [2, 3, 8–18]. As a first step, a pipelined IIR filter is used as a start-up solution for the second step, where the magnitude response is further improved with the aid of a proper nonlinear optimization algorithm. The ability of improving the magnitude response lies in the fact that by separating the cancelling pole-zero pairs gives extra degrees of freedom. The additional design margin can be allocated, e.g., for maximizing the stopband attenuation. Alternatively, it is possible to minimize the radius of the outermost pole, resulting in a smaller output roundoff noise due to multiplication errors. In addition, the proposed approach does not rely on the pole-zero cancellation. In finite-wordlength implementations, the inexact pole-zero cancellation will lead to errors in the realized spectrum and a time-variant behavior [19–21].

Furthermore, this contribution describes an algorithm for optimizing the filter coefficients to be expressible in very simple finite-precision representation forms, thereby decreasing the latency of the arithmetic operations. This results in a further increase in the maximal allowable sampling rate. If desired, an excess speed can also be traded for low power consumption through the use of power-supply voltage-scaling techniques [22].

2 Transformation Techniques

Transformation techniques generating high-speed IIR filters can be broadly classified into two categories, namely, *clustered look-ahead* (CLA) [8–10, 12, 16] and *scattered look-ahead* (SLA) [8, 15, 19, 23] transformations. The former technique will generally result in a lower computational complexity. It can be applied only for generating direct-form structures that generally share some poor finite-wordlength properties, such as a reasonably high coefficient sensitivity and a high output roundoff noise. In addition, the inexact pole-zero cancellation may increase the coefficient sensitivity even further. In the sequel, we concentrate on the optimization of only the high-speed IIR filters generated using CLA transformations. However, the proposed algorithm can be easily modified for optimizing also IIR filters obtained using SLA transformations.

In general, CLA transformation techniques involve the augmentation of an unpipelined stable filter described by

$$\widehat{H}(z) = \frac{A(z)}{B(z)} = \frac{\sum_{k=0}^{N} a_k z^{-k}}{1 + \sum_{k=1}^{N} b_k z^{-k}} \tag{1}$$

into the form

$$H(z) = \frac{C(z)}{D(z)} = \frac{A(z)Q(z)}{B(z)Q(z)} = \frac{\sum_{k=0}^{P} c_k z^{-k}}{1 + \sum_{k=M}^{P} d_k z^{-k}}, \tag{2}$$

*Institute of Signal Processing, Tampere University of Technology, P. O. Box 553, FIN-33101 Tampere, Finland. e-mail: juha.yli-kaakinen@tut.fi; ts@cs.tut.fi, Tel: +358 3 365 2930, Fax: +358 3 365 3087.

This work was supported by the Academy of Finland, project No. 44876 (Finnish Centre of Excellence Program (2000–2005)).

where M is the number of pipeline stages in the feedback loop. If the order of $Q(z)$ is K, then $C(z)$ as well as $D(z)$ are of order $P = N + K$. In most cases, $M < K$. The role of the cancelling pole-zero pairs determined by $Q(z)$ is to make the coefficients d_k of $D(z)$ for $k = 1, 2, \ldots, M-1$ equal to zero, resulting in a pipelined filter with M stages of pipelining. In order to guarantee the stability of $H(z)$, the roots included in $Q(z)$ must stay inside the unit circle. Since the first $M-1$ coefficients are zero-valued, the speedup factor of the resulting pipelined filter is roughly equal to M.

3 Optimization Problem

This section states the optimization problem for designing pipelined IIR filters with the transfer function given by Eq. (2). In this problem, the stopband attenuation is maximized subject to the constraint that in the passband the amplitude response of $H(z)$ stays within the same limits as that of $\widehat{H}(z)$ as given by Eq. (1). Efficient algorithms are then described for solving this problem.

3.1 Statement of the problem

For optimization purposes, it is beneficial to rewrite $H(z)$, as given by Eq. (2), as a cascade of first- and second-order sections as follows:

$$H(z) = k_0 \prod_{k=1}^{I_0} \frac{1 + a_{0k} z^{-1}}{1 + b_{0k} z^{-1}} \prod_{k=1}^{I_1} \frac{1 + a_{1k} z^{-1} + a_{2k} z^{-2}}{1 + b_{1k} z^{-1} + b_{2k} z^{-2}}, \quad (3)$$

where I_0 and I_1 are the number of first- and second-order sections, respectively, such that $P = I_0 + 2I_1$ and k_0 is a multiplicative constant for adjusting the desired level for the passband.

If $C(z)$ possesses I_0 real zeros at $z = r_k^{(\mathrm{zr})}$ for $k = 1, 2, \ldots, I_0$ and I_1 complex-conjugate zero pairs at $r_k^{(\mathrm{zc})} \exp(\pm j\theta_k^{(\mathrm{zc})})$ for $k = 1, 2, \ldots, I_1$ and $D(z)$ possesses I_0 real poles at $z = r_k^{(\mathrm{pr})}$ for $k = 1, 2, \ldots, I_0$ and I_1 complex-conjugate poles at $r_k^{(\mathrm{pc})} \exp(\pm j\theta_k^{(\mathrm{pc})})$ for $k = 1, 2, \ldots, I_1$, then

$$a_{0k} = -r_k^{(\mathrm{zr})} \quad \text{and} \quad b_{0k} = -r_k^{(\mathrm{pr})} \quad (4a)$$

for $k = 1, 2, \ldots, I_0$ and

$$a_{2k} = (r_k^{(\mathrm{zc})})^2, \qquad a_{1k} = -2r_k^{(\mathrm{zc})} \cos\theta_k^{(\mathrm{zc})}, \quad (4b)$$

$$b_{2k} = (r_k^{(\mathrm{pc})})^2, \quad \text{and} \quad b_{1k} = -2r_k^{(\mathrm{pc})} \cos\theta_k^{(\mathrm{pc})} \quad (4c)$$

for $k = 1, 2, \ldots, I_1$.

The squared-magnitude response of $H(z)$ is obtained by substituting $z = e^{j\omega}$ in $H(z)H(1/z)$, yielding

$$|H(\Phi, e^{j\omega})|^2 = k_0^2 \frac{|F_1(\omega)|^2 |F_2(\omega)|^2}{|G_1(\omega)|^2 |G_2(\omega)|^2}, \quad (5a)$$

where

$$|F_1(\omega)|^2 = \prod_{k=1}^{I_0} \left[1 - 2a_{0k}\cos\omega + a_{0k}^2\right], \quad (5b)$$

$$|F_2(\omega)|^2 = \prod_{k=1}^{I_1} \Big[\big((1+a_{2k})\cos\omega + a_{1k}\big)^2 + (1-a_{2k})^2 \sin\omega^2\Big], \quad (5c)$$

$$|G_1(\omega)|^2 = \prod_{k=1}^{I_0} \left[1 - 2b_{0k}\cos\omega + b_{0k}^2\right], \quad (5d)$$

$$|G_2(\omega)|^2 = \prod_{k=1}^{I_1} \Big[\big((1+b_{2k})\cos\omega + b_{1k}\big)^2 + (1-b_{2k})^2 \sin\omega^2\Big], \quad (5e)$$

and

$$\Phi = \big[k_0, a_{0k}, \ldots, a_{0I_0}, a_{1k}, \ldots, a_{1I_1}, a_{2k}, \ldots, a_{2I_1}, b_{0k}, \ldots, b_{0I_0}, b_{1k}, \ldots, b_{1I_1}, b_{2k}, \ldots, b_{2I_1}\big] \quad (6)$$

is the adjustable parameter vector containing the filter parameters.

The amplitude specifications are stated as follows:

$$1 - \delta_p \le |H(\Phi, e^{j\omega})| \le 1 + \delta_p \quad \text{for } \omega \in [0, \omega_p] \quad (7a)$$

$$|H(\Phi, e^{j\omega})| \le \delta_s \quad \text{for } \omega \in [\omega_s, \pi], \quad (7b)$$

where

$$|H(\Phi, e^{j\omega})| = \sqrt{|H(\Phi, e^{j\omega})|^2}. \quad (7c)$$

In order to the ensure the stability, all the poles must lie inside the unit circle. This implies that

$$|b_{0k}| \le 1 \quad \text{for } k = 1, 2, \ldots, I_0 \quad (8a)$$

and b_{1k} and b_{2k} for $k = 1, 2, \ldots, I_1$ lie in the triangular region defined by

$$b_{1k} - b_{2k} \le 1, \quad -b_{1k} - b_{2k} \le 1, \quad \text{and} \quad b_{2k} \le 1. \quad (8b)$$

Note that these constraints are linear with respect to the design parameters.

The resulting denominator polynomial $D(z)$ is obtained by convolving the first- and second-order sections as follows:

$$\begin{aligned} D(z) &= 1 + d_1 z^{-1} + \cdots + d_P z^{-P} \\ &= (1 + b_{01} z^{-1}) * \cdots * (1 + b_{0I_0} z^{-1}) \\ &\quad * (1 + b_{11} z^{-1} + b_{21} z^{-2}) * \cdots \\ &\quad * (1 + b_{1I_1} z^{-1} + b_{2I_1} z^{-2}). \end{aligned} \quad (9)$$

The optimization problem under consideration is the following:

Optimization problem: Given ω_p, ω_s, $P = I_0 + 2I_1$, and δ_p, as well as M, the number of pipelining stages in the feedback loop, find the adjustable parameter vector Φ, as given by Eq. (6), to minimize on $[\omega_s, \pi]$ the peak absolute value of $H(\Phi, \omega)$, as given by

$$\delta_s = \max_{\omega_s \le \omega \le \pi} |H(\Phi, e^{j\omega})|, \quad (10a)$$

subject to conditions of Eqs. (7a), (8), and

$$d_k = 0 \quad \text{for } k = 1, 2, \ldots, M-1. \quad (10b)$$

3.2 Optimization Algorithm

To solve this problem, we discretize the passband and stopband region into the frequency points $\omega_i \in [0, \omega_p]$ for $i = 1, 2, \ldots, L_p$ and $\omega_i \in [\omega_s, \pi]$ for $i = L_p+1, L_p+2, \ldots, L_p+L_s$. The resulting discrete optimization problem is to find Φ to minimize

$$\widehat{\delta}_s = \max_{L_p+1 \le i \le L_p+L_s} f_i(\Phi) \quad (11a)$$

subject to conditions

$$h_i(\Phi) \leq 0 \quad \text{for } i = 1, 2, \ldots, L_p, \tag{11b}$$

$$g_k(\Phi) = 0 \quad \text{for } k = 1, 2, \ldots, M-1, \tag{11c}$$

and

$$e_k(\Phi) = 0 \quad \text{for } k = 1, 2, \ldots, 2I_0 + 3I_1, \tag{11d}$$

where

$$f_i(\Phi) = \left|H(\Phi, e^{j\omega_i})\right|, \tag{11e}$$

$$h_i(\Phi) = \Big|\left|H(\Phi, e^{j\omega_i})\right| - 1\Big| - \delta_p, \tag{11f}$$

$$g_k(\Phi) = d_k, \tag{11g}$$

and

$$e_k(\Phi) = \begin{cases} b_{0k} - 1 & \text{for } k = 1, \ldots, I_0 \\ -b_{0k} - 1 & \text{for } k = I_0 + 1, \ldots, 2I_0 \\ b_{1k} - b_{2k} - 1 & \text{for } k = 2I_0 + 1, \ldots, 2I_0 + I_1 \\ -b_{1k} - b_{2k} - 1 & \text{for } k = 2I_0 + I_1 + 1, \ldots, \\ & \quad 2I_0 + 2I_1 \\ -b_{2k} - 1 & \text{for } k = 2I_0 + 2I_1 + 1, \ldots, \\ & \quad 2I_0 + 3I_1. \end{cases} \tag{11h}$$

The above-mentioned optimization problem can be solved using any of the three alternatives considered in Section 3 in [7]. Among these methods, the algorithm based on the use of sequential quadratic programming (SQP) methods is very efficient due to linearity of the constraints of Eq. (8). The convergence of the above algorithm to the optimum solution implies rather good initial values for the unknowns. A good initial starting-point filter for further optimization can be generated with the aid of the algorithm proposed by Lim and Liu in [12].

4 Numerical Examples

This section shows, by means of an example taken from the literature, the efficiency of the proposed algorithm. Another example is also included illustrating how to optimally quantize all the filter coefficients to have very simple representation forms.

4.1 Example 1

In order to illustrate the applicability of the proposed optimization algorithm, we consider the optimization of a sixth-order elliptic low-pass filter with five-stages of pipelining and $P = 13$ used by Lim and Liu [12] to illustrate their minimum order augmentation technique. The peak-to-peak passband ripple and stopband attenuation of the prototype filter are $A_p = 0.0419$ dB and $A_s = 47.6$ dB, respectively. The passband and stopband edges are at $\omega_p = 0.4\pi$ and at $\omega_s = 0.5\pi$, respectively.

The amplitude response and the pole-zero plot for the optimized filter is shown in Fig. 1. In this case, the optimization has been performed in such a manner that the stopband attenuation has been maximized subject to the constraint that the radius of the outermost pole is restricted to be less or equal to 0.928, the largest radius of the poles of the prototype filter.

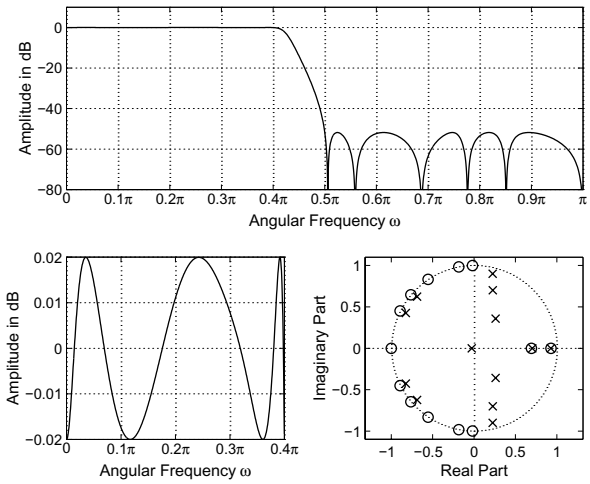

Figure 1: Some responses for the optimized pipelined IIR filter in Example 1.

$C(z)$	$D(z)$
$c_0 = 0.0294282$	$d_0 = 1$
$c_1 = 0.1238326$	$d_1 = 0$
$c_2 = 0.2718442$	$d_2 = 0$
$c_3 = 0.3742981$	$d_3 = 0$
$c_4 = 0.3124023$	$d_4 = 0$
$c_5 = 0.0620545$	$d_5 = -0.9522447$
$c_6 = -0.2415842$	$d_6 = 0.1445589$
$c_7 = -0.4033467$	$d_7 = 0.0039672$
$c_8 = -0.3388182$	$d_8 = -0.0274216$
$c_9 = -0.1388197$	$d_9 = -0.0978821$
$c_{10} = 0.0337677$	$d_{10} = 0.2721459$
$c_{11} = 0.0926421$	$d_{11} = -0.1251715$
$c_{12} = 0.0625382$	$d_{12} = 0.0405805$
$c_{13} = 0.0189096$	$d_{13} = 0.0012133$

Table 1: Optimized infinite-precision coefficients for the pipelined IIR filter in Example 1.

The stopband attenuation for the optimized filter is 51.83 dB. The values for the optimized coefficients are given in Table 1.

If the optimization is desired to be performed in such a manner that the radius of the outermost pole is minimized subject to the constraint that the filter still meets the given amplitude specifications, the radius of the outermost pole takes the value of 0.918.

4.2 Example 2

The proposed algorithm can also be modified for finding the minimum and maximum values for all the coefficients in such a manner that the given specifications are met. After finding these minimum and maximum values, it is straightforward to check whether in this parameter space there exist discrete coefficient values with which the given criteria are satisfied. This approach has been used by the authors of this paper for designing various kinds of digital filters [7].

As an example, we consider the specifications of Example 1. It is desired to use for coefficient representations four signed-powers-of-two terms with nine fractional bits. By sharing the common subexpressions within the coefficients, the

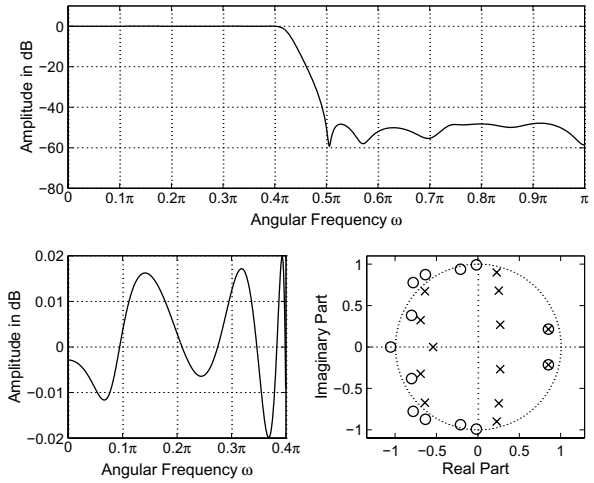

Figure 2: Some responses for the optimized finite-precision pipelined IIR filter in Example 2.

$C(z)$			$D(z)$		
c_a	$=$	$2^{-1} - 2^{-3}$	d_a	$=$	$2^{-1} + 2^{-3}$
c_b	$=$	$2^{-2} + 2^{-4}$	d_b	$=$	$2^{-2} - 2^{-4}$
c_c	$=$	$2^{-1} - 2^{-4}$			
c_0	$=$	$2^{-5} + 2^{-9}$	d_0	$=$	1
c_1	$=$	$2^{-3} + 2^{-6}$	d_1	$=$	0
c_2	$=$	c_b	d_2	$=$	0
c_3	$=$	$c_c - 2^{-5} \cdot c_b$	d_3	$=$	0
c_4	$=$	$c_a - 2^{-4} \cdot c_a$	d_4	$=$	0
c_5	$=$	$2^{-4} + 2^{-5} \cdot c_a$	d_5	$=$	$-d_a$
c_6	$=$	$-2^{-2} + 2^{-5} \cdot c_a$	d_6	$=$	0
c_7	$=$	$-c_a + 2^{-5} \cdot c_b$	d_7	$=$	$2^{-1} \cdot d_a - 2^{-4} \cdot d_b$
c_8	$=$	$-2^{-2} + 2^{-7}$	d_8	$=$	0
c_9	$=$	0	d_9	$=$	0
c_{10}	$=$	$2^{-1} \cdot c_b + 2^{-7}$	d_{10}	$=$	$d_b - 2^{-8}$
c_{11}	$=$	$2^{-1} \cdot c_a - 2^{-5} \cdot c_b$	d_{11}	$=$	0
c_{12}	$=$	$2^{-2} \cdot c_a + 2^{-6} \cdot c_a$	d_{12}	$=$	$-2^{-3} \cdot d_b$
c_{13}	$=$	$2^{-4} \cdot c_c$	d_{13}	$=$	$2^{-6} - 2^{-9}$

Table 2: Optimized finite-precision coefficients for the pipelined IIR filter in Example 2.

overall number of adders required to implement all the coefficient values becomes 19. In this case, the peak-to-peak passband ripple and stopband attenuation for the quantized filter are 0.0397 dB and 47.87 dB, respectively. The radius of the outermost pole is 0.928. The discrete coefficient values are given in Table 2. The magnitude response and the pole-zero plot for the finite-precision filter are shown in Fig. 2. Note that in this case the optimization has been performed in such a manner that the filter with scaled magnitude response meets the specifications.

References

[1] M. Renfors and Y. Neuvo, "The maximum sampling rate of digital filters under hardware speed constraints," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 196–202, Mar. 1981.

[2] K. K. Parhi and D. G. Messerschmitt, "Pipeline interleaving and parallelism in recursive digital filters — Part I: Pipelining using scattered look-ahead and decomposition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1099–1117, July 1989.

[3] K. K. Parhi and D. G. Messerschmitt, "Pipeline interleaving and parallelism in recursive digital filters — Part II: Pipelined incremental block filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1118–1134, July 1989.

[4] H. Johansson, *Synthesis and Realization of High-Speed Recursive Digital Filters*, Ph.D. dissertation no. 534, Linköping Studies Sci. Technol., Linköping Univ., Sweden, May 1998.

[5] H. Johansson and L. Wanhammar, "Filter structures composed of all-pass and FIR filters for interpolation and decimations by a factor of two," *IEEE Trans. Circuits Syst. II*, vol. 46, pp. 896–905, July 1999.

[6] H. Johansson and L. Wanhammar, "High-speed recursive digital filters based on the frequency-response masking approach," *IEEE Trans. Circuits Syst. II*, vol. 47, pp. 48–61, Jan. 2000.

[7] T. Saramäki and J. Yli-Kaakinen, "Design of digital filters and filter banks by optimization: Applications," in *Proc. X European Signal Processing Conf.*, Tampere, Finland, Sept. 5–8 2000, [Online]. Available HTTP: http://alpha.cc.tut.fi/~ylikaaki/EUSIPCO2000.pdf.

[8] H. B. Voelcker and E. E. Hartquist, "Digital filtering via block recursion," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 169–176, June 1970.

[9] P. M. Kogge and H. S. Stone, "A parallel algorithm for the efficient solution of a general class of recursive equations," *IEEE Trans. Comput.*, vol. C-22, pp. 786–793, Aug. 1973.

[10] H. H. Loomis and B. Sinha, "High-speed recursive digital filter realization," *IEEE Trans. Circuit, Syst., Signal Process.*, vol. 3, pp. 267–294, Sept. 1984.

[11] K. K. Parhi, C. Y. Wang, and A. P. Brown, "Synthesis of control circuits in folded pipelined DSP architectures," *IEEE J. Solid-State Circuits*, vol. 27, pp. 29–43, Jan. 1992.

[12] Y. C. Lim and B. Liu, "Pipelined recursive filter with minimum order augmentation," *IEEE Trans. Signal Processing*, vol. 40, pp. 1643–1651, July 1992.

[13] A. E. de la Serna, "Stability of time-domain pipelined IIR digital filters," Dipl. Eng. thesis, University of California, Davis, CA, 95616, Sept. 1993.

[14] C.-P. Lan and C.-W. Jen, "Efficient time domain synthesis of pipelined recursive filters," *IEEE Trans. Circuits Syst. II*, vol. 41, pp. 618–622, Sept. 1994.

[15] Y. C. Lim, "A new approach for deriving scattered coefficients of pipelined IIR filters," *IEEE Trans. Signal Processing*, vol. 43, pp. 2405–2406, Oct. 1995.

[16] M. A. Soderstrand and A. E. de la Serna, "Minimum denominator-multiplier pipelined recursive digital filters," *IEEE Trans. Circuits Syst. II*, vol. 42, pp. 666–672, Oct. 1995.

[17] Z. Jiang and A. N. Willson Jr., "Design and implementation of efficient pipelined IIR digital filters," *IEEE Trans. Signal Processing*, vol. 43, pp. 579–590, Mar. 1995.

[18] K. Chang, "Improved clustered look-ahead pipeling algorithm with minimum order augmentation," *IEEE Trans. Signal Processing*, vol. 45, pp. 2575–2579, Oct. 1997.

[19] K. K. Parhi, "Finite word effects in pipelined recursive filters," *IEEE Trans. Signal Processing*, vol. 39, pp. 1450–1454, June 1991.

[20] K. S. Arun and D. R. Wagner, "High-speed digital filtering: Structures and finite wordlength effects," *J. VLSI Signal Processing*, vol. 4, pp. 355–370, Nov. 1992.

[21] Y. Jang and S. P. Kim, "Block digital filter structures and their finite precision responses," *IEEE Trans. Circuits Syst. II*, vol. 43, pp. 495–506, July 1996.

[22] A. P. Chandrakasan and R. W. Brodersen, *Low Power Digital CMOS Desing*. Norwell, MA: Kluwer, 1995.

[23] K. K. Parhi, "Pipelining in dynamic programming architectures," *IEEE Trans. Signal Processing*, vol. 39, pp. 1442–1450, June 1991.