Finite Word-Length Effects of Pipelined Recursive Digital Filters

KyungHi Chang, *Member, IEEE*, and William G. Bliss, *Member, IEEE*

***Abstract*—Scattered look-ahead (SLA) pipelining is a new IIR filter structure that can achieve very high throughput, regardless of multiplier latency. However, the numerical properties of SLA have been largely unexplored. In this paper we analyze the finite word-length (FWL) performance of SLA filters under fixed-point arithmetic. To support this analysis, two new state variable descriptions (SVD) are introduced. First, a state variable based description especially suited for analysis of certain SLA structures, called the sectioned K and W description (SKWD), is defined by sectioning the noise contributions of the nonrecursive nodes. Second, a noncanonic state variable description (NCSVD) is introduced, which explicitly includes the pipelined delay variables in the state space. Roundoff noise (RON) and statistical coefficient quantization noise (SCQN) are derived under an independent pseudonoise source model. SCQN is shown to be interpretable as RON under the single length accumulation model, which enables the unification of RON and SCQN analysis. Analytic closed form solutions for first- and second-order direct form (DF) SLA filters are derived and compared to results from SLA minimum roundoff noise (MRON) structures. The SLA structures are all found to have generally good numerical properties, except that the DF SLA structure performs poorly in certain regions near the unit circle. However, the DF SLA structure actually performs better than the MRON form over much of the unit disk at a far smaller implementation cost.**

I. Introduction

Understanding finite word-length (FWL) effects, including roundoff noise (RON), coefficient quantization noise (CQN) and limit cycles (LC), is essential in designing IIR digital filters. IIR filters have often been avoided in real-time high throughput applications because of the problematic recursive updates of IIR algorithms. High speed digital filters, including FIR filters, often must be implemented using either a pipelining or parallel form of concurrent calculations. Concurrent algorithms typically must tradeoff latency for throughput, i.e., the transfer function is changed by the addition of a constant delay term. Because concurrent algorithms can have dramatically different internal calculations from those of their basis algorithms, it is important to study the numerical properties of these new algorithms.

Several stable look-ahead techniques for IIR filtering which break the requirement of multiply-accumulate (MAC) latency being less than the sampling period have been proposed [1]-[3]. In this paper we consider the pipelined technique called scattered look-ahead (SLA) with power-of-two decomposition [4]. This SLA technique relies on pole-zero cancellation in order to create a stable and pipelined (as opposed to block-parallel) class of algorithms. To achieve pipelining of depth M with this SLA technique requires an $O(log_2 M)$ increase in computational complexity. To reduce this computational overhead, Chung and Parhi [5] presented a design method which is not based on a rational polynomial transfer function, but rather on a bounded target frequency response. This method thus combines the approximation (pick $H(z)$) and synthesis (find structure that implements $H(z)$) problems into a joint step. Our paper will not address the numerical properties of this approach.

Little attention has been given to the numerical properties of the SLA filters. Parhi and Messerschmitt originally showed that for the state variable SLA structure, the RON due to the recursive section of the SLA filter is strictly decreasing with the number of loop pipelining stages M [4]. This result was based on comparing any non-SLA filter ($M = 1$) with state variable feedback matrix A, to the state variable description (SVD) SLA filter with feedback matrix A^M. We point out that their analysis is thus not applicable to the direct form (DF) SLA implementations, because the implied similarity transformation from A^M to companion form is not included in that analysis.

In [6] it was claimed that when considering RON from both the recursive and nonrecursive sections, the noise gain (NG) of the first-order filters always increases with M regardless of pole position, and that the total quantization noise decreases with M for poles near the unit circle, independent of θ, the angle of complex conjugate poles. However, these results were derived using the unscaled DF SLA structure, so they aren't relevant to real designs in fixed-point arithmetic. The importance of scaling in RON calculations is well documented in [7] and [8].

In this paper, we use the double-length accumulation RON model, i.e., the results of all multiplications are carried in full precision (double-length), are added in full precision, and are only then rounded to register-precision (single-length). In our RON model we explicitly include all RON sources, including those from the nonrecursive sections of the filter, being careful not to attribute noise to simple delay elements. We will demonstrate two state variable based methods to calculate scaling factors for SLA structures. The first method, which we call the sectioned **K** and **W** description (SKWD), is

Manuscript received June 24, 1992; revised December 20, 1993. This research was supported by the Texas Advanced Research Program under Grant number 999903-139. The associate editor coordinating the review of this paper amd approving it for publication was Prof. Tamal Bose.

K. H. Chang is with the Electronics and Telecommunications Research Institute, Taejon 305-600, Korea.

W. G. Bliss is with the Department of Electrical Engineering, Texas A&M University, College Station, TX 77843-3128 USA.

IEEE Log Number 9401906.

1053-587X/94$04.00 © 1994 IEEE

a special factored SVD (FSVD) method. The second method, which we call the noncanonic SVD (NCSVD), explicitly includes pipeline delays in the description. Quantization of filter coefficients from their ideal design values to fixed-point numbers is known to cause the realized filter to deviate from the desired response. The state-space approach has been applied to the CQN problem as well as the RON problem. In this paper we derive sufficient conditions for a statistical CQN analysis, and explicitly consider the dynamic range of coefficient values, not just the sensitivities of coefficients. This realistic hardware level modeling allows a unified analysis of the sum of RON and CQN.

II. Scattered Look-ahead Pipelining

The dynamic equation of an Nth-order canonic discrete time system is described by

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{b}u(k), \tag{1}$$

$$y(k) = \mathbf{c}\mathbf{x}(k) + du(k) \tag{2}$$

where $\mathbf{x}(k)$ is $N \times 1$, $\mathbf{A}$ is $N \times N$, $\mathbf{b}$ is $N \times 1$, $\mathbf{c}$ is $1 \times N$ and d, $u(k)$, and $y(k)$ are scalars.

First, we review the SLA algorithm with power-of-two decomposition [4]. The basis of all look-ahead algorithms is to unwind loop recursions. Applying M levels of look-ahead gives

$$\mathbf{x}(k+M) = \mathbf{A}^M\mathbf{x}(k) + \sum_{i=0}^{M-1} \mathbf{A}^i\mathbf{b}u(k+M-1-i). \tag{3}$$

The output calculations in the SLA algorithm are thus identical to those of the nonpipelined basis structure (2). The SLA structure achieves its pipelining by inserting $(M-1)$ extra poles for each original pole in $H(z)$. Note that throughout this paper, we will assume stable $H(z)$, i.e., system eigenvalues inside the unit circle. The extra poles are at the same radius, r, as the original pole and are at angles, θ, that are equally spaced around 2π

$$\theta_i = \theta + i2\pi/M, \qquad \text{where } i = 1, 2, \cdots, (M-1). \tag{4}$$

The apparent denominator polynomial of the SLA filter is thus a polynomial in z^{-M}, which allows latency M in the recursion. The sparsity of the denominator polynomial also means that its computational complexity is low, i.e., there are only N coefficients for an Nth-order filter. To achieve I/O equivalence (in perfect arithmetic), $N(M-1)$ zeros must be inserted to exactly cancel the added poles. This numerator polynomial isn't sparse, and thus it represents a large number of calculations. However, by choosing M to be a power-of-two, the numerator was factored as $\log_2 M$ polynomials with N nontrivial coefficients each [4]. The computational complexity of SLA with power-of-two decomposition is thus $N(1+\log_2 M)$ for the DF all-pole filter. The nonrecursive portion of the filter which implements the pole-canceling zeros can be pipelined to any required degree to facilitate MAC implementation. Similar computation-saving numerator decompositions have been shown for any composite $M = M_1M_2$, so it isn't required that M be a power of two [3], [4].

In contrast to the SLA technique, the clustered look-ahead (CLA) technique, whether DF or general SVD form, doesn't necessarily map its extra poles inside the unit disk, so the recursive part of a CLA filter may be unstable, regardless of the stability of the original filter. However, recent development of CLA algorithm in DF by Lim and Liu [9] solves the problem of the instability of resultant filters with the minimum augmentation of pipelining stages. We don't consider CLA algorithms in this paper.

III. Roundoff Noise and Scaling

Roundoff noise (RON) is an important FWL effect due to the quantization of internal accumulations in a digital filter. We follow the notation and assumptions of Roberts and Mullis [7]. In particular, trivial summations in the SFG which correspond to only passing a state variable from one delay to another, with no added inputs, do not generate any RON. As the SLA structure has many of these "pure delays," it is important to recognize that they don't generate additional noise.

Denote the response at the ith state variable, $x_i(k)$, due to a unit pulse at the input, $u(k)$, as $f_i(k)$, and the response at the output, $y(k)$, due to a unit pulse at the i'th state variable as $g_i(k)$. Define $\mathbf{K}$ and $\mathbf{W}$ as

$$\mathbf{K} = \sum_{k=1}^{\infty} \mathbf{f}(k)\mathbf{f}^T(k), \tag{5}$$

$$\mathbf{W} = \sum_{k=1}^{\infty} \mathbf{g}^T(k)\mathbf{g}(k). \tag{6}$$

Then $\mathbf{K}$ is also the covariance matrix of the state-space vector when the input to the filter is a stationary white noise with zero mean and unit variance. Assuming that double-length internal accumulations are used before rounding to single-length at state variables, the total output RON for the l_2-scaled filter is (following convention and omitting the RON due to the output node)

$$\sigma_{RON}^2 = \sum_{n=1}^{N} \sigma_n^2 = \sigma_e^2 \sum_{n=1}^{N} \mathbf{W}_s(n,n)$$
$$= \delta^2\left(\frac{q^2}{12}\right) \sum_{n=1}^{N} \mathbf{W}(n,n)\mathbf{K}(n,n) \tag{7}$$

where the sum from 1 to N implies that all N state variables do in fact introduce RON. Roundoff noise gain, which drops the multiplicative factor $\delta^2 q^2/12$ from σ_{RON}^2, becomes

$$NG = \sum_{n=1}^{N} \mathbf{W}(n,n)\mathbf{K}(n,n). \tag{8}$$

The fallacy of attempting to calculate RON on an unscaled filter by

$$\left(\frac{q^2}{12}\right) \sum_{n=1}^{N} \mathbf{W}(n,n) \tag{9}$$

can be demonstrated by considering a diagonal transformation, T, where δ is chosen to be a small (and unreasonable) number, e.g., $\delta = 0.5$. Such a choice reduces the apparent RON power by 20 dB when compared to the reasonable choice of $\delta = 5$. The apparent gain is illusory because the finite-length registers really overflow or saturate, or alternately more bits are really used at these registers.

In a theoretical model of VLSI computing, the bound for integer multiplication of B bits has been shown to be $AT^2 = O(B^2)$, where A is area, and T is processing time [10]. In many practical technologies, for reasonable B, and where the multiplier can be pipelined, the product of area and throughput is proportional to B^2. Thus a moderate increase in B can cause a large increase in area.

IV. Roundoff Noise of SLA Filters

In this section we derive analytic forms for the RON of first- and second-order SLA structures. To simplify the analysis, we introduce two new techniques, called the sectioned K and W description (SKWD) and the noncanonic state variable description (NCSVD).

A. First-Order Pipelined Filter

We first analyze the RON of the single pole DF SLA structure. The SKWD technique simplifies the analysis of certain SLA structures by sectioning the noise gain contribution of multiple paths to the output from a given state variable. Again, define $f_i(k)$ as the impulse response at summing node i. But for the SKWD, define $g_{ij}(k)$ as the response due to an impulse at node i taking the jth path to $y(k)$. Thus, $\sum_j g_{ij}(k) = g_i(k)$ in the normal SVD notation. Define $K_i = \sum_k f_i^2(k)$, the noise power gain from the input to node i. Define $W_{ij} = \sum_k g_{ij}^2(k)$ as the noise power gain from node i through the jth path to the output. Certain SLA structures, including the single pole DF structure, by inspection have multiple output paths which are orthogonal, i.e., $g_{ij_1}(k) \cdot g_{ij_2}(k) = 0$ if $j_1 \neq j_2$. This orthogonality allows the summation of powers over the j index, so the noise gain becomes

$$NG = \sum_i \sum_{j(i)} W_{ij} K_i. \tag{10}$$

Consider the conventional DF filter ($M = 1$) with a single pole at $z = a$. Then a RON source $e_1(k)$ is located at the summing node.

$$f_1(k) = \{0\, b\, ab\, a^2b \cdots\} \tag{11}$$

$$g_{11}(k) = \{c\, ca\, ca^2\, ca^3 \cdots\} \tag{12}$$

$$K_1 = b^2 + a^2b^2 + a^4b^2 + \cdots \tag{13}$$

$$W_{11} = c^2 + c^2a^2 + c^2a^4 + \cdots \tag{14}$$

The scaled input gain is

$$b_s = b/\delta K_1^{1/2} = (1-a^2)^{1/2}/\delta \tag{15}$$

and the scaled output gain is

$$c_s = c\delta K_1^{1/2} = bc\delta(1-a^2)^{-1/2} \tag{16}$$

while a and d are unchanged by scaling. Then the noise gain of l_2-scaled first-order conventional DF filter assuming $|a| < 1$ is

$$\begin{aligned} NG(M=1) &= W_{11}K_1 \,|_{M=1}, \\ &= \frac{b^2c^2}{(1-a^2)^2}. \end{aligned} \tag{17}$$

For the first-order pipelined filter with two loop pipelining stages ($M = 2$), the noise gain is

$$\begin{aligned} NG(M=2) &= W_{11}K_1 \,|_{M=2}, \\ &= \frac{b^2c^2}{(1-a^2)(1-a^4)}. \end{aligned} \tag{18}$$

For the first-order pipelined filter with four loop pipelining stages, we can have l_2-scaled filter coefficients and SFG by using the partitioned scaling, SKWD. There is a summing node in the nonrecursive section of the filter, so there are now two uncorrelated RON sources $e_1(k)$ and $e_2(k)$ (under the assumption of small q and sufficiently energetic input). Among four total summation nodes, $e_1(k)$ and $e_2(k)$ are located at the second and the fourth summation nodes, from the output node, respectively. Observe that the two paths from the single noise source $e_2(k)$ in the nonrecursive section of the filter to the output have orthogonal sequences, so the total noise power contribution is the sum of the two powers. Let p denote the number of the paths from each noise source in nonrecursive section to the output.

$$\begin{aligned} \sigma^2_{RON} &= \left(\frac{q^2}{12}\right) \sum_{i=1}^{\log_2 M} \delta_i^2 \left(\sum_{j=1}^{p} W_{ij}\right) K_i \\ &= \left(\frac{q^2}{12}\right) \delta^2 [W_{11}K_1 + (W_{21} + W_{22})K_2] \,|_{M=4} \end{aligned} \tag{19}$$

Note that we are defining separate K and W for each state variable group associated with a RON source. By inspection, the sectioned response sequences are

$$f_1(k) = \{0\,0\,0\,0\, b\, ab\, a^2b\, a^3b \cdots\}, \tag{20}$$

$$f_2(k) = \{b\, ab\, 0\,0\,0 \cdots\}, \tag{21}$$

$$g_{11}(k) = \{c\,0\,0\,0\, ca^4 \cdots\}, \tag{22}$$

$$g_{21}(k) = \{0\,0\,0\,0\,0\,0\, ca^2\, 0\,0\,0\, ca^6 \cdots\}, \tag{23}$$

$$g_{22}(k) = \{0\,0\,0\,0\, c\, 0\,0\,0\, ca^4 \cdots\}. \tag{24}$$

Therefore,

$$\begin{aligned} NG(M=4) &= W_{11}K_1 + (W_{21} + W_{22})K_2 \,|_{M=4}, \\ &= b^2c^2 \frac{2 - a^8}{(1-a^2)(1-a^8)}. \end{aligned} \tag{25}$$

For M=8, we have three true RON sources $e_1(k)$, $e_2(k)$, and $e_3(k)$. By using the SKWD again, the roundoff noise gain is obtained as

$$\begin{aligned} NG(M=8) &= W_{11}K_1 + (W_{21} + W_{22})K_2 \\ &\quad + (W_{31} + W_{32} + W_{33})K_3 \,|_{M=8}, \\ &= b^2c^2 \frac{3 - 2a^{16}}{(1-a^2)(1-a^{16})}. \end{aligned} \tag{26}$$

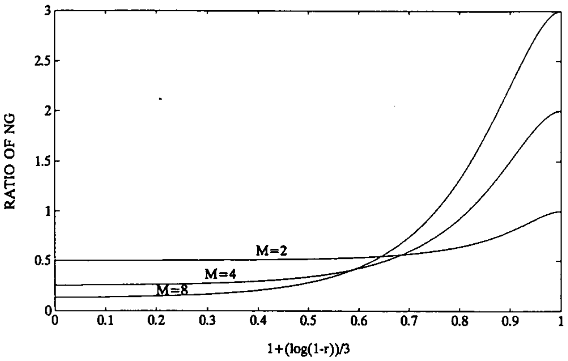


Fig. 1. Ratios of NG in the first-order pipelined filters with M=1 base.

Following the preceding development, the general form of the NG for $M \geq 2$ is inductively found to be

$$NG(M) = b^2c^2\frac{\log_2 M - (\log_2 M - 1)a^{2M}}{(1-a^2)(1-a^{2M})}. \tag{27}$$

The NG goes to infinity as $|a|$ goes to one, but a larger M slows the approach. For ease of comparison, the noise gains normalized to the conventional filter ($M = 1$) are plotted in Fig. 1. Note that throughout this paper, the abscissa of noise plots is $1 + (\log_{10}(1 - r))/3$, where r is the pole radius of a basis filter. Thus, the abscissa of one at the far right corresponds to $r = 0$, the abscissa of zero at the far left corresponds to $r = 0.999$, and the scale is an offset logarithmic. When the pole goes to zero, the NG ratio becomes

$$\lim_{a\to 0}\frac{NG(M)}{NG(M=1)} = \log_2 M. \tag{28}$$

When the pole goes to ±1, the NG ratio becomes

$$\lim_{a\to \pm 1}\frac{NG(M)}{NG(M=1)} = \frac{1}{M}. \tag{29}$$

It is thus clear that the $M = 2$ SLA first-order filter has lower RON than the conventional filter ($M = 1$) at all pole location. But for $M = 4, 8, \cdots$, it depends on the position of the pole. That is, for a pole near the origin the conventional filter is better than the pipelined one except $M = 2$ case, but the pipelined filters always have better RON characteristics for pole location near $z = \pm 1$.

Let

$$\sigma^2_{RON}(M=1) = (\frac{\Delta^2 2^{-2B_1}}{12})NG \tag{30}$$

where B_1 represents the number of bits in internal state registers of $M = 1$ case. When the pole is near the origin, then

$$\begin{aligned}\sigma^2_{RON}(M) &= \log_2 M \cdot \sigma^2_{RON}(M=1),\\ &= (\frac{\Delta^2 2^{-2(B_1-\frac{1}{2}\log_2(\log_2 M))}}{12})NG. \end{aligned}\tag{31}$$

Therefore, for a pole near the origin, the use of an SLA filter effectively costs an extra $\frac{1}{2}\log_2(\log_2 M)$ bits of internal registers to achieve the same RON as $M = 1$. Note that this is not a very large penalty, as M=16 costs one extra bit, M=65 536 costs two extra bits, etc. However near $z = \pm 1$

$$\begin{aligned}\sigma^2_{RON}(M) &= \frac{1}{M}\cdot\sigma^2_{RON}(M=1),\\ &= (\frac{\Delta^2 2^{-2(B_1+\frac{1}{2}\log_2 M)}}{12})NG. \end{aligned}\tag{32}$$

Therefore using an SLA filter near $z = \pm 1$ allows the use of $\frac{1}{2}\log_2 M$ less bits of internal registers to achieve the same RON as the $M = 1$ conventional filter. For example, using $M = 4$ saves one bit, using $M = 16$ saves two bits, etc. So it is important to note that for reasonable values of M, the performance of these first-order DF SLA filters is not much different from the conventional $M = 1$ filter. Where the conventional filter works well, the SLA filter requires around one bit extra. Where the conventional filter doesn't work too well (near $z = \pm 1$), the SLA filter requires one or two less bits.

B. Second-Order Pipelined Filters

In this subsection we find the RON of second-order SLA filters for the following two structures; the minimum roundoff noise (MRON) and true direct form (DF). If we start with a conventional $M = 1$ DF filter and apply the SVD form of SLA without any similarity transformation, the resultant SLA filters become apparent direct form (ADF) whose system matrix is no longer in companion form. For a second-order filter (or cascade or parallel connection), the ADF has more multiplications than the DF, and its NG is always worse than that of the MRON. Therefore, in this paper we do not consider the ADF structure, which has no area of application.

To have overall and explicit **K** and **W** matrices for the second-order case, we should solve the set of linear equations, Liapunov equations. To reduce the complexity of solving Liapunov equations and derive the explicit formulae for **K** and **W** of overall structure, we employ the SKWD introduced in Section IV-A. Under the assumption of system stability, i.e., eigenvalues of **A** being inside the unit circle, this analysis can be extended to the higher-order structures. Note that summing nodes and signal paths are now vector sequences. Therefore,

$$NG(M) = \sum_{n=1}^{N}\sum_{i=1}^{\log_2 M}(\sum_{j=1}^{p}\mathbf{W}_{ij}(n,n))\mathbf{K}_i(n,n)\mid_M . \tag{33}$$

Notice that using the SKWD technique, the dimensions of the **K** and **W** matrices are the same as that of the system matrix **A**, e.g., 2 × 2 matrices for second-order systems.

The RON of the second-order structures will be analyzed as a function of the radius and angle of the complex conjugate pole pair. Real poles can be implemented as a parallel or cascade connection of the first-order filters, which were analyzed previously.

Minimum Roundoff Noise Structure (Sectional Optimal) In this subsection we start with the second-order sectional optimal

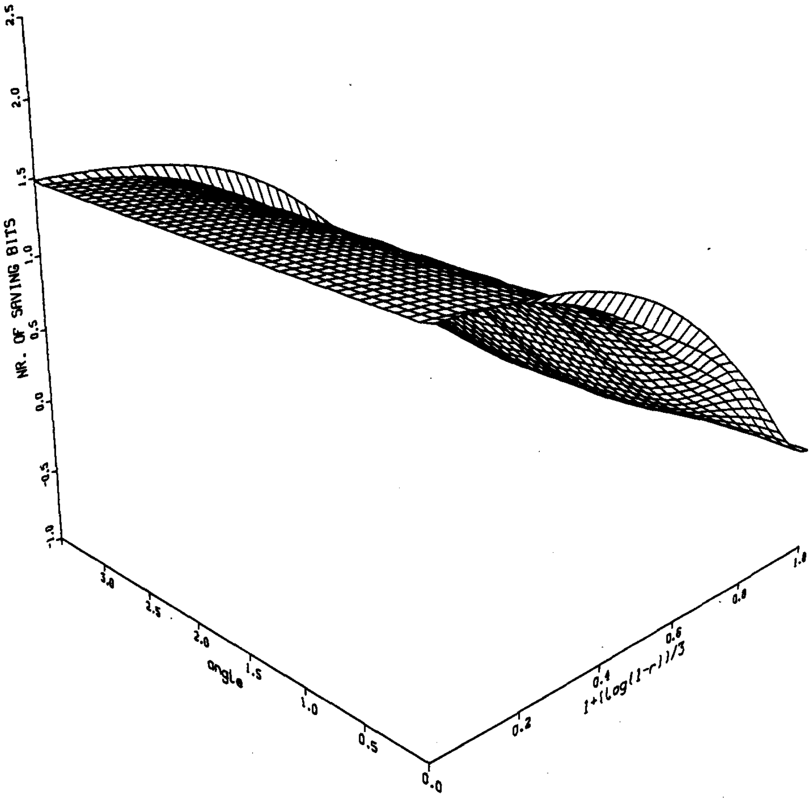


Fig. 2. Number of saving bits in the pipelined MRON filters ($M = 8$) to match the conventional MRON filter's RON performance.

MRON structure [11]-[13], apply the SVD SLA synthesis technique, and then analyze the resultant RON. We note that the explicit synthesis equations of Bomar [14] can't be used in our case (due to trivial zeros), so modified equations are

$$a_{11} = a_{22} = r\cos\theta, \tag{34}$$

$$b_1 = c_2 = 0, \tag{35}$$

$$a_{12} = \{\frac{1 + r^2(1 - 3\cos^2\theta) + r^4\cos^2\theta}{1 + r^2}\}^{\frac{1}{2}}, \tag{36}$$

$$a_{21} = \frac{-r^2\sin^2\theta}{a_{12}}, \tag{37}$$

$$b_2 = \left\{1 - \frac{a_{21}^2(1 - a_{12}a_{21} + a_{11}a_{22}) + a_{22}^2(1 + a_{12}a_{21} - a_{11}a_{22})}{1 - a_{12}a_{21} - a_{11}a_{22}}\right\}^{\frac{1}{2}}, \tag{38}$$

$$c_1 = \frac{1}{b_2 a_{12}} \tag{39}$$

where a_{12} and b_2 are real under stability.

After calculating **A**, **b**, **c** and d, the SVD SLA synthesis technique is applied. We note that the resultant SLA structure is not canonic (like all SLA structures), and thus is not MRON structure in the sense of the original definition. Fig. 2 plots the number of bits saved by the pipelined MRON filters ($M = 8$) to match the conventional MRON filter's RON performance. The RON of the MRON structure is symmetric about $\theta = \pi/2$. It is worth noting that for conventional filters ($M = 1$), the MRON filter actually has higher NG than the DF filter, except near $z = \pm 1$. This result is due to the fact that the MRON filter is optimized under the assumption that all potential parameters (two RON sources) in the structure generate RON, while our analysis of the DF structure only includes RON at nontrivial multipliers (one RON source).

Noticeable characteristics are as follows:

1) As r goes to zero, increasing $M > 2$ causes an increase in NG. The number of bits lost, i.e., the number of extra bits required to equal the NG performance with $M = 1$, is given by $\frac{1}{2}[\log_2(\log_2 M) - 1]$.
2) As r goes to one, increasing M decreases the NG. Each doubling of M saves one-half bit, and is independent on θ.

Direct Form Structure. We now consider a true DF SLA implementation [4], where each section (including nonrecursive) has a companion form of state update matrices $\mathbf{A}, \mathbf{A}^2, \mathbf{A}^4$ and $\mathbf{A}^8$, as shown in Fig. 3 for $M = 8$. The companion form of the matrices is important because the sparsity translates directly to a much smaller number of multiplications. And we will show that in many applications the numerical performance is adequate, even better than more complex structures.

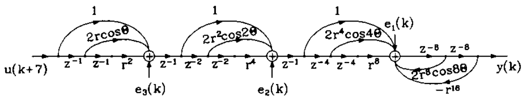


Fig. 3. Second-order pipelined DF filter with 8 levels of look-ahead.

The SKWD method of analysis is not convenient for this DF structure, so we demonstrate that the traditional SVD-based approach can be easily extended to include intermediate calculations, a method we call the noncanonic state variable description (NCSVD). For SLA structures, the intermediate calculations in the nonrecursive sections of the filter can include pipeline delays, and in practical implementations do so, as shown in Fig. 3 for $M = 8$. These pipeline delays, and the pure delays inside each nonrecursive section are now explicitly included in the state-space model of the filter. The well known SVD techniques are then used to calculate $\mathbf{K}$ and $\mathbf{W}$, which are of large dimension because these noncanonic state variables have been included in the state space. To calculate the NG, we then only consider those states which truly have RON sources, i.e., that are not pure delays. For the $M = 8$ true DF SLA structure, the state update and output equations are

$$\begin{aligned} x_1(k+1) &= x_2(k), \\ x_2(k+1) &= x_3(k), \\ &\cdots \\ x_{16}(k+1) &= -r^{16}x_1(k) + 2r^8\cos 8\theta x_9(k) + r^8 x_{17}(k) \\ &\quad + 2r^4\cos 4\theta x_{21}(k) + x_{25}(k), \\ &\cdots \qquad (40) \\ x_{25}(k+1) &= r^4 x_{26}(k) + 2r^2\cos 2\theta x_{28}(k) + x_{30}(k), \\ &\cdots \\ x_{30}(k+1) &= r^2 x_{31}(k) + 2r\cos\theta x_{32}(k) + u(k), \\ x_{31}(k+1) &= x_{32}(k), \\ x_{32}(k+1) &= u(k), \\ y(k-16) &= x_1(k) \qquad (41) \end{aligned}$$

where $\mathbf{A}$ is 32×32, $\mathbf{b}$ is 32×1, $\mathbf{c}$ is 1×32, and d is 1×1. Selecting only noise generating state variables, the NG can be calculated as

$$\begin{aligned} NG(M=8) = \mathbf{W}(16,16)\mathbf{K}(16,16) + \mathbf{W}(25,25)\mathbf{K}(25,25) \\ + \mathbf{W}(30,30)\mathbf{K}(30,30). \qquad (42) \end{aligned}$$

General Formula for Noise Gain. When considering the DF SLA filters, there are many common terms between filters of different M, so we have derived the following relationships:

$$\begin{aligned} \mathbf{K}(2,2)\,|_{M=1} &= \mathbf{K}(4,4)\,|_{M=2} = \mathbf{K}(8,8)\,|_{M=4} \\ &= \mathbf{K}(16,16)\,|_{M=8} \qquad (43) \\ \mathbf{K}(13,13)\,|_{M=4} &= \mathbf{K}(30,30)\,|_{M=8} \qquad (44) \\ \mathbf{W}(2,2)\,|_{M=1} &= \mathbf{K}(2,2)\,|_{M=1} \qquad (45) \\ \mathbf{W}(4,4)\,|_{M=2} &= \mathbf{W}(13,13)\,|_{M=4} \\ &= \mathbf{W}(30,30)\,|_{M=8} \qquad (46) \\ \mathbf{W}(8,8)\,|_{M=4} &= \mathbf{W}(25,25)\,|_{M=8} \qquad (47) \end{aligned}$$

The terms $\mathbf{W}(2,2)\,|_{M=1}, \mathbf{W}(4,4)\,|_{M=2}, \mathbf{W}(8,8)\,|_{M=4}$, and $\mathbf{W}(16,16)\,|_{M=8}$ can all be calculated by

$$P(M) = \frac{1+r^{2M}}{1+r^{2M}-4r^{2M}\cos^2 M\theta - r^{4M} + 4r^{4M}\cos^2 M\theta - r^{6M}}. \qquad (48)$$

Therefore, the NG of the second-order pipelined DF filters can be computed by

$$\begin{aligned} NG(M) = P(1, P(M) + K_{\log_2 M}P(M/2) + K_{\log_2 M-1}P(M/4) \\ + \cdots + K_2 P(2), \qquad \text{for } M \geq 2. \qquad (49) \end{aligned}$$

Here, states are numbered from the output to the input, and so K_i denotes the power gain from the input to the ith node from the output. That is,

$$\begin{aligned} K_1\,|_{M=1} &= \mathbf{K}(2,2)\,|_{M=1}, \qquad (50) \\ K_1\,|_{M=2} &= \mathbf{K}(4,4)\,|_{M=2}, \qquad (51) \\ K_2 &= \mathbf{K}(13,13)\,|_{M=4}, \qquad (52) \\ K_3 &= \mathbf{K}(25,25)\,|_{M=8}\,. \qquad (53) \end{aligned}$$

The results above have allowed us to derive the analytic closed form for the NG near the unit circle [15]. In the limit as r approaches one, the ratio of NG as a function of M to the NG of the conventional DF filter is

$$\frac{NG(M)}{NG(1)} = \frac{\sin^2\theta}{M\sin^2 M\theta}. \qquad (54)$$

Therefore, the number of bits saved is

$$-\frac{1}{2}\log_2 \frac{\sin^2\theta}{M\sin^2 M\theta}. \qquad (55)$$

Notice that because the number of savings bits is negative, extra bits are actually required in the SLA structure to achieve performance equal to the conventional. Numerical limitations in the calculations has rendered the true negative infinities at about -50 bits. These negative infinities occur at $\theta = i\pi/M$, where $i = \pm 1 \cdots \pm (M-1)$, so the number of peaks in NG near the unit circle is $(M-1)$. At $\theta = 0$ and π, ratio of NG is $8^{-\log_2 M}$, i.e. M^{-3}. The huge peaks in NG can be understood by noting that the poles added by the SLA algorithm have clustered with the original poles of the filter, and the poor numerical performance of DF structures with clustered and high Q poles is well known. The number of bits saved over the conventional DF structure ($M = 1$) as a function of radius and angle is shown in Fig. 4 for $M = 8$. The RON characteristic is again symmetric about $\theta = \pi/2$.

Notable characteristics of the DF pipelined filters are as follows:

1) As r goes to zero, the NG worsens with increasing $M > 2$. The ratio of noise gains is $\log_2 M$, so the number of bits lost is $\frac{1}{2}\log_2(\log_2 M)$.

2) As r goes to one, the very poor behavior around $\theta = i\pi/M$, which was previously described, is seen in context.

Values of M (among 1, 2, 4, or 8) that minimize RON of the DF SLA filter are plotted as a function of r and θ in Fig. 6 (a).

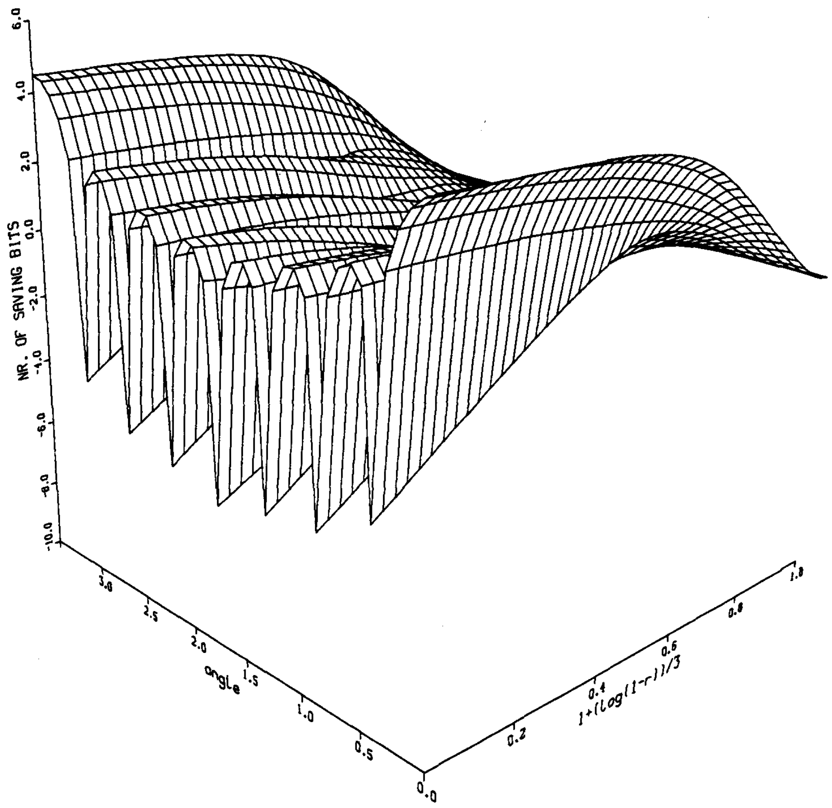


Fig. 4. Number of saving bits in the pipelined DF filters (M = 8) to match the conventional DF filter's RON performance.

Results for RON of Second-Order Structures. In this subsection we compare the two second-order SLA structures described previously; the MRON and the DF. The number of multiplications for each structure is

$$\text{MRON Structure}: N^2(1+\log_2 M), \tag{56}$$

$$\text{DF Structure}: N(1+\log_2 M). \tag{57}$$

Comparing the NG of DF and MRON structures, the region where the MRON structure has less NG is increasing as M increases.

In conclusion, the NG of second-order SLA structures is generally quite good when compared to the underlying basis filter, with the notable exception of the DF structure for high Q poles near $\theta = i\pi/M$. There is a slight increase in NG with increasing M for very low Q poles, but the cost for the realistic cases of $M \leq 16$ is less than one bit, and grows very slowly beyond that. The question of whether the computationally simpler DF SLA filter can be used retains the same character as that question for a non-SLA DF filter, except for high Q poles near $\theta = i\pi/M$. In the later case, either a more complex structure, such as MRON, can be used, or alternately a composite M (not a power of 2) can be chosen to achieve the required speedup without clustering poles (but with slightly higher complexity, depending on the composite factorization).

V. Statistical Coefficient Quantization Model

The effect of coefficient errors on the performance of digital filters was first noticed by Kaiser [16]. There also have been several attempts to observe CQN by a deterministic way, including [17] and [18]. However, in [17] the concept of scaling is not treated, and in [18] the assumptions about approximate deterministic output error variance due to coefficient quantization are not well founded, i.e., we cannot simply ignore the higher-order terms related to the quantized values of $\mathbf{A}$. So the output error variance due to coefficient quantization should be replaced by

$$E[\Delta y^2] \simeq tr[\mathbf{ZW}] + \Delta\mathbf{c}\mathbf{V}_{\tilde{x}\tilde{x}}\Delta\mathbf{c}^T + \Delta d^2 \tag{58}$$

where

$$\begin{aligned}\mathbf{Z} = \Delta\mathbf{A}\mathbf{V}_{\tilde{x}\tilde{x}}\Delta\mathbf{A}^T + \mathbf{A}(\mathbf{V}_{\tilde{x}\tilde{x}} - \mathbf{V}_{x\tilde{x}})\Delta\mathbf{A}^T \\ + \Delta\mathbf{A}(\mathbf{V}_{\tilde{x}\tilde{x}} - \mathbf{V}_{\tilde{x}x})\mathbf{A}^T + \Delta\mathbf{b}\Delta\mathbf{b}^T.\end{aligned} \tag{59}$$

Here, $\mathbf{Z}$ is closely related to the filter structure from the input to the state variables, similar to the role of $\mathbf{K}$ in RON analysis. In contrast to RON, CQN has deterministic property because of preset coefficients of digital filter. However we cannot measure the output error, caused only by the coefficient quantization, by deterministic way in real situations, because numerical error comes from the mixture of RON and CQN sources at the output node. This is one of the reasons we need statistical tools to analyze CQN. Another reason is the necessity of unifying analytical and synthetic tools for both errors. Methods using statistical analysis for coefficient quantization include [18]-[22]. Fettweis noticed there exists similarity of generations between RON and CQN [19]. According to Jackson [13], RON and CQN are found to form the bound of each other. Recently two noticeable researches about statistical CQN (SCQN) have been done in [20] and [21], and they reach the same conclusion even though they employed different approaches.

In this paper, we provide another viewpoint for the analysis of CQN. That is, to calculate the reflected CQN on the output node, output noise power due to coefficient quantization, σ^2_{CQN}, is considered instead of coefficient sensitivity. In σ^2_{CQN} analysis with supporting register model, the amount of output error can be directly found. There also has been an attempt to find the linear dependence between the weighted sensitivity and the weighted noise measure [22].

In the following sections, we apply the state-space linear algebraic approach for RON analysis of [11] to the coefficient quantization problem. Statistical assumptions including hardware model to unify the CQN and RON analysis are also presented. The SLA filters with the power-of-two decomposition in [4] are used as an example structure to demonstrate a unified approach. These results contradict previous claims [6] about the numerical performance of the SLA structure, where register scaling [8], [23] was not considered. The total output error variance after applying a proper scaling is derived by combining RON and CQN [24].

In this section we state sufficient conditions to show that SCQN source scan be interpreted just like single-length accumulation RON source. This result was also achieved in [21] by assuming certain approximations, and as a different form in [20]. In [21] by using the state error vector $e_x(k) = \tilde{x}(k) - x(k)$ and the output error $e_y(k) = \tilde{y}(k) - y(k)$, the statistical sensitivity S, output error variance normalized by the variance σ^2 of coefficient variations, is described by

$$S = \sigma^2_{CQN}/\sigma^2 = tr[\mathbf{K}]tr[\mathbf{W}] + tr[\mathbf{W}] + tr[\mathbf{K}] \qquad (60)$$

where output branch is omitted as usual. Here, it is assumed all the elements of the system matrices are varied and

$$\mathbf{K} = \tilde{\mathbf{K}} \qquad (61)$$

where $\mathbf{K}$ is the state covariance matrix of the ideal filter and $\tilde{\mathbf{K}}$ is that of the quantized filter. For (61), the condition $\sigma^2 \ll \sigma^2_u$, where σ^2_u is an upper bound of variance that satisfies mean square asymptotical stability of the virtual system, should be satisfied. Another underlying assumption in (60) is the same q for state vector and system matrices $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}$ and d, that means there are the same length of bits below decimal point in every register. Δ, in defining quantization step size q, is a constant chosen to fit the dynamic range of signal. However, Δ for state vector and system matrices does not usually have the same values (Δ for state vector is 1 by using the l_2-scaling with $\delta = 1$). That is, Δ and B combination of state vector and system matrices should be different to guarantee the same q. Therefore, the number of bits above the decimal point for system matrices should be determined by the maximum dynamic range of system matrices, assuming all system matrices use registers of the same length. There can also be a tradeoff between the usage of registers of different length for each system matrix and extra circuitry to support that.

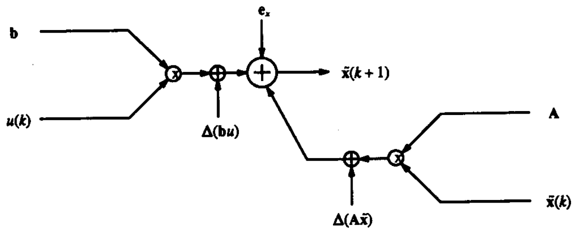


Fig. 5. Noise source model of statistical coefficient quantization.

Considering the above register model, a general equation of output noise variance due to the coefficient quantization of scaled filter is

$$\sigma^2_{CQN} = \frac{q^2}{12}\{tr[\mathbf{K}_s]tr[\mathbf{W}_s] + tr[\mathbf{W}_s] + tr[\mathbf{K}_s]\}. \qquad (62)$$

In contrast to the previous researches of coefficient sensitivity which is normalized by σ^2, the above equation considers a practical register model by Δ and B, and this phenomena will be detailed for the DF and the MRON structures. To make the δ-dependence of our register model clear, we recall a larger δ brings a larger RON (CQN also) and smaller overflow. In addition, a larger δ causes shorter coefficient registers for $\mathbf{b}$ and longer registers for $\mathbf{c}$ in our model. From now on, we assume $\delta = 1$ without any loss of generality. That is, the decimal point is assumed to be the right of the most significant bit (MSB) and the other bits are reserved for the fraction. Negative numbers are represented in two's complement form. It is similar to $Q15$ format in TMS320C25 [25].

Theorem 1: Noise source of coefficient quantization can be interpreted as the roundoff noise source of $q^2/12$ variance in the single-length accumulation register model under the following three assumptions.

1) Coefficients are rounded to the least significant bit (LSB) = q, the same as state variables, but noting that many structures require additional bits to represent coefficients whose magnitude exceeds one.
2) Coefficient quantized errors, $\Delta\mathbf{b}_s$ and Δd_s, are uncorrelated with standard white Gaussian input $u(k)$. Coefficient quantized errors, $\Delta\mathbf{A}_s$ and $\Delta\mathbf{c}_s$, are uncorrelated with quantized state vector $\tilde{\mathbf{x}}(k)$.
3) $\mathbf{K}_s = \tilde{\mathbf{K}}_s$, the state covariance matrix of the quantized system.

Proof: Noise source model of statistical coefficient quantization is shown in Fig. 5 for state equation. To apply the concept of single-length accumulation register RON model, we transfer real CQN sources for $\mathbf{b}_s$ and $\mathbf{A}_s$ after each multiplier. Let's label the two transferred SCQN sources as $\Delta(\mathbf{b}_s u)$ and $\Delta(\mathbf{A}_s\tilde{\mathbf{x}})$ respectively. For the input section, after the scaling and under the steady-state (i.e. time index $k \rightarrow \infty$)

$$(\mathbf{b}_s + \Delta\mathbf{b}_s)u(k) = \mathbf{b}_s u(k) + \Delta(\mathbf{b}_s u(k)). \qquad (63)$$

That is,

$$\Delta(\mathbf{b}_s u(k)) = \Delta\mathbf{b}_s \cdot u(k). \tag{64}$$

Therefore, the variance of $\Delta(\mathbf{b}_s u(k))$ becomes

$$\begin{aligned}
&\sigma^2_{\Delta(\mathbf{b}_s u(k))}\\
&= E[\Delta\mathbf{b}_s\Delta\mathbf{b}_s^T \cdot u^2\mathbf{I}] - E[\Delta\mathbf{b}_s \cdot u]E[\Delta\mathbf{b}_s \cdot u]^T,\\
&\simeq E[\Delta\mathbf{b}_s\Delta\mathbf{b}_s^T]E[u^2\mathbf{I}] - E[\Delta\mathbf{b}_s]E[u]E[u]^T E[\Delta\mathbf{b}_s]^T,\\
&= \frac{q^2}{12}\mathbf{I}.
\end{aligned} \tag{65}$$

Here, we used the assumption of uncorrelation between $\Delta\mathbf{b}_s$ and u. For the state recurrence section,

$$(\mathbf{A}_s + \Delta\mathbf{A}_s)\tilde{\mathbf{x}}(k) = \mathbf{A}_s\tilde{\mathbf{x}}(k) + \Delta(\mathbf{A}_s\tilde{\mathbf{x}}(k)). \tag{66}$$

Then

$$\Delta(\mathbf{A}_s\tilde{\mathbf{x}}(k)) = \Delta\mathbf{A}_s \cdot \tilde{\mathbf{x}}(k). \tag{67}$$

Therefore, the variance of $\Delta(\mathbf{A}_s\tilde{\mathbf{x}}(k))$ becomes

$$\begin{aligned}
&\sigma^2_{\Delta(\mathbf{A}_s\tilde{\mathbf{x}}(k))}\\
&= E[\Delta\mathbf{A}_s\Delta\mathbf{A}_s^T \cdot \tilde{\mathbf{x}}\tilde{\mathbf{x}}^T] - E[\Delta\mathbf{A}_s \cdot \tilde{\mathbf{x}}]E[\Delta\mathbf{A}_s \cdot \tilde{\mathbf{x}}]^T,\\
&\simeq E[\Delta\mathbf{A}_s\Delta\mathbf{A}_s^T]E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T]\\
&- E[\Delta\mathbf{A}_s]E[\tilde{\mathbf{x}}]E[\tilde{\mathbf{x}}]^T E[\Delta\mathbf{A}_s]^T,\\
&= \frac{q^2}{12}E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T],\\
&\simeq \frac{q^2}{12}E[\mathbf{x}\mathbf{x}^T],\\
&= \frac{q^2}{12}\mathbf{I}.
\end{aligned} \tag{68}$$

Here, we used the assumption that $\Delta\mathbf{A}_s$ and $\tilde{\mathbf{x}}$ are uncorrelated with each other and $\mathbf{K}_s$ is equal to $\tilde{\mathbf{K}}_s$, the state covariance matrix of quantized filter. Above derivations are similarly applied to the output equation. Therefore, the variances of all noise sources are proved to be $q^2/12$ in our SCQN model. □

The above three assumptions assure the variance of each noise source of single-length accumulation register is equal to $q^2/12$, and are the same assumptions as those in [21] except the first one. Actually, the first assumption has not been noticed before, but it was represented as a normalized output error variance instead. For the compactness of the theorem, we assumed state variables and coefficients are all rounded to the $LSB = q$ together. However, even though state variables and each coefficient have different quantization step sizes, the actual value of σ^2_{CQN} can be directly calculated in our model under the validity of assumptions 2 and 3. In the previous researches of CQN sensitivity, it was difficult to calculate the value of σ^2_{CQN}, even in the case of same q, unless we know the variance of coefficient variations, which cannot easily be obtained from the specifications of a digital filter.

The main point of the above analysis is the unification of analysis methods for RON and CQN. Moreover, the minimization of RON also ensures the minimization of CQN, that is, the minimization of numerical noises can be achieved simultaneously under the assumption of full system matrix. The above analyses are based on the condition of mean square asymptotical stability of the quantized filter which guarantees the convergence of state covariance matrix and state error covariance matrix of quantized filter [21].

The accuracy of (60) was considered in detail in [21]. For the DF filter, we consider the sparsity of $\mathbf{A}_s$ matrix also. The largest values of coefficients come from $\mathbf{c}$ if $\delta \geq 1$,which is a usual case. For the our all-pole second-order filters, it is

$$\sqrt{\frac{1 + r^2}{1 + r^2 - 4r^2\cos^2\theta - r^4 + 4r^4\cos^2\theta - r^6}} \tag{69}$$

which is $\sqrt{K_{11}}$. It is the same for the DF and the MRON filters and also the same for all levels of loop pipelined filters. That is, (69) can also be the extra bits for the pipelined filters. It is interesting to note that the maximum value of $\mathbf{A}_s$ in the DF filter is always larger than 1 (0 for extra bits), but always smaller than 1 in the MRON filter. We observe the dynamic range of $\mathbf{b}_s$ over the z-plane in the DF filter has significantly larger value than in the MRON filter, and the maximum values of $\mathbf{b}_s$ of both filters are less than 1.

VI. Coefficient Quantization Noise Of SLA Filters

A. First-Order Pipelined Filter.

By using the SKWD

$$\begin{aligned}
\sigma^2_{CQN}(M = 1) &= \frac{q^2}{12}(K_sW_s + W_s + K_s),\\
&= \frac{q^2}{12}(2NG + 1),\\
&= \frac{q^2}{12}\{\frac{2b^2c^2}{(1 - a^2)^2} + 1\},
\end{aligned} \tag{70}$$

$$\begin{aligned}
\sigma^2_{CQN}(M) = \frac{q^2}{12}\{&\frac{2(\log_2 M - 1)b^2c^2(1 - a^{2M}) + 3b^2c^2}{(1 - a^2)(1 - a^{2M})}\\
&+ 1\}, \qquad \text{where } M \geq 2.
\end{aligned} \tag{71}$$

The σ^2_{CQN} goes to infinity as $|a|$ goes to one, but a larger M slows the approach. When the pole goes to the origin, the σ^2_{CQN} ratio becomes

$$\lim_{|a|\to 0}\frac{\sigma^2_{CQN}(M)}{\sigma^2_{CQN}(M = 1)} = \frac{2}{3}(1 + \log_2 M) \tag{72}$$

where $M \geq 2$. When the pole goes to ±1, the σ^2_{CQN} ratio becomes

$$\lim_{|a|\to 1}\frac{\sigma^2_{CQN}(M)}{\sigma^2_{CQN}(M = 1)} = \frac{3}{2M} \tag{73}$$

where $M \geq 2$. The effect of M here is similar to the RON case.

The extra bits for coefficient registers to guarantee the same q (here, $\delta = 1$) come from the maximum dynamic range of the scaled coefficients, output branch usually, that is

$$\Delta_{coff}\mid_{max} = \sqrt{\frac{1}{1-r^2}} \tag{74}$$

which is the same for all levels of loop pipelined first-order filters. This can also be the number of bits we should move the decimal point to the right side in coefficient registers when we use different q, i.e., different Δ_{coff} and same B.

B. Second-Order Pipelined Filters

Minimum Roundoff Noise Structure. By using the same analysis method in the first-order case and considering sparsity in $\mathbf{b}$ and $\mathbf{c}$ of our all-pole filter

$$\begin{aligned}\sigma^2_{CQN}(M=1) = \frac{q^2}{12}\{&2NG\mid_{M=1}\\ &+ \mathbf{W}_{11}(2,2)\mathbf{K}_1(2,2)+1\},\end{aligned} \tag{75}$$

$$\begin{aligned}\sigma^2_{CQN}(M=2) = \frac{q^2}{12}\{&3NG\mid_{M=2}\\ &+ \mathbf{W}_{11}(2,2)\mathbf{K}_1(2,2)+1\},\end{aligned} \tag{76}$$

$$\begin{aligned}\sigma^2_{CQN}(M=4) = \frac{q^2}{12}\{&NG_2\mid_{M=4} + \mathbf{W}_2(2,2)\mathbf{K}_2(2,2)\\ &+ 5NG_1\mid_{M=4} +1\},\end{aligned} \tag{77}$$

$$\begin{aligned}\sigma^2_{CQN}(M=8) = \frac{q^2}{12}\{&NG_3\mid_{M=8} + \mathbf{W}_3(2,2)\mathbf{K}_3(2,2)\\ &+ 3NG_2\mid_{M=8} + 5NG_1\mid_{M=8} +1\}\end{aligned} \tag{78}$$

where $\mathbf{W}_2 = \mathbf{W}_{21} + \mathbf{W}_{22}$ and $\mathbf{W}_3 = \mathbf{W}_{31} + \mathbf{W}_{32} + \mathbf{W}_{33} + \mathbf{W}_{34}$. The σ^2_{CQN} ratio for the MRON pipelined filters is

$$\lim_{r\to 0}\frac{\sigma^2_{CQN}(M)}{\sigma^2_{CQN}(M=1)}\mid_{MRON} = \frac{1}{6}\{4+3(\log_2 M - 1)\} \tag{79}$$

where $M \geq 2$.

The absolute coefficient noise gain (CNG) in the second-order conventional DF and MRON filters and the number of saving bits to match the conventional filter's performance for the pipelined MRON filters take the form of plots quite similar to the RON case. The ratio near the unit circle is $2^{-M/4}$, except at $\theta = \pi/2\mid_{M=2}$. Combining with RON results, we observe a large M is recommended for the pipelined MRON filters regardless of θ to implement high Q filter.

Direct Form Structure. After applying the NCSVD, we consider the sparsity of the system matrices of the pipelined DF filters. Then,

$$\begin{aligned}\sigma^2_{CQN}(M=1) &= \frac{q^2}{12}\{\sum_{n=1}^{2}(m_{\mathbf{A}_s}(n)\\ &\quad + m_{\mathbf{b}_s^T}(n))\mathbf{W}_s(n,n) + m_{\mathbf{c}_s}(n)\},\\ &= \frac{q^2}{12}\{\mathbf{W}_s(1,1) + 2\mathbf{W}_s(2,2)+1\},\\ &= \frac{q^2}{12}\{(r^4+2)P(1)^2 + P(1) + 1\},\end{aligned} \tag{80}$$

$$\begin{aligned}\sigma^2_{CQN}(M=2) = \frac{q^2}{12}\{&\mathbf{W}_s(1,1) + \mathbf{W}_s(3,3)\\ &+ \mathbf{W}_s(4,4) + \mathbf{W}_s(5,5)\\ &+ \mathbf{W}_s(6,6) + 1\},\\ = \frac{q^2}{12}\{&(r^8+2)P(1)P(2)\\ &+ 2r^2(r^2+2\cos^2\theta)P(2)\\ &+ P(1) + 1\},\end{aligned} \tag{81}$$

$$\begin{aligned}\sigma^2_{CQN}(M=4) = \frac{q^2}{12}\{&\mathbf{W}_s(1,1) + \mathbf{W}_s(5,5) + \mathbf{W}_s(9,9)\\ &+ \mathbf{W}_s(11,11)\\ &+ \mathbf{W}_s(13,13) + \mathbf{W}_s(14,14)\\ &+ \mathbf{W}_s(15,15) + 1\},\\ = \frac{q^2}{12}\{&(r^{16}+1)P(1)P(4)\\ &+ 2r^4(r^4+2\cos^2 2\theta)\mathbf{K}(13,13)P(4)\\ &+ \mathbf{K}(13,13)P(2) + 2r^2(r^2+2\cos^2\theta)P(2)\\ &+ P(1) + 1\},\end{aligned} \tag{82}$$

$$\begin{aligned}\sigma^2_{CQN}(M=8) = \frac{q^2}{12}\{&\mathbf{W}_s(1,1) + \mathbf{W}_s(9,9)\\ &+ \mathbf{W}_s(17,17) + \mathbf{W}_s(21,21)\\ &+ \mathbf{W}_s(26,26) + \mathbf{W}_s(28,28)\\ &+ \mathbf{W}_s(30,30) + \mathbf{W}_s(31,31)\\ &+ \mathbf{W}_s(32,32) + 1\},\\ = \frac{q^2}{12}\{&(r^{32}+1)P(1)P(8)\\ &+ 2r^8(r^8+2\cos^2 4\theta)\mathbf{K}(25,25)P(8)\\ &+ 2r^4(r^4+2\cos^2 2\theta)\mathbf{K}(13,13)P(4)\\ &+ \mathbf{K}(13,13)P(2)\\ &+ 2r^2(r^2+2\cos^2\theta)P(2) + P(1) + 1\}\end{aligned} \tag{83}$$

where $m(i)$ is the number of the varied elements in the ith column of corresponding matrix, and $P(M)$, $K(13,13)$, and $K(25,25)$ are defined in Section IV-B.

Therefore, a general formula of output CQN variance ($M \geq 4$) is

$$\begin{aligned}\sigma^2_{CQN}(M) = \frac{q^2}{12}\{&(r^{4M}+1)P(1)P(M)\\ &+ 2r^M(r^M + 2\cos^2\frac{M}{2}\theta)K_{\log_2 M}P(M)\\ &+ 2r^{M/2}(r^{M/2} + 2\cos^2\frac{M}{4}\theta)K_{\log_2 M-1}P(M/2)\\ &+ \cdots\cdots\cdots\\ &+ 2r^4(r^4+2\cos^2 2\theta)K_2P(4) + K_2P(2)\\ &+ 2r^2(r^2+2\cos^2\theta)P(2) + P(1) + 1\}.\end{aligned} \tag{84}$$

Then the σ^2_{CQN} ratio becomes

$$\lim_{r\to 0}\frac{\sigma^2_{CQN}(M)}{\sigma^2_{CQN}(M=1)} = 1, \tag{85}$$

$$\lim_{r\to 1}\frac{\sigma^2_{CQN}(M=2)}{\sigma^2_{CQN}(M=1)} = \frac{\sin^2\theta}{2\sin^2 2\theta}, \tag{86}$$

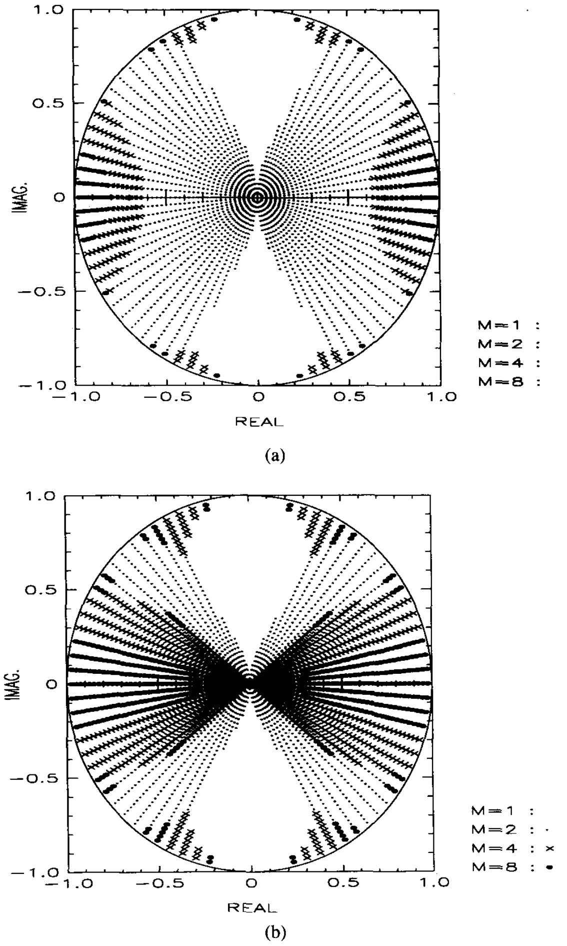


Fig. 6. Values of M to minimize quantization noise in the pipelined DF filters: (a) RON; (b) CQN.

$$\lim_{r\to 1}\frac{\sigma_{CQN}^2(M)}{\sigma_{CQN}^2(M=1)}=\frac{\sin^2\theta}{6(\log_2 M-1)\sin^2 M\theta},\quad \text{where } M\geq 4. \tag{87}$$

Here, CQN due to the feedforward part becomes zero near the origin, and that is why we have constant CQN ratio near the origin. As the poles approach the unit circle, however, unlike the MRON structure, the σ_{CQN}^2 ratios of the pipelined DF filters are seen to be highly dependent on the pole angle. In particular, if $\theta = \pm i\pi/M$ for $i \in \{1, 2, \cdots, (M-1)\}$, then the ratios become infinite as in the RON case. At $\theta = 0$ and π, ratio of coefficient noise gain (CNG) becomes $\{6M^2(\log_2 M - 1)\}^{-1}$, $M \geq 4$. For the pipelined DF filters, the number of saving bits to match the conventional DF filter's CQN performance also takes a form of plots similar to the RON case. Values of M at each r and θ to minimize CQN are shown in Fig. 6(b) for the DF. To design a high Q pipelined DF filter, a smaller M is better around the imaginary axis, but a larger M is recommended around the real axis. With M larger, the DF structure is getting better than the MRON structure in CQN, which is contrary to the RON case. That is, they have different improving velocity about RON and CQN. With M larger, the pipelined DF filter is better in decreasing CQN and the pipelined MRON filter is better in decreasing RON compared to each other. Again, the pipelined DF and MRON filters are the structures to be selected for the purpose of less CQN or less complexity, as in the conventional case.

TABLE I
$\sigma_{total}^2(M)/\sigma_{total}^2(M=1)$ OF THE FIRST AND SECOND-ORDER PIPELINED FILTERS

Pole Loc.	First - Order	Second - Order
r ⟶ 0	(5+3(log2M-1)) / 4	(4+log2M) / 5
r ⟶ 1	4 / 3M	$\sin^2\theta / 2\sin^2 2\theta$, M = 2 $3\sin^2\theta / 4M\sin^2 M\theta$, M ≥ 4

VII. Total Quantization Noise of SLA Filters

From the uncorrelation of noise sources or by the deterministic approach [18], we can verify

$$\sigma_{total}^2 = \sigma_{RON}^2 + \sigma_{CQN}^2. \tag{88}$$

In the pipelined filters,

1) First-Order Case

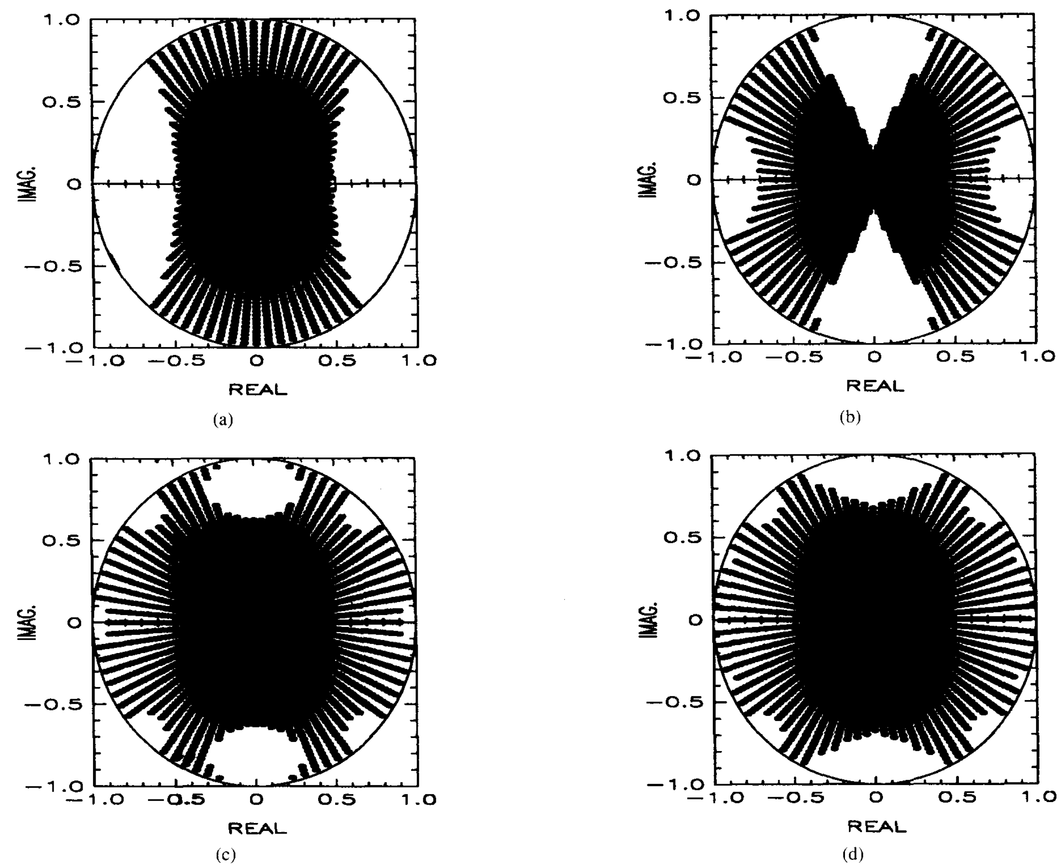
$$\sigma_{total}^2 = \frac{q^2}{12}\{\frac{\log_2 M-(\log_2 M-1)r^{2M}}{(1-r^2)(1-r^{2M})} + \frac{2(\log_2 M-1)(1-r^{2M})+3}{(1-r^2)(1-r^{2M})}+1\}\,|_{M\geq 2} \tag{89}$$

2) Second-Order DF Case

$$\begin{aligned}\sigma_{total}^2 = \frac{q^2}{12}\{&P(1)P(M)+K_{\log_2 M}P(M/2)\\ &+K_{\log_2 M-1}P(M/4)\\ &+\cdots\cdots+K_2P(2)\\ &+(r^{4M}+1)P(1)P(M)\\ &+2r^M(r^M+2\cos^2\frac{M}{2}\theta)K_{\log_2 M}P(M)\\ &+2r^{M/2}(r^{M/2}+2\cos^2\frac{M}{4}\theta)K_{\log_2 M-1}P(M/2)\\ &+\cdots\cdots\cdots\\ &+2r^4(r^4+2\cos^2 2\theta)K_2P(4)+K_2P(2)\\ &+2r^2(r^2+2\cos^2\theta)P(2)+P(1)+1\}\,|_{M\geq 4}.\end{aligned} \tag{90}$$

The ratios of each σ_{total}^2 with $M = 1$ base are in Table I.

Fig. 7 shows the regions of less total quantization noise (TQN) for the pipelined DF filters compared with the pipelined MRON filters. As a conclusion, the DF and the MRON structures should be considered together to implement the pipelined filter for the purpose of better numerical performance or less complexity.



(a) (b) (c) (d)

Fig. 7. Region of less TQN for the pipelined DF filters compared with the pipelined MRON filters: (a) M=1; (b) M=2; (c) M=4; (d) M=8.

VIII. Conclusions

The FWL effects of the first-order and the second-order DF and MRON-based SLA filters using fixed-point arithmetic have been investigated using the unified statistical methods presented in this paper. Two methods to calculate scaling parameters and NG ratios with respect to the $M = 1$ (non-SLA) were demonstrated. The first-order SLA forms increase NG by $\log_2 M$ near $z = 0$, and decrease NG by M near $z = \pm 1$. The second-order SLA MRON form has a well behaved NG, being very close to the first-order SLA. Near the origin the second-order DF SLA has constant NG, but depends strongly on pole angle near the unit circle, where NG ratios vary between infinite increase and decrease by M^3. The overall RON performance of the SLA pipelining is quite good, except for small regions of the z-plane near the unit circle, where the second-order DF SLA filters behave poorly. Except for these regions, the need to use the MRON structure instead of the DF structure is analogous to the same choice with non-SLA filters. Because the DF structure has fewer summing nodes than the MRON structure, the DF can have slightly smaller NG in certain regions of the z-plane. As M increases, this region of the z-plane atrophies. By applying the unified statistical tools developed in this paper, CQN is proved to be interpretable as RON under the single-length accumulation model and usually larger than RON in digital system. In our modeling with given specifications of a digital filter, the actual value of σ^2_{CQN} can be directly calculated, which is not possible in the previous researches of coefficient sensitivity. It can even be done in the case of the different quantization step sizes for state variables and each coefficient. With increasing M, the regions in the z-plane where the DF SLA filter has lower CQN than the MRON structure are actually increasing; however, for RON the analogous regions are decreasing. In conclusion, both the DF and the MRON basis filters should be considered when implementing high throughput digital filters. Contrary to widespread belief, the DF structure including the DF SLA, sometimes exhibits lower RON and CQN with lower implementation cost. The limit cycle behavior of the various SLA pipelined filters is analyzed in [26] and [27].

References

[1] A. L. Moyer, "An efficient parallel algorithm for digital IIR filters," in *Proc. IEEE Int. Conf. ASSP*, Apr. 1976, pp. 525-528.

[2] C. W. Barnes and S. Shinnaka, "Block shift invariance and block implementation of discrete-time filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 667-672, Aug. 1980.

[3] R. P. Brent and Z. B. Bing, "A stabilized parallel algorithm for direct-form recursive filters," *IEEE Trans. Comput.*, vol. 40, pp. 333-336, Mar. 1991.

[4] 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. 37, pp. 1099-1117, July 1989.

[5] J. G. Chung and K. K. Parhi, "Design of pipelined lattice IIR digital filters," in *Proc. Asilomar Conf.*, Nov. 1991, pp. 1021-1025.

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

[7] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*. Reading, MA: Addison-Wesley, 1987.

[8] W. G. Bliss and K. H. Chang, "The roundoff noise of pipelined scattered look-ahead IIR digital filters with decomposition," in *Proc. Asilomar Conf.*, Nov. 1991, pp. 1026-1030.

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

[10] C. D. Thompson, "Area-time complexity for VLSI," in *Proc. Caltech Conf. on VLSI*, Jan. 1979, pp. 495-508.

[11] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551-562, Sept. 1976.

[12] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273-281, Aug. 1977.

[13] L. B. Jackson, A. G. Lindgren, and Y. Kim, "Optimal synthesis of second-order state-space structures for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 149-153, Mar. 1979.

[14] B. W. Bomar, "New second-order state-space structures for realizing low roundoff noise digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*. vol. ASSP-33, pp. 106-110, Feb. 1985.

[15] K. H. Chang, "Finite word-length effects of pipelined recursive digital filters," *Ph.D. Dissertation*, Dept. Elec. Eng., Texas A&M Univ., Aug. 1992.

[16] J. F. Kaiser, "Some practical considerations in the realization of linear digital filters," in *Proc. 3rd Allerton Conf. on Circuit and Syst. Theory*, pp. 621-633, Oct. 1965.

[17] J. B. Knowles and E. M. Olcayto, "Coefficient accuracy and digital filter response," *IEEE Trans. Circuit Theory*, vol. CT-15, pp. 31-41, Mar. 1968.

[18] M. Kawamata and T. Higuchi, "A unified approach to the optimal synthesis of fixed-point state-space digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 911-920, Aug. 1985.

[19] A. Fettweis, "On the connection between multiplier word length limitation and roundoff noise in digital filters," *IEEE Trans. Circuit Theory,*, vol. CT-19, pp. 486-491, Sept. 1972.

[20] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 884-888, Oct. 1984.

[21] M. Iwatsuki, M. Kawamata, and T. Higuchi, "Statistical sensitivity and minimum sensitivity structures with fewer coefficients in discrete time linear systems," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 72-80, Jan. 1990.

[22] L. Thiele, "On the sensitivity of linear state-space systems," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 502-510, May 1986.

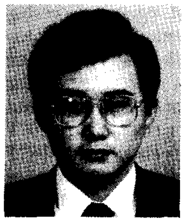
[23] K. H. Chang and W. G. Bliss, "Register modeling for the unification of quantization analysis in IIR digital filters," in *Proc. Asilomar Conf.*, Oct. 1992, pp. 278-282.

[24] ——, "Roundoff and coefficient quantization noise of pipelined scattered look-ahead filters with decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1992, vol. 4, pp. 433-436.

[25] Texas Instruments, *TMS320C25 Digital Signal Processor Design Workshop*, ver. 4.0, Nov. 1989.

[26] K. H. Chang and W. G. Bliss, "Limit cycles in state-space pipelined scattered look-ahead filters," in *Proc. IEEE 35th Midwest Symp. on Circuits Syst.*, Aug. 1992, pp. 1128-1131.

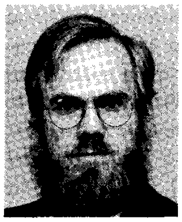
[27] ——, "Limit cycle behavior of pipelined recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. 41, pp. 351-355, May 1994.



KyungHi Chang (S'87-M'92) was born in Yesan, Korea, in 1962. He received the B.S. and M.S. degrees in electronics engineering from the Yonsei University, Seoul, in 1985 and 1987, respectively. He received the Ph.D. degree in electrical engineering from the Texas A&M University, College Station, in 1992.

From 1989 to 1990, he was with Samsung Advanced Institute of Technology (SAIT), Kihung, Korea, as a Researcher, and was involved in the DSP system design of a home-use digital VCR. In 1992, he joined the Integrated Circuits Development Section of Electronics and Telecommunications Research Institute (ETRI), Taejon, Korea, as a Senior Researcher, and is currently implementing CDMA system. His research interests include digital filter design and performance analysis, parallel architecture, VLSI signal processing, VLSI systems design, and mobile communication system design.

Dr. Chang is a member of the IEEE Signal Processing, Circuits and Systems, and Communications Societies.



William G. Bliss (S'88-M'89) received the B.S. degree from South Dakota State University in 1975, the M.S. degree from the University of Minnesota in 1975, and the Ph.D. degree from the University of Colorado at Boulder in 1988, all in electrical engineering.

He was with the IBM Development Labs in Rochester, MN, and Boulder, CO from 1977-1983, working on various aspects of magnetic recording including modulation coding, error correction coding, phase-locked loop design, and optimal reciever design. He was a consultant to Saxpy Computer Corp. from 1987 to 1988 in the areas of fault-tolerant array computing and parallel algorithms for moving image sequence estimation. He served a four-year term on the faculty of Texas A&M University, Department of Electrical Engineering, from 1988 to 1992, where his research included parallel algorithms for digital signal processing, finite arithmetic effects in digital signal processing, and digital control of analog signal processing circuits. He joined the technology group of Cirrus Logic, Broomfield, CO, in 1992, where he works on VLSI architectures for magnetic recording read channels.