A Matrix Factorization-Based Structure for Digital Filters

Gang Li, Jian Chu, and Jun Wu

***Abstract*—In this correspondence, a novel digital filter structure is derived using matrix factorization. This structure, called the LCW-structure, is an improved version of the so-called LGS-structure proposed by Li, Gevers, and Sun ["Performance Analysis of a New Structure for Digital Filter Implementation," *IEEE Transactions on Circuits and Systems I*, vol. 47, no. 4, pp. 474–482, April 2000]. For a digital filter of order N, this new structure requires $4N-1$ multiplications and $4N-1$ additions for computing one filter output sample, instead of $7N-3$ and $6N-3$, respectively, which are needed for the LGS-structure. An expression of the roundoff noise gain is derived for the LCW-structure. Design examples are presented to demonstrate the behavior of this structure and to compare with several well-known digital filter structures in terms of minimizing roundoff noise and implementation efficiency.**

***Index Terms*—Digital filter structures, finite-word-length (FWL) effects, roundoff noise, state-space realizations.**

I. Introduction

CONSIDER a well-designed time-invariant linear digital filter of transfer function $H(z)$. Such a filter can be implemented with many different *structures*[1] such as its state-space equations:

$$\begin{cases} x(n+1) = Ax(n) + Bu(n) \\ y(n) = Cx(n) + du(n) \end{cases} \tag{1}$$

where $u(n)$ and $y(n)$ are the input and output of the filter, respectively. (A, B, C, d) is called a state-space realization of $H(z)$ with $A \in \mathcal{R}^{N\times N}, B \in \mathcal{R}^{N\times 1}, C \in \mathcal{R}^{1\times N}$ and $d \in \mathcal{R}$, satisfying $H(z) = d + C(zI - A)^{-1}B$. The state-space realizations are not unique. Denote S_H as the set of all the realizations of $H(z)$. S_H is characterized by

$$A = T^{-1}A_0T \quad B = T^{-1}B_0 \quad C = C_0T \tag{2}$$

where $(A_0, B_0, C_0, d) \in S_H$ and $T \in \mathcal{R}^{N\times N}$ is *any* nonsingular matrix.

One of the serious problems in filter implementation is the finite-word-length (FWL) effects. The classical optimal state-space realization problem in terms of minimizing roundoff noise was originally solved in [1] and [2]. The optimal realizations, on the one hand, can reduce the roundoff noise significantly, they, on the other hand, are fully parametrized, that is, they are full of nontrivial parameters.[2] This implies that to compute one output sample, $(N+1)^2$ multiplications and $N(N+1)$ additions are required.

A lot of effort has been made to achieve sparse optimal or quasi-optimal realizations [4]–[6]. Based on the δ-operator, sparse structures were derived for the class of low-pass narrow bandwidth filters [7]–[9]. These structures have recently been extended to a wider range of digital filters by optimizing the so-called polynomial operators in [10]–[12].

A class of orthogonal polynomial-based structures, known as lattice structures, was used to reduce roundoff noise in [13], where three efficient structures were analyzed. These three structures have some interesting properties [14], among which the normalized lattice structure is shown to be vastly superior to the others and the direct form-based structures in terms of minimizing roundoff. Such a structure contains $5N+1$ multipliers and has been considered as one of the best structures in many applications such as filter implementation and adaptive IIR filtering [15]–[17]. Another interesting realization is an input balanced realization, denoted as $(A_{\rm ib}, B_{\rm ib}, C_{\rm ib}, d)$, which is fully parametrized. Based on a factorization of $A_{\rm ib}$, an efficient structure, called LGS-structure, was proposed in [18]. Such structure, besides the excellent performance against the FWL effects, requires only $7N-3$ multiplications and $6N-3$ additions for computing one output sample.

The main objective in this correspondence is to derive an alternative filter structure to the LGS-structure, that is the LCW-structure, and to analyze its behavior towards roundoff noise. This proposed structure is more efficient and as to be seen, has a better performance against roundoff noise than the LGS-structure and the normalized lattice structure.

II. Preliminaries

For a given digital signal processor that is used to implement the filter, the signal dynamical range is fixed. The state variables in $x(n)$ have to be stored in order to compute the next state vector $x(n+1)$. The magnitude of these variables is structure dependent. To keep these signals within a certain dynamical range, the actually implemented realization (A, B, C, d) has to be properly scaled in order to avoid signal overflow (see [1] and [2]). There exist several scaling schemes. The popularly used l_2-scaling means that each state variable should have a unit variance when the input is a white noise with a unit variance. The controllability and observability Gramians, denoted as W_c and W_o, respectively, of the realization (A, B, C, d) are the solution of the following equations:

$$W_c = AW_cA^{\mathcal{T}} + BB^{\mathcal{T}}, \quad W_o = A^{\mathcal{T}}W_oA + C^{\mathcal{T}}C. \tag{3}$$

The l_2-scaling can be achieved if [1], [2]

$$W_c(k,k) = 1 \quad \forall k. \tag{4}$$

It is easy to see that if the realization (A, B, C, d) is transformed from an initial realization (A_0, B_0, C_0, d) with T, as specified by (2), then

$$W_c = T^{-1}W_c^0T^{-\mathcal{T}}, \quad W_o = T^{\mathcal{T}}W_o^0T \tag{5}$$

where W_c^0, W_o^0 are the Gramians of the initial realization and $\mathcal{T}$ is the transpose operation.

Let $F(z)$ be the transfer function of a multiple-input single-output (MISO) system. When the input signal $w(n)$ is a white process, that is $E[w(n)w^{\mathcal{T}}(n-m)] = R_w\delta(m)$, where $E[\cdot]$ denotes the statistical average operation, R_w is a constant matrix, and $\delta(m)$ is the unit sample

Manuscript received December 16, 2006; revised February 5, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mariane R. Petraglia. G. Li would like to acknowledge the financial support provided by Zhejiang University for his visit during summer 2006.

G. Li is with the College of Information Engineering, Zhejiang University of Technology, Zhejiang, 310014, China. (e-mail: lig_61@yahoo.com and egli@ntu.edu.sg).

J. Chu and J. Wu are with the National Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Zhejiang 310027, China (e-mail: chuj@iipc.zju.edu.cn; jwu@iipc.zju.edu.cn).

Digital Object Identifier 10.1109/TSP.2007.896111

[1]Here, a structure of a digital filter means a way how the filter output is computed with a given input signal.

[2]By *trivial parameters*, we mean those that are 0 and ± 1. Other parameters are, therefore, referred to *nontrivial parameters*.

function, it is well known that the corresponding output $v(n)$ of $F(z)$ is stationary and

$$E[v^2(n)] = \frac{1}{j2\pi}\oint_{|z|=1} F(z)R_w F^{\mathcal{H}}(z)z^{-1}\,dz$$

with $\mathcal{H}$ denoting the conjugate-transpose. Let (A, B, C, D) be a realization of $F(z)$ and W_o is the corresponding observability Gramian. According to the residue theory, it can be shown that

$$E[v^2(n)] = \mathrm{tr}[(B^{\mathcal{T}}W_oB + D^{\mathcal{T}}D)R_w] \tag{6}$$

where $\mathrm{tr}[\,\cdot\,]$ is the trace operator.

Another FWL related issue is the rounding operations due to multiplications between signals and nontrivial parameters. There exist several rounding schemes, among which the fixed-point arithmetic is the simplest one. This scheme is popularly used in real-time applications, where the processing speed is one of the main concerns, and is assumed in the sequel. Let τ be a nontrivial parameter (multiplier) in a filter structure, say (1). In an actual implementation of less-than-double precision with *rounding after multiplication*, the product $\tau s(n)$ has to be rounded by a quantizer $q[\,.\,]$. Denote $\epsilon_\tau(n) \triangleq q[\tau s(n)] - \tau s(n)$ as the roundoff noise. As is well known [1], [2], roundoff noises can be modeled as statistically independent white processes and $E[\epsilon_\tau^2(n)] = \sigma_0^2$ is constant, uniquely determined by the word length used for representing the states. Denote $\Delta y(n)$ as the corresponding output deviation of the filter due to $\epsilon_\tau(n)$ and $H_\tau(z)$ as the transfer function between $\epsilon_\tau(n)$ and $\Delta y(n)$. The roundoff noise gain for the parameter τ is defined as

$$G_\tau \triangleq \frac{E[(\Delta y(n))^2]}{E[\epsilon_\tau^2(n)]}.$$

If S_τ is the set of all the nontrivial parameters for a given structure, the overall roundoff noise gain is defined as $G \triangleq \sum_{\tau\in S_\tau} G_\tau$.

It can be shown (see [1], [2], and [18]) that the overall roundoff noise gain of a fully parametrized state-space realization (1) is given by

$$G = [\mathrm{tr}(W_o) + 1](N+1) \tag{7}$$

where W_o is the observability Gramian of the realization (A, B, C, d) of digital filter $H(z)$. The classical optimal roundoff realization problem is to identify those (fully parametrized) realizations, denoted as R_f^{opt}, which minimize G given by (7) subject to (4). This problem was solved in [1] and [2]. The corresponding roundoff noise gain is given by

$$G_f^{\mathrm{opt}} = \left[1 + \frac{1}{N}\left(\sum_{k=1}^{N}\sigma_k\right)^2\right](N+1) \tag{8}$$

where $\sigma_k\ \forall k$ are the singular values of the filter with σ_k equal to the square root of the kth eigenvalue of W_cW_o.

A very interesting state-space realization, denoted as (Φ, K, L, d_s), was proposed by Johns, Snelgrove, and Sedra in [19] for implementing stable analog filters. This realization, called JSS-realization, takes the following sparse form (for $N = 3$ as example):

$$\Phi = \begin{bmatrix} 0 & \alpha_1 & 0 \\ -\alpha_1 & 0 & \alpha_2 \\ 0 & -\alpha_2 & -\alpha_3 \end{bmatrix} \quad K = \begin{bmatrix} 0 \\ 0 \\ \sqrt{2\alpha_3} \end{bmatrix} \tag{9}$$

with $\alpha_k > 0\ \forall k$ and L having no special structure. Using the bilinear transformation $s = (z-1)/(z+1)$, one can show that any stable Nth digital filter $H(z)$ can be implemented with the following realization $(A_{\mathrm{ib}}, B_{\mathrm{ib}}, C_{\mathrm{ib}}, d)$:

$$A_{\mathrm{ib}} = (I+\Phi)(I-\Phi)^{-1}, \quad B_{\mathrm{ib}} = \frac{\sqrt{2}}{2}(I + A_{\mathrm{ib}})K,$$
$$C_{\mathrm{ib}} = \frac{\sqrt{2}}{2}L(I + A_{\mathrm{ib}}). \tag{10}$$

It is interesting to note that this is an input balanced realization as the corresponding controllability Gramian W_c^{ib} is equal to the identity matrix I and it will be denoted as R_{ib} in the sequel. Please refer to [18] for how to compute the corresponding $\{\alpha_k\}$ for a given $H(z)$.

Like the JSS-realization, R_{ib} has very good numerical properties against the FWL effects [18]. It, however, is fully parametrized. The corresponding roundoff noise gain is given by

$$G_{\mathrm{ib}} = \left[1 + \mathrm{tr}\left(W_o^{\mathrm{ib}}\right)\right](N+1) = \left(1 + \sum_{k=1}^{N}\sigma_k^2\right)(N+1) \tag{11}$$

with $\{\sigma_k\}$, as defined before, the singular values of the filter, and W_o^{ib} is the observability Gramian of R_{ib}.

III. LCW-Structure

First of all, noting $I - \Phi^2 = (I+\Phi)(I-\Phi) = (I-\Phi)(I+\Phi)$, one can show that

$$A_{\mathrm{ib}} = (I+\Phi)(I-\Phi)^{-1} = (I-\Phi)^{-1}(I+\Phi) = 2(I-\Phi)^{-1} - I \tag{12}$$

and hence $B_{\mathrm{ib}} = (\sqrt{2})/(2)(I + A_{\mathrm{ib}})K = \sqrt{2}(I-\Phi)^{-1}K$.

It is well known that the dual realization (A_t, B_t, C_t, d) of R_{ib} is also a realization of the same filter $H(z)$: $A_t = A_{\mathrm{ib}}^{\mathcal{T}} = (I+\Phi)^{\mathcal{T}}(I-\Phi)^{-\mathcal{T}}$, $B_t = C_{\mathrm{ib}}^{\mathcal{T}}$, $C_t = B_{\mathrm{ib}}^{\mathcal{T}} = \sqrt{2}K_{\mathrm{ib}}^{\mathcal{T}}(I-\Phi)^{-\mathcal{T}}$.

Using the similarity transformation $T \triangleq (I-\Phi)^{\mathcal{T}}$, the dual realization is transformed as (A^*, B^*, C^*, d)

$$A^* = (I-\Phi)^{-\mathcal{T}}(I+\Phi)^{\mathcal{T}} = A_{\mathrm{ib}}^{\mathcal{T}},$$
$$B^* = (I-\Phi)^{-\mathcal{T}}C_{\mathrm{ib}}^{\mathcal{T}}, \quad C^* = \sqrt{2}K^{\mathcal{T}}. \tag{13}$$

One can see that this realization is sparser than R_{ib} due to the special structure of K and hence C^*.

According to (5), the Gramians of the realization (A^*, B^*, C^*, d) are

$$W_c^* = (I-\Phi)^{-\mathcal{T}}W_o^{\mathrm{ib}}(I-\Phi)^{-1},$$
$$W_o^* = (I-\Phi)(I-\Phi)^{\mathcal{T}}. \tag{14}$$

Define $\beta_{k+1} \triangleq -\alpha_{k+1}\gamma_k$, $\gamma_k \triangleq (1)/(1-\alpha_k\beta_k)$ for $k = 1, 2, 3, \ldots, N-2$ with $\gamma_{N-1} = (1)/(1+\alpha_N - \alpha_{N-1}\beta_{N-1})$, $\beta_1 = -\alpha_1$, and $U(i, j, x)$ as the unit matrix of dimension N except that its (i, j)th element is $x\ \forall(i, j)$. With

$$T_k = U(k+1, k+1, \gamma_k)U(k+1, k, \alpha_k) \triangleq A_{2k}A_{2k-1} \tag{15}$$

for $k = 1, 2, \ldots, N-1$, it can be shown that $T_{N-1}T_{N-2}\ldots T_k\ldots T_2T_1(I-\Phi)^{\mathcal{T}} = \Gamma$, where Γ is a matrix whose elements are all zero except $\Gamma(k, k) = 1, \Gamma(k, k+1) = -\beta_k\ \forall k$.

Since Γ^{-1} can be decomposed as

$$U(1,2,\beta_1)U(2,3,\beta_2)\dots U(k,k+1,\beta_k)\dots U(N-1,N,\beta_{N-1}) \triangleq A_{3(N-1)}A_{3(N-1)-2}\dots A_{2(N-1)+2}A_{2(N-1)+1} \quad (16)$$

one has the following:

$$(I-\Phi)^{-\mathcal{T}} = A_{3(N-1)}A_{3(N-1)-1}\dots A_m\dots A_2A_1 \triangleq \prod_{m=1}^{3(N-1)} A_m \quad (17)$$

with $A_m\ \forall m$ defined in (15) and (16). Therefore, $A^* = 2(I-\Phi)^{-\mathcal{T}} - I = 2\prod_{m=1}^{3(N-1)} A_m - I$.

Looking at (14), one can see that the realization (A^*, B^*, C^*, d) is generally not l_2-scaled, which means that $s_k^2 \triangleq W_c^*(k,k)$ may not be equal to one for certain index k. With the diagonal similarity transformation $T_s \triangleq \mathrm{diag}(s_1, s_2, \dots, s_N)$, the realization (A^*, B^*, C^*, d) is transformed as $(\tilde{A}, \tilde{B}, \tilde{C}, d)$

$$\begin{aligned}\tilde{A} &\triangleq T_s^{-1}A^*T_s = 2\prod_{m=1}^{3(N-1)}\tilde{A}_m - I\\ \tilde{B} &\triangleq T_s^{-1}B^* = T_s^{-1}C_{\mathrm{ib}}^{\mathcal{T}}\\ \tilde{C} &\triangleq C^*T_s = \sqrt{2}s_N K^{\mathcal{T}}\end{aligned} \quad (18)$$

where $\tilde{A}_m \triangleq T_s^{-1}A_mT_s, \forall m, \tilde{C}$, and $\tilde{A}_m$ have the same structure as $K^{\mathcal{T}}$ and A_m, respectively. In fact, $\tilde{A}_m$ is equal to A_m with $\{\alpha_k, \beta_k, \gamma_k\}$ replaced with $\{\tilde{\alpha}_k, \tilde{\beta}_k, \tilde{\gamma}_k\}$, as follows:

$$\begin{aligned}&\tilde{\alpha}_k \triangleq s_{k+1}^{-1}s_k\alpha_k,\\ &\tilde{\beta}_k \triangleq s_k^{-1}s_{k+1}\beta_k, \tilde{\gamma}_k \triangleq \gamma_k\end{aligned} \quad (19)$$

for $k = 1, 2, \dots, N-1$.

Denote $x^{(0)}(n) \triangleq 2x(n)$. Based on the factorization of $\tilde{A}$, one can see that the l_2-scaled state-space realization (18) can be implemented with the following iterative way:

$$\begin{cases} x^{(m)}(n) = \tilde{A}_m x^{(m-1)}(n)\\ x(n+1) = x^{(3(N-1))}(n) - x(n) + \tilde{B}u(n)\\ y(n) = \tilde{C}x(n) + du(n).\end{cases} \quad (20)$$

This structure is referred as LCW-structure.

The LSG-structure in [18] was derived directly from R_{ib}, where A_{ib} is factorized into $N_0 \triangleq 3(N-1)+1$ sparse matrices: $A_{\mathrm{ib}} = \prod_{m=1}^{N_0} A^{(m)}$, where $A^{(N_0)} = I + \Phi$ and $A^{(m)} = SA_mS$ for $m = 1, 2, \dots, 3(N-1)$ with S a sign matrix defined as $S \triangleq \mathrm{diag}(\eta_1, \eta_2, \dots, \eta_k, \dots, \eta_N), \eta_k = (-1)^k\ \forall k$. Letting $x^{(0)}(n) = x(n)$, the LGS-structure is then specified as

$$\begin{cases} x^{(m)}(n) = A^{(m)}x^{(m-1)}(n)\\ x(n+1) = x^{(N_0)}(n) + B_{\mathrm{ib}}u(n)\\ y(n) = C_{\mathrm{ib}}x(n) + du(n).\end{cases} \quad (21)$$

Simple calculation shows that the LSG-structure (21) needs $7N-3$ multiplications and $6N-3$ additions for computing one output sample $y(n)$. Noting that $x^{(0)}(n) = 2x(n)$ in (20) can be implemented using shifting operations rather than multiplications, one can see that the LCW-structure needs only $4N-1$ multiplications and $4N-1$ additions. In that sense, the proposed LCW-structure is much simpler in terms of implementation. Another important advantage of the LCW-structure is the sparseness of $\tilde{C}$, whose contribution to the overall roundoff noise gain is one, while for the LGS-structure, C_{ib} is fully parametrized and hence its contribution is N. Therefore, the proposed LCW-structure is also expected to have a smaller roundoff gain. These are actually the main motivations of this new structure.

IV. Roundoff Noise Analysis

In the LCW-structure, the nontrivial parameters are $\{\tilde{\alpha}_k, \gamma_k, \tilde{\beta}_k\}$, the N elements in $\tilde{B}$, the Nth element of $\tilde{C}$, and d. Now, let us consider the quantization errors caused by the nontrivial parameters $\tilde{\alpha}_k$ and γ_k, which occur in $\tilde{A}_{2k-1}$ and $\tilde{A}_{2k}$, respectively. See (15). The actual model of the LCW-structure (20) is then

$$\begin{cases} \hat{x}^{(m)}(n) = \tilde{A}_m\hat{x}^{(m-1)}(n), m \neq 2k-1, 2k\\ \hat{x}^{(2k)}(n) = q\left[\tilde{A}_{2k}q\left[\tilde{A}_{2k-1}\hat{x}^{(2k-2)}(n)\right]\right]\\ \hat{x}(n+1) = \hat{x}^{(3(N-1))}(n) - \hat{x}(n) + \tilde{B}u(n)\\ \hat{y}(n) = \tilde{C}\hat{x}(n) + du(n)\end{cases} \quad (22)$$

where $\hat{x}^{(0)}(n) \triangleq 2\hat{x}(n)$, $q[Mv(n)]$ is the quantizer rounding all products that occur in the multiplication of a matrix M and a vector $v(n)$ of proper dimension.

Denote $e_\alpha(n) \triangleq q[\tilde{A}_{2k-1}\hat{x}^{(2k-2)}(n)] - \tilde{A}_{2k-1}\hat{x}^{(2k-2)}(n)$ and $e_\gamma(n) \triangleq q[\tilde{A}_{2k}q[\tilde{A}_{2k-1}\hat{x}^{(2k-2)}(n)]] - \tilde{A}_{2k}q[\tilde{A}_{2k-1}\hat{x}^{(2k-2)}(n)]$. The second equation in (22) can then be rewritten as

$$\hat{x}^{(2k)}(n) = \tilde{A}_{2k}\tilde{A}_{2k-1}\hat{x}^{(2k-2)}(n) + \tilde{A}_{2k}e_\alpha(n) + e_\gamma(n).$$

Let $\Delta x^{(m)}(n) \triangleq \hat{x}^{(m)}(n) - x^{(m)}(n), \forall m, \Delta x(n) \triangleq \hat{x}(n) - x(n)$, and $\Delta y(n) \triangleq \hat{y}(n) - y(n)$. It follows from (20) and (22) that

$$\begin{cases} \Delta x^{(m)}(n) = \tilde{A}_m\Delta x^{(m-1)}(n), \quad m \neq 2k-1, 2k\\ \Delta x^{(2k)}(n) = \tilde{A}_{2k}\tilde{A}_{2k-1}\Delta x^{(2k-2)}(n) + \tilde{A}_{2k}e_\alpha(n) + e_\gamma(n)\\ \Delta x(n+1) = \Delta x^{(3(N-1))}(n) - \Delta x(n)\\ \Delta y(n) = \tilde{C}\Delta x(n)\end{cases}$$

with $\Delta x^{(0)}(n) = 2\Delta x(n)$.

Noting the special form of $\tilde{A}_{2k-1}, \tilde{A}_{2k}$, one has $e_\alpha(n) = \epsilon_\alpha(n)v_{k+1}, e_\gamma = \epsilon_\gamma(n)v_{k+1}$, where v_{k+1} is the $(k+1)$th elementary (column) vector, whose elements are all zero except the $(k+1)$th which is one, and $\epsilon_\alpha(n), \epsilon_\gamma(n)$ are the roundoff noise produced by $\tilde{\alpha}_k$ and γ_k, respectively. It follows from $\tilde{A}_{2k}v_{k+1} = \gamma_k v_{k+1}$ that

$$\begin{cases} \Delta x(n+1) = \tilde{A}\Delta x(n) + \tilde{P}_k^{(1)}[\tilde{\gamma}_k\epsilon_\alpha(n) + \epsilon_\gamma(n)]\\ \Delta y(n) = \tilde{C}\Delta x(n)\end{cases}$$

where

$$\tilde{P}_k^{(1)} \triangleq \left(\prod_{m=2k+1}^{3(N-1)}\tilde{A}_m\right)v_{k+1}, \quad k = 1, 2, \dots, N-1. \quad (23)$$

According to (6), the output error variance due to $\tilde{\alpha}_k$ and γ_k is given by

$$\begin{aligned}E[(\Delta y(n))^2] &= \mathrm{tr}\left[\left(\tilde{P}_k^{(1)}\right)^{\mathcal{T}}\tilde{W}_o\tilde{P}_k^{(1)}I\left(1+\gamma_k^2\right)\sigma_0^2\right]\\ &= \mathrm{tr}\left[\tilde{W}_o\tilde{P}_k^{(1)}\left(\tilde{P}_k^{(1)}\right)^{\mathcal{T}}\left(1+\gamma_k^2\right)\right]\sigma_0^2\end{aligned}$$

with σ_0^2 defined before and $\tilde{W}_o$, the observability Gramian of the realization (18). Therefore, the total output error variance due to all $\{\tilde{\alpha}_k, \gamma_k\}$ is $\mathrm{tr}[\tilde{W}_o\tilde{R}_1]\sigma_0^2$ with $\tilde{R}_1 \triangleq \sum_{k=1}^{N-1}(1+\gamma_k^2)\tilde{P}_k^{(1)}(\tilde{P}_k^{(1)})^{\mathcal{T}}$.

Using the same procedure, one can derive the expression for the output error variance due to $\tilde{\beta}_k$ which occurs in $\tilde{A}_{3(N-1)+1-k}$. See (16). In fact, it can be shown that the output error due to $\tilde{\beta}_k$ is given by

$$\begin{cases} \Delta x(n+1) = \tilde{A}\Delta x(n) + \tilde{P}_k^{(2)}\epsilon_\beta(n)\\ \Delta y(n) = \tilde{C}\Delta x(n)\end{cases}$$

where

$$\tilde{P}_k^{(2)} \triangleq \begin{cases}\left(\prod_{m=3(N-1)+2-k}^{3(N-1)}\tilde{A}_m\right)v_k, & k \neq 1\\ v_1, & k = 1.\end{cases} \quad (24)$$

Therefore, the total output error variance due to all $\{\tilde{\beta}_k\}$ is $\mathrm{tr}[\tilde{W}_o\tilde{R}_2]\sigma_0^2$ with $\tilde{R}_2 \triangleq \sum_{k=1}^{N-1}\tilde{P}_k^{(2)}(\tilde{P}_k^{(2)})^{\mathcal{T}}$. Similarly, the error variance due to $\tilde{B}, \tilde{C}$, and d can be shown to be given by $[2+\mathrm{tr}(\tilde{W}_o)]\sigma_0^2$. The overall roundoff noise gain, denoted as G_{lcw}, of our proposed structure is given by

$$G_{\mathrm{lcw}} = \mathrm{tr}[\tilde{W}_o(\tilde{R}_1+\tilde{R}_2+I)]+2. \tag{25}$$

For the LGS-structure specified by (21), it can be shown in the same way that the corresponding roundoff noise gain is given by

$$G_{\mathrm{lgs}} = \mathrm{tr}\left[W_o^{\mathrm{ib}}R_{\mathrm{lgs}}\right]+(N+1) \tag{26}$$

where $R_{\mathrm{lgs}} \triangleq (I+\Phi)(R_1+R_2)(I+\Phi)^{\mathcal{T}}+R_3$ with R_1, R_2 of exactly the same form as $\tilde{R}_1, \tilde{R}_2$, respectively, except $\tilde{A}_m$ is replaced by $A^{(m)}$, and $R_3 = \mathrm{diag}(2,3,3,\ldots,3,3)$.

The normalized lattice structure contains $5N+1$ nontrivial parameters. It is shown [17] that the normalized lattice filter is equivalent to the following revised state-space realization:

$$\begin{bmatrix} x(n+1) \\ w(n) \end{bmatrix} = Q\begin{bmatrix} x(n) \\ u(n) \end{bmatrix}, \quad y(n) = \bar{h}^{\mathcal{T}}\begin{bmatrix} x(n+1) \\ w(n) \end{bmatrix} \tag{27}$$

where $w(n)$ is an intermediate (auxiliary) signal, $\bar{h} = \{h_k\} \in \mathcal{R}^{(N+1)\times 1}$ and $Q \in \mathcal{R}^{(N+1)\times(N+1)}$ is an orthogonal matrix, having the following factorization form:

$$Q = Q_N Q_{N-1}\ldots Q_{N+1-k}\ldots Q_2 Q_1 \triangleq \prod_{m=1}^{N} Q_m. \tag{28}$$

The equivalent (standard) state-space realization to (27), denoted as $(A_{nl}, B_{nl}, C_{nl}, d)$, can be easily obtained. Applying the same approach, one can analyze the roundoff noise behavior of (27). It can be shown that the roundoff noise gain of the normalized lattice is given by[3]

$$G_{nl} = \mathrm{tr}\left[([I\ \mathbf{0}]^{\mathcal{T}}W_o^{nl}[I\ \mathbf{0}]+\bar{h}\bar{h}^{\mathcal{T}})\sum_{k=1}^{N}R_k^{nl}\right]+(N+1) \tag{29}$$

where W_o^{nl} is the observability Gramian of $(A_{nl}, B_{nl}, C_{nl}, d)$ and $R_k^{nl} \triangleq 2P_k[v_k v_k^{\mathcal{T}}+v_{k+1}v_{k+1}^{\mathcal{T}}]P_k^{\mathcal{T}}$ with

$$P_k \triangleq \begin{cases} \prod_{m=N+2-k}^{N} Q_m, & k \neq 1 \\ I_{N+1}, & k = 1. \end{cases}$$

It should be pointed out that an alternative expression for G_{nl} was derived in [13], and ours has the advantages in terms of conciseness and computation efficiency.

V. Numerical Examples and Simulation Results

In this section, we will present two numerical examples to examine the roundoff noise performance of six structures, which are the proposed LCW-structure R_{lcw}, the l_2-scaled controllable realization R_c, R_{ib}, and a fully parametrized optimal realization R_f^{opt}, the LGS-structure R_{lgs}, and the normalized lattice structure R_{nl}. The structure complexity of a structure is measured by the numbers of multiplications and additions, denoted as N_M and N_A, respectively, which are required to compute one output sample. Table I provides the information about the structure complexity, where N is the filter order.

[3]Note that the roundoff noise gain due to the $(N+1)$ parameters in $\bar{h}$, which are usually nontrivial, is $(N+1)$.

TABLE I
Structure Complexity Comparison

	R_c	R_f^{opt}	R_{ib}	R_{lgs}	R_{nl}	R_{lcw}
N_M	$2N+2$	$(N+1)^2$	$(N+1)^2$	$7N-3$	$5N+1$	$4N-1$
N_A	$2N$	N^2+N	N^2+N	$6N-3$	$3N+1$	$4N-1$

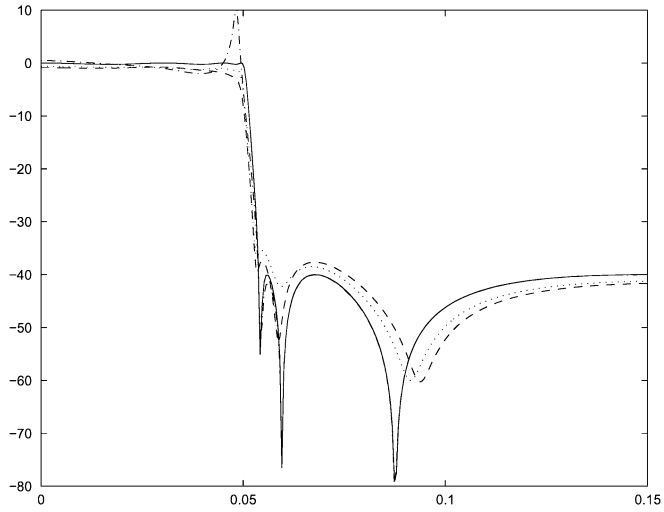

Fig. 1. Magnitude responses for Example I: solid line: ideal; dotted line: R_{lcw} with $B_c = 8$; dashed line: R_{nl} with $B_c = 8$; and dashed–dotted line: R_c with $B_c = 20$. The x axis is for the normalized frequency.

TABLE II
Roundoff Noise Gain for Each of the Six Structures

	R_c	R_f^{opt}	R_{ib}	R_{lgs}	R_{nl}	R_{lcw}
Example I	1.0253×10^{11}	19.2149	26.0154	20.7617	17.2683	10.1027
Example II	3.1315×10^{9}	23.5817	30.8367	23.2032	19.3118	10.7685

Example I: A seventh-order low-pass elliptic filter is generated with Matlab command

$$\mathrm{ellip}(7, 0.25, 40, 0.1).$$

The corresponding magnitude response is depicted in Fig. 1 with solid line.

Table II shows the roundoff noise gain for each of the six structures in this example, from which one can see that R_c, though the simplest, has a huge roundoff noise, while R_f^{opt} and R_{ib} have a much smaller roundoff noise, but both require many more multiplications and additions, which will slow down the processing. The other three structures, however, not only have a very small roundoff noise but also are very efficient. In fact, for this example both R_{nl} and R_{lcw} have a roundoff noise gain even smaller than R_f^{opt} with a structure complexity comparable with that of R_c.

Remark: Such a filter $H(z)$, as proposed in [15], can be implemented using a parallel interconnection of two all-pass sub-filters $H_1(z)$ and $H_2(z)$ with each implemented in the normalized lattice structure. Since $H(z) = (1/2)[H_1(z)+H_2(z)]$, the roundoff noise gain for such a structure is $G = G_1+G_2+1$ with G_k the roundoff noise gain of $H_k(z)$, for $k = 1, 2$. For this particular example, we have $G = 16+12+1 = 29$, which implies that such a structure cannot do better than the normalized lattice structure of the whole filter $H(z)$ in terms of reducing roundoff noise and implementation complexity.

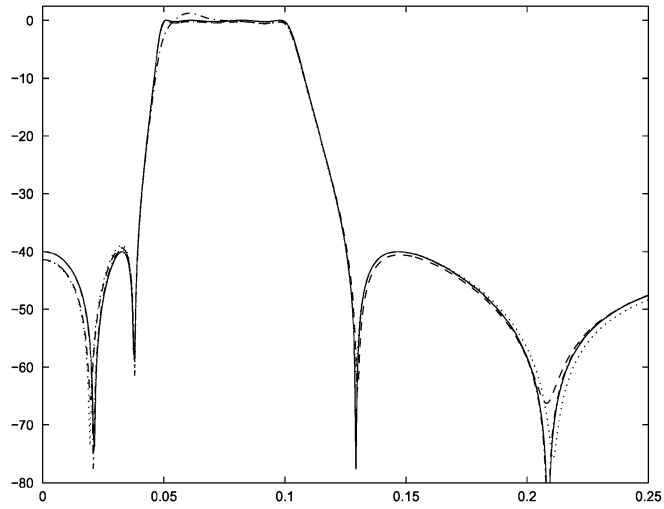

Fig. 2. Magnitude responses for Example II: solid line: ideal; dotted line: $R_{\rm lcw}$ with $B_c = 10$; dashed line: R_{nl} with $B_c = 10$; and dashed–dotted line: R_c with $B_c = 20$. The x axis is for the normalized frequency.

To confirm the theoretical results, simulations are done. The fractional part of each nontrivial parameter in a structure is truncated into a B_c-bit format, then the magnitude response of the truncated structure is computed. In Fig. 1, three responses are presented for R_c, R_{nl} and $R_{\rm lcw}$, respectively. One observes that with only $B_c = 8$ bits, R_{nl} and $R_{\rm lcw}$ yields a much better response than that of R_c truncated with $B_c = 20$ bits. It is also seen that $R_{\rm lcw}$ is slightly better than R_{nl} during the pass-band.

Example II: The second example is an eighth-order bandpass elliptic filter, which is generated with MATLAB command $\mathrm{ellip}(4, 0.25, 40, [0.10\ 0.20])$. The corresponding magnitude response is depicted in Fig. 2 with solid line.

The statistics for each of the six structures are shown in Table II and similar simulations are presented in Fig. 2. The same comments as those for Example I apply to this example.

VI. Conclusion

In this correspondence, the LCW-structure has been derived and analyzed. Two numerical examples and simulations have carried out, from which it is shown that the proposed structure yields the best performance among the six existing structures in terms of minimizing roundoff noise and an implementation efficiency comparable with the simplest R_c.

Acknowledgment

The authors would like to thank the reviewers for the constructive comments and suggestions.

References

[1] 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, Sep. 1976.

[2] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 8, pp. 273–281, Aug. 1977.

[3] C. Xiao, "Improved L_2-sensitivity for state-space digital system," *IEEE Trans. Signal Process.*, vol. 45, no. 4, pp. 837–840, Apr. 1997.

[4] 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, no. 1, pp. 72–80, Jan. 1989.

[5] B. W. Bomar and J. C. Hung, "Minimum roundoff noise digital filters with some power-of-two coefficients," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 833–840, Oct. 1984.

[6] G. Amit and U. Shaked, "Small roundoff realization of fixed-point digital filters and controllers," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-36, no. 6, pp. 880–891, Jun. 1988.

[7] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 41, no. 2, pp. 629–637, Feb. 1993.

[8] J. Kauraniemi, T. I. Laakso, I. Harttimo, and S. J. Ovaska, "Delta operator realizations of direct-form IIR filters," *IEEE Trans. Circuits Syst. II*, vol. 45, no. 1, pp. 41–45, Jan. 1998.

[9] N. Wong and T. S. Ng, "A generalized direct-form delta operator-based IIR filter with minimum noise gain and sensitivity," *IEEE Trans. Circuits Syst. II*, vol. 48, pp. 425–431, Apr. 2001.

[10] G. Li, "A polynomial-operator-based DFIIt structure for IIR filters," *IEEE Trans. Circuits Syst. II*, vol. 51, pp. 147–151, Mar. 2004.

[11] G. Li and Z. X. Zhao, "On the generalized DFIIt structure and its state-space realization in digital filter implementation," *IEEE Trans. Circuits Syst. I*, vol. 51, pp. 769–778, Apr. 2004.

[12] G. Li, C. R. Wan, and G. A. Bi, "An improved ρDFIIt structure for digital filters with minimum roundoff noise," *IEEE Trans. Circuts Syst. II*, vol. 52, no. 4, pp. 199–203, Apr. 2005.

[13] J. D. Markel and A. H. Gray Jr., "Roundoff noise characteristics of a class orthogonal polynomial structures," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 5, pp. 473–486, Oct. 1975.

[14] A. H. Gray Jr., "Passive cascaded lattice digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, no. 5, pp. 337–344, May 1980.

[15] P. P. Vaidyanathan, S. K. Mitra, and Y. Neuvo, "A new approach to the realization of low sensitivity IIR digital filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 2, pp. 350–361, Apr. 1986.

[16] P. A. Regalia, S. K. Mitra, and P. P. Vaidyanathan, "The digital all-pass filter: A versatile signal processing building block," *Proc. IEEE*, vol. 76, no. 1, pp. 19–37, Jan. 1988.

[17] P. A. Regalia, "Stable and efficient lattice algorithms for adaptive IIR filtering," *IEEE Trans. Signal Process.*, vol. 40, no. 2, pp. 375–388, Feb. 1992.

[18] G. Li, M. Gevers, and Y. X. Sun, "Performance analysis of a new structure for digital filter implementation," *IEEE Trans. Circuits Syst. I*, vol. 47, pp. 474–482, Apr. 2000.

[19] D. A. Johns, W. M. Snelgrove, and A. S. Sedra, "Orthonormal ladder filters," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 337–343, Mar. 1989.