Performance Analysis of a New Structure for Digital Filter Implementation

Gang Li, *Member, IEEE*, Michel Gevers, *Fellow, IEEE*, and Youxian Sun

***Abstract*—It is well known that for a digital filter of order n the number of nontrivial parameters in the classical optimal state-space realizations is proportional to n^2. In this paper, a new structure is proposed for digital filter implementation. This structure possesses $5n - 1$ nontrivial parameters and is shown to be equivalent to the discrete-time counterpart of the well-known *orthonormal ladder filter* in [13]. A new property of this ladder filter is revealed. The performance of the proposed structure is analyzed. Expressions for evaluating the sensitivity measure and the roundoff noise gain are derived. A numerical example is presented to illustrate the design procedure. Simulations show that the performance of the proposed structure is almost the same as that of the fully parameterized optimal realizations.**

***Index Terms*—Realizations, roundoff noise, sensitivity, structures.**

I. Introduction

MUCH attention has been paid to the numerical problems caused by the finite word length (FWL) effects in digital filter implementation for more than two decades. The optimal FWL state-space realization design has been considered as one of the most effective methods (see, e.g., [1]–[6]) to minimize the effects of FWL errors on the performance of digital filters and controllers. It is well known that for a digital filter there exist a number of different representations/parameterizations. Theoretically, they are equivalent since they represent the same system transfer function. However, different representations have different numerical properties and for a given application (measure or criterion) one representation can be better than another. A digital filter transfer function can be implemented with anyone of its state-space realizations. In digital filter implementation, the optimal FWL state-space design is to compute those realizations that minimize the degradation of the filter due to the FWL effects. These effects are classified into two categories: parameter perturbation and roundoff noise. During the last fifteen years, many results have been achieved in terms of finding optimal realizations that optimize a certain FWL effect related cost function such as sensitivity measure and roundoff noise gain; see, [1]–[6].

It has been noted that optimal realizations are usually fully parameterized. In practice, it is desired that the filter have a nice performance as well as a very simple structure [1] that possesses many trivial parameters[2], which can be implemented exactly and produce no rounding errors. Noting this fact, some modified measures were proposed (see, e.g., [7]–[9]). A lot of effort has been made to achieve sparse optimal or quasi-optimal realizations. For minimal sensitivity realizations, even though the number of nontrivial parameters can be much reduced using the degrees of freedom in the optimal realizations, the amount of nontrivial parameters is still proportional to n^2, where n is the order of filter (see, e.g., [10], [11]). Noting that the degrees of freedom in the optimal roundoff noise realizations are very limited, sophisticated numerical algorithms were developed to obtain the so-called quasi-optimal realizations that have a very sparse structure [7], [12]. Besides the numerical difficulty involved in those algorithms, the position of each trivial parameter in the obtained structure is not predictable. In fact, the quasi-optimal sparse structure for one digital filter may be very different from that for another.

Johns *et al.* [13] introduced a state-space structure for implementing *analog* filters. This structure, referred as *JSS-structure* in the sequel, has a very good sensitivity and dynamic range performance comparable to a cascade of biquads. Filters implemented with such structure are called *orthonormal ladder filters* and ensure optimum dynamic range scaling with an L_2-norm. Another interesting property of the JSS-structure is that it is very sparse. Its discrete-time counterpart, obtained with the *bilinear* transformation, is called *DJSS-structure* and can be shown to be inherently l_2-scaled. The FWL effects can be greatly reduced when a digital filter is implemented with the DJSS-structure. Unfortunately, the DJSS-structure, unlike the JSS-structure, is fully parameterized.

In this paper, we develop a new structure for digital filter implementation. This structure, referred to as *LGS-structure*, is very sparse. The basic idea in the development is to factorize the transition matrix of the DJSS-structure into a series of very sparse matrices. With such a factorization, the LGS-structure possesses $5n-1$ nontrivial parameters instead of n^2+2n+1. For each output, $7n-3$ multiplications and $6n-3$ additions are required, compared with n^2+2n+1 and n^2+n, respectively. One of the interesting properties of this structure is that the position of each parameter is always fixed. Performance analysis on this structure is conducted by deriving the expressions for the sensitivity measure and the roundoff noise gain. It should be pointed out that the LGS-structure does *not* belong to the

Manuscript received February 8, 1999; revised July 12, 1999. This paper was recommended by Associate Editor G. Martinelli.

G. Li is with the School of EEE, Nanyang Technological University, Singapore 639798 (e-mail: egli@ntu.edu.sg).

M. Gevers is with CESAME, Louvain University, Louvain-la-Neuve, Belgium (e-mail: gevers@csam.ucl.ac.be).

Y. Sun is with the Institute of Control Science and Engineering, Zhejiang University, 310027, Hangzhou, Zhejiang, China.

Publisher Item Identifier S 1057-7122(00)02959-7.

[1] Throughout this paper, a *structure* means a way in which the digital filter is implemented.

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

state-space realization set though it is theoretically equivalent to the DJSS-structure. The sparseness of this structure makes it a very attractive candidate for digital filter implementation. Simulations show that its performance can be even better than that of the fully parameterized DJSS-structure and of the optimal roundoff noise realizations.

An outline of this paper is given as follows. Section II provides some preliminaries on the optimal FWL state-space realization design. The JSS-structure is introduced in Section III. As our first contribution in this paper, a new stability property of this structure is also revealed in this section. Based on the DJSS-structure, the LGS-structure is developed in Section IV. Section V is devoted to analyzing the performance of the LGS-structure. In this section, the analytical expressions for computing the sensitivity measure and the roundoff noise gain are derived. A design example is given in Section VI to illustrate the design procedure and to compare the performance of the proposed structure with that of five well-known realizations.

II. Preliminaries

Consider a single-input\single-output time-invariant linear digital filter $H(z)$ implemented with its state space equations

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + du(t) \end{aligned} \tag{1}$$

where $u(t)$ and $y(t)$ are the scalar input and output of the filter, respectively.[3] $R \triangleq (A, B, C, d)$ with $A \in \mathcal{R}^{n\times n}, B \in \mathcal{R}^{n\times 1}, C \in \mathcal{R}^{1\times n}$ and $d \in \mathcal{R}$ is called a realization of $H(z)$, satisfying

$$H(z) = d + C(zI - A)^{-1}B. \tag{2}$$

Denote $S_H \triangleq \{R\text{: } R \text{ satisfies (2)}\}$. S_H is characterized by

$$A = T^{-1}A_0T, \quad B = T^{-1}B_0, \qquad C = C_0T \tag{3}$$

where $R_0 \triangleq (A_0, B_0, C_0, d) \in S_H$ and $T \in \mathcal{R}^{n\times n}$ is *any* nonsingular matrix.

A. Sensitivity Measure

For an actual implementation, the ideal parameters in (A, B, C, d) have to be truncated into FWL coefficients. Therefore, the actually implemented transfer function, denoted by $\tilde{H}(z)$, is different from $H(z)$. In the traditional FWL analysis, the parameter errors are modeled as zero mean uniformly distributed independent random variables [6], [18], [19]. Keeping this in mind, one can show (see, e.g., [6] and [14]) that the variance of the degradation $\Delta H(z) \triangleq \tilde{H}(z) - H(z)$ is proportional to the following sensitivity measure:

$$M_{L_2} \triangleq \sum_{k=1}^{N_p} \left\| \frac{\partial H(z)}{\partial p_k} \right\|_2^2 \tag{4}$$

where $\{p_k\}$ are the nontrivial parameters in the realization (A, B, C, d) and $\|\cdot\|_2$ is the L_2-norm defined below with $q = 2$.

Definition 1: Let $f(z) \in \mathcal{C}^{n\times m}$ be any complex matrix-valued function of the complex variable z. The L_q-norm of $f(z)$ is defined as

$$\|f\|_q \triangleq \left(\frac{1}{2\pi} \int_0^{2\pi} \|f(e^{j\omega})\|_F^q \, d\omega \right)^{1/q} \tag{5}$$

where $\|f(e^{j\omega})\|_F$ is the Frobenius norm of the matrix $f(e^{j\omega})$

$$\begin{aligned} \|f(e^{j\omega})\|_F &\triangleq \left(\sum_{i=1}^{n} \sum_{k=1}^{m} |f_{ik}(e^{j\omega})|^2 \right)^{1/2} \\ &= \{\mathrm{tr}[f^{\mathcal{T}}(e^{-j\omega})f(e^{j\omega})]\}^{1/2} \end{aligned} \tag{6}$$

with $\mathrm{tr}(\cdot)$ and $\mathcal{T}$ denoting the trace and transpose operations, respectively.

The parameter sensitivity $(\partial H/\partial p_k)$ can be found from

$$\begin{aligned} S_A(z) &\triangleq \frac{\partial H(z)}{\partial A} = G(z)F^{\mathcal{T}}(z) \\ S_B(z) &\triangleq \frac{\partial H(z)}{\partial B} = G(z) \\ S_C(z) &\triangleq \frac{\partial H(z)}{\partial C^{\mathcal{T}}} = F(z), \qquad S_d(z) \triangleq \frac{\partial H(z)}{\partial d} = 1 \end{aligned} \tag{7}$$

where

$$\begin{aligned} F(z) &\triangleq (zI - A)^{-1}B = [f_1(z) \ldots f_n(z)]^{\mathcal{T}} \\ G^{\mathcal{T}}(z) &\triangleq C(zI - A)^{-1} = [g_1(z) \ldots g_n(z)]. \end{aligned} \tag{8}$$

It is easy to see that different realizations have different sensitivity measures. The optimal sensitivity realization design problem is to identify those realizations that minimize M_{L_2}.

Since the minimization of M_{L2} was a hard problem, it was initially replaced by the minimization of the following L_1/L_2 mixed sensitivity measure [4]:

$$M_{L_1/L_2} \triangleq \left\| \frac{\partial H}{\partial A} \right\|_1^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C^{\mathcal{T}}} \right\|_2^2 + \left\| \frac{\partial H}{\partial d} \right\|_2^2. \tag{9}$$

It can be shown [4] that

$$\begin{aligned} M_{L_1/L_2} &\le \mathrm{tr}(W_c)\,\mathrm{tr}(W_o) + \mathrm{tr}(W_c) + \mathrm{tr}(W_o) + 1 \\ &\triangleq \bar{M}_{L_1/L_2} \end{aligned} \tag{10}$$

and

$$\min_{R\in S_H} M_{L_1/L_2} \Leftrightarrow \min_{R\in S_H} \bar{M}_{L_1/L_2} \tag{11}$$

where W_c, W_o are the controllability and observability gramians of (A, B, C, d), respectively, which are the solutions of the Lyapunov equations

$$W_c = AW_cA^{\mathcal{T}} + BB^{\mathcal{T}}, \qquad W_o = A^{\mathcal{T}}W_oA + C^{\mathcal{T}}C. \tag{12}$$

The optimal realizations, denoted by R_{L_1/L_2}, are characterized by the following identity:

$$W_c = W_o. \tag{13}$$

[3] It should be pointed out that the (time) index t is normally used for continuous-time systems. Throughout this paper, since the letters such as i, j, k, m, and n will be used for defining other variables, $t \in \{0, 1, 2, \ldots\}$ is defined as the time index for discrete-time systems.

The use of the L_1/L_2 mixed sensitivity measure leads to an easy treatment of the optimal realization problem. The pure L_2-based sensitivity measure is more logical since it has a clear physical interpretation [6], [8], [14].

Returning to the measure M_{L_2}, for fully parameterized realizations one has

$$M_{L_2} = \left\|\frac{\partial H}{\partial A}\right\|_2^2 + \left\|\frac{\partial H}{\partial B}\right\|_2^2 + \left\|\frac{\partial H}{\partial C^{\mathcal{T}}}\right\|_2^2 + \left\|\frac{\partial H}{\partial d}\right\|_2^2. \tag{14}$$

The problem of minimizing M_{L_2} within the fully parameterized realization set, denoted with S_H^f, has been solved; see, e.g., [6], [14], and [15]. The corresponding realizations are denoted by R_{L_2}.

For sparse realizations, it can be shown [8] that M_{L_2} can be evaluated with

$$\begin{aligned} M_{L_2} = &\sum_{l=1}^{n}\sum_{k=1}^{n}\phi_{lk}(C \quad \mathbf{0})R_{lk}(C \quad \mathbf{0})^{\mathcal{T}} \\ &+ \sum_{k=1}^{n}[\varphi_k W_o(k,k) + \psi_k W_c(k,k)] + \upsilon \end{aligned} \tag{15}$$

where W_c, W_o are defined in (12), R_{lk} is the solution of the following Lyapunov equation:

$$\begin{aligned} R_{lk} - \begin{pmatrix} A & e_l e_k^{\mathcal{T}} \\ \mathbf{0} & A \end{pmatrix} R_{lk} \begin{pmatrix} A & e_l e_k^{\mathcal{T}} \\ \mathbf{0} & A \end{pmatrix}^{\mathcal{T}} \\ = \begin{pmatrix} \mathbf{0} \\ B \end{pmatrix}\begin{pmatrix} \mathbf{0} \\ B \end{pmatrix}^{\mathcal{T}}, \qquad \forall(l,k) \end{aligned} \tag{16}$$

with e_i denoting the ith elementary (column) vector, and

$$\begin{aligned} \phi_{lk} &= \begin{cases} 0, & \text{if } A(l,k) \text{ is trivial} \\ 1, & \text{if } A(l,k) \text{ is nontrivial} \end{cases} \\ \varphi_k &= \begin{cases} 0, & \text{if } B(k) \text{ is trivial} \\ 1, & \text{if } B(k) \text{ is nontrivial} \end{cases} \\ \psi_k &= \begin{cases} 0, & \text{if } C(k) \text{ is trivial} \\ 1, & \text{if } C(k) \text{ is nontrivial} \end{cases} \\ \upsilon &= \begin{cases} 0, & \text{if } d \text{ is trivial} \\ 1, & \text{if } d \text{ is nontrivial.} \end{cases} \end{aligned} \tag{17}$$

Though for a given realization one can compute this sensitivity measure, how to minimize M_{L_2} within the realization space of sparse realizations, denoted with S_H^s, is very difficult, and as far as we know this is still an open problem.

B. Roundoff Noise Gain

Another FWL effect is due to the rounding of the states in (1). Consider the situation where the rounding operation is performed after a multiplication (RAM). It can be shown [1], [2], [6] that the variance of the output error due to the roundoff is given by

$$\sigma_{\Delta y}^2 = [\mathrm{tr}(W_o Q) + m_{n+1}]\sigma_0^2 \tag{18}$$

where σ_0^2 is a constant, determined by the word length used for representing the states, $Q = \mathrm{diag}(m_1, \ldots, m_k, \ldots, m_n)$ with m_k the number of nontrivial parameters in the row vector kth row of $(A \;\; B)$ for $k = 1, \ldots, n$, and m_{n+1} is the number of nontrivial parameters in the row vector $(C \;\; d)$.

In the actual implementation, all states in (1) have to be maintained in the same dynamic range, which is achieved with the following l_2-scaling [1], [2]:

$$W_c(i,i) = 1, \quad \forall i. \tag{19}$$

Equation (19) defines a subset, denoted $S_H^{l_2}$, of the realization set S_H defined in (3). The optimal roundoff noise realization problem is to find those realizations in $S_H^{l_2}$ that minimize $\sigma_{\Delta y}^2$ or, equivalently, the roundoff noise gain defined as

$$G \triangleq \frac{\sigma_{\Delta y}^2}{\sigma_0^2} = \mathrm{tr}(W_o Q) + m_{n+1}. \tag{20}$$

If the realization is fully parameterized, that is $m_k = n+1$, $\forall k$, one has

$$G = [\mathrm{tr}(W_o) + 1](n+1). \tag{21}$$

The following optimal realization problem:

$$\min_{R \in S_H^f} G = [\mathrm{tr}(W_o) + 1](n+1) \tag{22}$$

was solved in [1] and [2] independently. The corresponding realizations are denoted by R_G.

The optimal realization problem defined over the set of sparse realizations

$$\min_{R \in S_H^s} G \tag{23}$$

just as the optimal sensitivity realization problem, seems very difficult and is still an open problem. In [7], a numerical algorithm was proposed to solve a suboptimal problem.

It is interesting to note that for all l_2-scaled realizations in S_H^f, one has

$$\bar{M}_{L_1/L_2} = G \tag{24}$$

where $\bar{M}_{L_1/L_2}$ and G are defined in (10) and (21), respectively. Equation (24) implies that any minimal roundoff realization R_G is also a minimum sensitivity realization in this subset. Therefore, R_G should yield a very good performance in terms of reducing FWL effects. R_G, however, possesses $n^2 + 2n + 1$ nontrivial parameters, which limits its applications in digital filter implementation. In what follows, based on a well-known continuous-time realization, we will develop a new structure which is sparse and yields a performance comparable to the fully parameterized R_G.

III. The JSS Structure

In this section, we introduce the JSS-structure used in [13] for analog filter implementation. A new stability property of this structure is revealed. Based on the JSS-structure, a new structure will be developed for digital filter implementation in the next section.

It is well known that the continuous-time counterpart of $H(z)$, denoted with $F(s)$, can be obtained with the *bilinear* *transformation* $s = (z-1/z+1)$, or $z = (1+s/1-s)$, such that

$$F(s) = H(z)|_{z=\frac{1+s}{1-s}}. \tag{25}$$

Clearly, *$H(z)$ has all its poles in $|z|<1$ if and only if $F(s)$ has all its poles in* $\mathrm{Re}(s)<0$, where $\mathrm{Re}(x)$ denotes the real part of any complex number x.

Let $(A,B,C,d) \in S_H$ and let $F(s)$ be defined from $H(z)$ as in (25). It can be shown that there exists a realization of $F(s)$ given by

$$\begin{aligned} \Phi &= (I+A)^{-1}(A-I), & K &= \sqrt{2}(I+A)^{-1}B \\ L &= \sqrt{2}C(I+A)^{-1}, & D &= d - C(I+A)^{-1}B \end{aligned} \tag{26}$$

such that

$$F(s) = D + L(sI-\Phi)^{-1}K. \tag{27}$$

The continuous-time gramian pair (P,Q) of (Φ,K,L,D) satisfies the following equations:

$$\Phi P + P\Phi^{\mathcal{T}} = -KK^{\mathcal{T}}, \qquad \Phi^{\mathcal{T}}Q + Q\Phi = -L^{\mathcal{T}}L \tag{28}$$

It is interesting to note that

$$P = W_c, \qquad Q = W_o \tag{29}$$

where (W_c, W_o), as defined before, is the gramian pair of (A,B,C,d).

Like $H(z), F(s)$ also has an infinite number of realizations. Consider the following continuous-time realization $(\Phi_{\text{in}}, K_{\text{in}}, L_{\text{in}}, D)$ of $F(s)$:

$$\Phi_{\text{in}} = \begin{pmatrix} 0 & \alpha_1 & 0 & 0 & \cdots & 0 & 0 \\ -\alpha_1 & 0 & \alpha_2 & 0 & \cdots & 0 & 0 \\ 0 & -\alpha_2 & 0 & \alpha_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \alpha_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & -\alpha_{n-1} & -\alpha_n \end{pmatrix}$$

$$K_{\text{in}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \sqrt{2\alpha_n} \end{pmatrix} \tag{30}$$

where α_k are real $\forall k$ and L_{in} has no special structure.

This state-space structure, proposed by Johns *et al.*[4] for implementing a given continuous-time transfer function $F(s)$, was shown to have a very good behavior against FWL effects [13] and leads to the so-called *orthonormal ladder filters*. We note that $P = I$ if (Φ, K) is replaced with $(\Phi_{\text{in}}, K_{\text{in}})$ in (28), and with (29) one can see that the JSS-structure is inherently L_2-scaled [13] and its discrete-time counterpart $(A_{\text{in}}, B_{\text{in}}, C_{\text{in}}, d)$, called *DJSS-structure*, is automatically l_2 scaled. For other nice properties of the JSS-structure, we refer to [13].

Before going to the issue of computing the JSS-structure, let us reveal a new property of this structure.

A. A New Property of the JSS-Structure

Let λ be any eigenvalue of Φ_{in}. We show that Φ_{in} is (strictly) stable, that is the real part of λ is negative, *iff* $\alpha_k \neq 0, k = 1,2,\ldots,n-1$ with $\alpha_n > 0$. To prove this claim, one needs the following lemma.

Lemma 1: Let $M \in \mathcal{R}^{n\times n}$ and $M = M_s + M_{sk}$ with $M_s = M_s^{\mathcal{T}}$ and $M_{sk} = -M_{sk}^{\mathcal{T}}$. Denote $\{\lambda_k\} = \lambda(M)$ and as the eigenvalue set of M. Then

$$\mathrm{Re}(\lambda_k) = \frac{\bar{x}_k^{\mathcal{H}} M_s \bar{x}_k}{\bar{x}_k^{\mathcal{H}} \bar{x}_k} \quad \forall k \tag{31}$$

where $\bar{x}_k$ is an eigenvector corresponding to λ_k and $\mathcal{H}$ denotes the transpose-conjugate operator.

Proof: Let $\bar{x}_k$ be an eigenvector of λ_k, that is $M\bar{x}_k = \lambda_k \bar{x}_k$. Clearly, $\bar{x}_k^{\mathcal{H}} M \bar{x}_k = \lambda_k \bar{x}_k^{\mathcal{H}} \bar{x}_k$. With the decomposition $M = M_s + M_{sk}$, one has $\bar{x}_k^{\mathcal{H}} M_s \bar{x}_k = \mathrm{Re}(\lambda_k)\bar{x}_k^{\mathcal{H}}\bar{x}_k$, which leads to (31). □

Theorem 1: Let Φ_{in} be given by (30). Denote $\{\lambda_k\} = \lambda(\Phi_{\text{in}})$, then Φ_{in} is strictly stable *iff* $\alpha_k \neq 0$ for $k = 1,2,\ldots,n-1$ and $\alpha_n > 0$.

Proof: First of all, let $M = \Phi_{\text{in}}$ in Lemma 1. Noting all the elements of the symmetric matrix M_s are zero except the last diagonal element, which is $-\alpha_n$, it follows from (31) that

$$\mathrm{Re}(\lambda) = \frac{-\alpha_n |x_n|^2}{x^{\mathcal{H}} x} \tag{32}$$

where $x = (x_1 x_2 \ldots x_k x_{n-1} x_n)^{\mathcal{T}}$ is an eigenvector corresponding to $\lambda \in \lambda(\Phi_{\text{in}})$. Clearly, the sufficiency part of the theorem can be proved if one can show that $\alpha_k \neq 0$ for $k = 1,2,\ldots,n-1$ and $\alpha_n > 0$ implies $x_n \neq 0$. In fact, it follows from $\Phi_{\text{in}} x = \lambda x$ that

$$\begin{aligned} \alpha_1 x_2 &= \lambda x_1 \\ -\alpha_k x_k + \alpha_{k+1} x_{k+2} &= \lambda x_{k+1}, \quad k = 1,2,\ldots,n-2 \\ -\alpha_{n-1} x_{n-1} - \alpha_n x_n &= \lambda x_n. \end{aligned} \tag{33}$$

It is easy to see that with all $\alpha_k \neq 0$, $x_n = 0$ means $x = 0$, which contradicts the assumption that x is an eigenvector corresponding to $\lambda \in \lambda(\Phi_{\text{in}})$.

Now, assume that Φ_{in} is strictly stable. The necessity part of the theorem is proved if one can show that no α_k is zero and that $\alpha_n > 0$.

It is easy to see from (32) that α_n is necessarily positive and from (30) that $\alpha_1 \neq 0$ if Φ_{in} is stable. Assume $\alpha_k = 0$ for all $k = 2,\ldots,n-1$. Then $\Phi_{\text{in}} = \left(\begin{smallmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{smallmatrix}\right)$, where $\Phi_1 \in \mathcal{R}^{(k+1)\times(k+1)}$ and $\Phi_2 \in \mathcal{R}^{(n-k-1)\times(n-k-1)}$ and k is any number between 2 and $n-1$. Noting the fact that Φ_1 is a skew-symmetric matrix, it follows from (31) that all eigenvalues of Φ_1 are imaginary. This means that Φ_{in} is marginally stable, which is against the assumption. □

B. Computing the JSS-Structure

Now, let us consider the problem of computing the JSS-structure for a given $H(z)$. First of all, suppose $F(s) = N(s)/D(s)$

[4] In the sequel, we will refer to this structure as the *JSS-structure*.

is obtained by (25). Let the denominator be decomposed as $D(s) = E(s) + O(s)$, where $E(s)$ and $O(s)$ are the even- and odd-power terms of $D(s)$, respectively. Define $Z(s) = O(s)/E(s)$, if the order of $D(s)$ is odd, or $Z(s) = E(s)/O(s)$, if the order of $D(s)$ is even. Then we have the following continued fraction form:

$$Z(s) = r_n s + \cfrac{1}{r_{n-1}s + \cfrac{1}{r_{n-2}s + \cfrac{1}{r_{n-3}s + \cfrac{1}{\ddots \, \cfrac{1}{r_1 s}}}}}. \tag{34}$$

The *Hurwitz* test (see, e.g., [16] and [17]) says that $D(s)$ is (strictly) stable *iff* $r_k > 0, \forall k$. It was shown in [13] that (30) can be obtained by diagonally transforming the realization of an orthonormal ladder filter derived from a singly terminated LC ladder filter with a unit valued resistor, where the states correspond to the capacitor voltages and inductor currents, and r_k defined in (34) is either the capacitor or inductor value for all k. The corresponding diagonal similarity transformation is $T = \mathrm{Diag}(\sqrt{(1/2r_1)}, \sqrt{(1/2r_2)}, \ldots, \sqrt{(1/2r_n)})$, which leads to

$$\alpha_k = \sqrt{\frac{1}{r_k r_{k+1}}}, \qquad k = 1, 2, \ldots, n-1$$
$$\alpha_n = \sqrt{\frac{1}{r_n}}. \tag{35}$$

Φ_{in} and K_{in} can directly be defined from the α_k obtained above.

Let (Φ_c, K_c, L_c, d) be any realization of $F(s)$, say the controllable realization. With Φ_{in} and K_{in}, one can find the similarity transformation T_{in} that transforms (Φ_c, K_c, L_c, d) into the JSS-structure and hence also L_{in} with $L_{\text{in}} = L_c T_{\text{in}}$.

IV. The LGS Structure

It follows from (26) that the DJSS-structure is given by

$$\begin{aligned} A_{\text{in}} &= (I + \Phi_{\text{in}})(I - \Phi_{\text{in}})^{-1} \\ B_{\text{in}} &= \frac{\sqrt{2}}{2}(I + A_{\text{in}}) = K_{\text{in}} \\ C_{\text{in}} &= \frac{\sqrt{2}}{2} L_{\text{in}}(I + A_{\text{in}}) \\ d &= D + C_{\text{in}}(I + A_{\text{in}})^{-1} B_{\text{in}}. \end{aligned} \tag{36}$$

It can be shown that the expression for computing the sensitivity measure for *analog* filters has the same form as that given in (10) [20]. Let R_k^s be a continuous-time realization and R_k the corresponding digital realization, for $k = 1, 2$. Equation (29) implies that the JSS-structure and the DJSS-structure have exactly the same sensitivity behavior in the sense that if R_1^s has a smaller sensitivity measure than R_2^s, then the same holds for R_1 vis-à-vis R_2. Noting the fact that the JSS-structure has a very small sensitivity measure [13], it follows from (24) that the the DJSS-structure has a very small sensitivity as well as a very small roundoff noise gain.

One would therefore suggest using this DJSS-structure for digital filter implementation. However, it has been noted that, unlike $\Phi_{\text{in}}, A_{\text{in}}$ obtained with (36) is fully parameterized due to the matrix inversion involved in (36). A direct implementation of A_{in} leads to n^2 multiplications for computing $A_{\text{in}}x(t)$ in (1). In this section, we show that A_{in} can be factorized into a series of simple (sparse) matrices. Using this factorization, computation for $A_{\text{in}}x(t)$ can be much simplified.

Denote by $U(i, j, x)$ the unit matrix except that its (i, j)th element is $x, \forall(i, j)$, and let $T_1 = U(2, 2, \gamma_1)U(2, 1, -\alpha_1)$ with $\gamma_1 = 1/(1 + \alpha_1^2)$. Now we note that

$$\begin{aligned} I - \Phi_{\text{in}} &= T_1^{-1} T_1 (I - \Phi_{\text{in}}) \\ &= T_1^{-1} \begin{pmatrix} 1 & -\alpha_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \beta_2 & 0 & \cdots & 0 & 0 \\ 0 & \alpha_2 & 1 & -\alpha_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -\alpha_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & \alpha_{n-1} & 1 + \alpha_n \end{pmatrix} \end{aligned}$$

where $\beta_2 = -\alpha_2/(1 + \alpha_1^2)$. Repeating this procedure, one can show that

$$I - \Phi_{\text{in}} = T_1^{-1} T_2^{-1} \ldots T_k^{-1} \ldots T_{n-1}^{-1} \Psi \tag{37}$$

where Ψ is the unit matrix except $\Psi(k, k+1) = \beta_k, \forall k$ with β_k given by the following recursive equations:

$$\begin{aligned} \beta_{k+1} &= -\frac{\alpha_{k+1}}{s_k}, \quad s_k = 1 - \alpha_k \beta_k, \qquad k = 1, \ldots, n-2 \\ \beta_1 &= -\alpha_1, \qquad s_{n-1} = 1 + \alpha_n + \alpha_{n-1}\beta_{n-1} \end{aligned} \tag{38}$$

and

$$T_k = U(k+1, k+1, \gamma_k) U(k+1, k, -\alpha_k), \qquad k = 1, 2, \ldots, n-1 \tag{39}$$

where $\gamma_k = s_k^{-1}$ with s_k defined in (38).

Noting that

$$\begin{aligned} \Psi^{-1} = {} & U(1, 2, -\beta_1) \ldots U(k+1, k+1, -\beta_k) \\ & \ldots U(n-1, n, -\beta_{n-1}) \end{aligned} \tag{40}$$

one has the following:

$$\begin{aligned} A_{\text{in}} &= (I + \Phi_{\text{in}})(I - \Phi_{\text{in}})^{-1} \\ &= (I + \Phi_{\text{in}}) U(1, 2, -\beta_1) \ldots U(n-1, n, -\beta_{n-1}) \\ &\quad \times U(n, n, \gamma_{n-1}) U(n, n-1, -\alpha_{n-1}) \\ &\quad \ldots U(2, 2, \gamma_1) U(2, 1, -\alpha_1) \end{aligned} \tag{41}$$

which shows that A_{in} is factorized with $N \triangleq 1 + 3(n-1)$ sparse matrices

$$A_{\text{in}} = A^{(N)} A^{(N-1)} \ldots A^{(2)} A^{(1)} \triangleq \prod_{k=1}^{N} A^{(k)} \tag{42}$$

where $A^{(N)} = I + \Phi_{\text{in}}$. Clearly, A_{in} is parameterized with $3(n-1)$ different nontrivial parameters $\{\alpha_k, \beta_k, \gamma_k\}$ in (41) or (42) (noting $\beta_1 = \alpha_1$).

Taking advantage of this factorization, one can see that with the DJSS-structure (1) can be rewritten as

$$\begin{aligned} x^{(k)}(t) &= A^{(k)} x^{(k-1)}(t), \quad x^{(0)}(t) = x(t), \\ &\qquad k = 1, 2, \ldots, N \\ x(t+1) &= x^{(N)}(t) + B_{\text{in}} u(t) \\ y(t) &= C_{\text{in}} x(t) + d u(t). \end{aligned} \tag{43}$$

Note that $x^{(N)}(t) = A_{\text{in}}x(t)$. Simple calculations show that computing $x^{(N)}(t)$ with (43) requires *only* $5n-4$ multiplications and $4n-3$ additions, rather than n^2 and $n(n-1)$, respectively, as required in the DJSS-structure. This is a significant reduction of computational complexity.

For convenience, (43) is referred to as the *LGS-structure*. Before turning to the next section, we point out that the LGS-structure, though equivalent to the state-space DJSS-structure, does *not* belong to the state-space realization set. Equation (43) yields a different class of implementation structures.

V. Performance Analysis of the LGS Structure

In this section, we analyze the performance of the structure proposed in the previous section in terms of sensitivity measure and roundoff noise.

A. Sensitivity Analysis

First of all, it follows from (41) and (42) that

$$A^{(2k)}A^{(2k-1)} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & \gamma_k & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix} \times \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & -\alpha_k & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}$$

$$A^{(N-k)} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & -\beta_k & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix} \tag{44}$$

for $k = 1, 2, \ldots, n-1$, and $A^{(N)} = I + \Phi_{\text{in}}$. Therefore, the LGS-structure (43) is parameterized with $B_{\text{in}}, C_{\text{in}}, d$, and three parameter sets $\{\alpha_k\}, \{\beta_k\}, \{\gamma_k\}$. Noting that B_{in} and C_{in} are generally fully parameterized with nontrivial parameters, one can see that the LGS-structure has $5n-1$ parameters to be implemented with $\beta_1 = \alpha_1$. The transfer function sensitivities $\partial H/\partial B_{\text{in}}, \partial H/\partial C_{\text{in}}$, and $\partial H/\partial d$ can be obtained with (7)–(8).

Now, let us consider $\partial H/\partial p, \forall p \in S_p$, where S_p denotes the parameter set containing the parameters that define A_{in}: α_k for $k = 1, \ldots, n, \beta_k$ for $k = 2, \ldots, n-1$, and γ_k for $k = 1, \ldots, n-1$. Denote

$$A_{\text{in}} = A^{(N)} \prod_{i=N-(n-1)}^{N-1} A^{(i)} \prod_{i=1}^{2(n-1)} A^{(i)} \triangleq A^{(N)}A_2A_1.$$

It follows from $H(z) = d + C_{\text{in}}(zIA_{\text{in}})^{-1}B_{\text{in}}$ that

$$\frac{\partial H}{\partial p} = C_{\text{in}}(zI - A_{\text{in}})^{-1}\frac{\partial A_{\text{in}}}{\partial p}(zI - A_{\text{in}})^{-1}B_{\text{in}} \tag{45}$$

where

$$\frac{\partial A_{\text{in}}}{\partial p} = \frac{\partial A^{(N)}}{\partial p}A_2A_1 + A^{(N)}\frac{\partial A_2}{\partial p}A_1 + A^{(N)}A_2\frac{\partial A_1}{\partial p}. \tag{46}$$

Noting $\beta_1 = \alpha_1$, one can show (47), as shown at the bottom of the page. Similarly, we have (48) and (49), shown at the bottom of the next page. Noting that B_{in} and C_{in} are generally fully parameterized, and assuming that d is nontrivial, the L_2-sensitivity measure for the proposed structure is

$$\begin{aligned} M_{\text{LGS}} &= \sum_{p\in S_p}\left\|\frac{\partial H}{\partial p}\right\|_2^2 + \left\|\frac{\partial H}{\partial B_{\text{in}}}\right\|_2^2 + \left\|\frac{\partial H}{\partial C_{\text{in}}}\right\|_2^2 + \left\|\frac{\partial H}{\partial d}\right\|_2^2 \\ &= \sum_{p\in S_p}\left\|\frac{\partial H}{\partial p}\right\|_2^2 + \operatorname{tr}\left(W_o^{(\text{in})}\right) + (n+1) \end{aligned} \tag{50}$$

where $W_o^{(\text{in})}$ is the observability gramian of $(A_{\text{in}}, B_{\text{in}}, C_{\text{in}})$ for which $W_c^{(\text{in})} = I$.

Noting the following equality:

$$M_1^{-1}M_2M_3^{-1} = (I \quad \mathbf{0})\begin{pmatrix} M_1 & -M_2 \\ \mathbf{0} & M_3 \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{0} \\ I \end{pmatrix}$$

it follows from (45) that

$$\begin{aligned} \frac{\partial H}{\partial p} &= C_{\text{in}}(I_n \quad \mathbf{0})\left(zI_{2n} - \begin{pmatrix} A_{\text{in}} & \frac{\partial A_{\text{in}}}{\partial p} \\ \mathbf{0} & A_{\text{in}} \end{pmatrix}\right)^{-1}\begin{pmatrix} \mathbf{0} \\ I_n \end{pmatrix}B_{\text{in}} \\ &\triangleq \bar{C}_{\text{in}}(zI_{2n} - \bar{A}_{\text{in}})^{-1}\bar{B}_{\text{in}} \end{aligned} \tag{51}$$

where I_m denotes the unit matrix of order m and $\mathbf{0}$s are zero matrices of proper dimension. It is easy to show that

$$\left\|\frac{\partial H}{\partial p}\right\|_2^2 = \bar{C}_{\text{in}}\bar{W}_c\bar{C}_{\text{in}}^{\mathcal{T}} = \bar{B}_{\text{in}}^{\mathcal{T}}\bar{W}_o\bar{B}_{\text{in}} \tag{52}$$

where $(\bar{W}_c, \bar{W}_o)$ is the gramian pair of the realization $(\bar{A}_{\text{in}}, \bar{B}_{\text{in}}, \bar{C}_{\text{in}})$ defined in (31).

$$\begin{aligned} \frac{\partial A^{(N)}}{\partial \alpha_k} &= \begin{cases} e_ke_{k+1}^{\mathcal{T}} - e_{k+1}e_k^{\mathcal{T}} & k = 1, 2, \ldots, n-1 \\ -e_ne_n^{\mathcal{T}} & k = n \end{cases} \\ \frac{\partial A_1}{\partial \alpha_k} &= \begin{cases} -\prod_{i=2}^{2(n-1)} A^{(i)}e_2e_1^{\mathcal{T}} & k = 1 \\ -\prod_{i=2k}^{2(n-1)} A^{(i)}e_{k+1}e_k^{\mathcal{T}}\prod_{i=1}^{2(k-1)} A^{(i)} & 2 \le k \le n-1 \\ 0 & k = n \end{cases} \\ \frac{\partial A_2}{\partial \alpha_k} &= \begin{cases} -e_1e_2^{\mathcal{T}}\prod_{i=N-(n-1)}^{(N-2)} A^{(i)} & k = 1 \\ 0 & k \ne 1 \end{cases} \end{aligned} \tag{47}$$

Therefore, $\sum_{p\in S_p} \|(\partial H/\partial p)\|_2^2$ and hence M_{LGS} can be computed using (52) with (46)–(50).

B. Roundoff Noise Analysis

In an actual implementation, all intermediate variables should be maintained to a certain wordlength, say B_s bits. Therefore, the product of any nontrivial parameter and an intermediate variable has to be rounded to B_s bits. So, the actual model of the LGS-structure (43) with all intermediate variables rounded is

$$\begin{aligned} x^*(t+1) &= Q\left[A^{(N)}Q\left[\cdots Q\left[A^{(2)}Q\left[A^{(1)}x^*(t)\right]\right]\cdots\right]\right] \\ &\quad + Q[B_{\mathrm{in}}u(t)] \\ y^*(t) &= Q[C_{\mathrm{in}}x^*(t)] + Q[du(t)] \end{aligned} \tag{53}$$

where $Q[Mv]$ is the quantizer that rounds all products occurring in the multiplication Mv into B_s bits.

Denoting

$$\begin{aligned} Z_1(t) &\triangleq Q\left[A^{(2(n-1))}\cdots Q\left[A^{(1)}x^*(t)\right]\cdots\right] \\ Z_2(t) &\triangleq Q\left[A^{(N-1)}\cdots Q\left[A^{(N-(n-1))}Z_1(t)\right]\cdots\right] \end{aligned} \tag{54}$$

it then follows from (53) that

$$\begin{aligned} x^*(t+1) &= A^{(N)}Z_2(t) + B_{\mathrm{in}}u(t) + e_3(t) \\ y^*(t) &= C_{\mathrm{in}}x^*(t) + du(t) + \eta(t) \end{aligned} \tag{55}$$

where $e_3(t)$ and $\eta(t)$ are the roundoff noises in $Q[A^{(N)}Z_2(t)] + Q[B_{\mathrm{in}}u(t)]$ and $Q[C_{\mathrm{in}}x^*(t)] + Q[du(t)]$, respectively. Their covariance is given by

$$\begin{aligned} R_{e_3} &= E\left[e_3(t)e_3^{\mathcal{T}}(t)\right] = \mathrm{diag}(2,3,3,\ldots,3,3)\sigma_0^2 \triangleq D_3\sigma_0^2 \\ R_\eta &= e[\eta(t)\eta^{\mathcal{T}}(t)] = (n+1)\sigma_0^2 \end{aligned} \tag{56}$$

with σ_0^2 a constant depending on B_s.

It is easy to see that

$$Z_2(t) = A_2Z_1(t) + e_2(t) \tag{57}$$

where $e_2(t)$ is the quantization error of $Z_2(t)$ given by (54). Noting the structure specified in (44), one can show that

$$R_{e_2} = E\left[e_2(t)e_2^{\mathcal{T}}(t)\right] = \mathrm{diag}(1,1,\ldots,1,0)\sigma_0^2 \triangleq D_2\sigma_0^2. \tag{58}$$

Similarly

$$Z_1(t) = A_1x^*(t) + e_1(t) \tag{59}$$

where $e_1(t)$ is the roundoff noise, satisfying

$$\begin{aligned} R_{e_1} &= E\left[e_1(t)e_1^{\mathcal{T}}(t)\right] \\ &= \mathrm{diag}\left(0, 1+\gamma_1^2, 1+\gamma_2^2, \ldots, 1+\gamma_{n-1}^2\right)\sigma_0^2 \triangleq D_1\sigma_0^2. \end{aligned} \tag{60}$$

Combining (55), (57) and (59), one has

$$\begin{aligned} x^*(t+1) &= A_{\mathrm{in}}x^*(t) + B_{\mathrm{in}}u(t) + \epsilon(t) \\ y^*(t) &= C_{\mathrm{in}}x^*(t) + du(t) + \eta(t) \end{aligned} \tag{61}$$

where

$$\epsilon(t) = A^{(N)}A_2e_1(t) + A^{(N)}e_2(t) + e_3(t). \tag{62}$$

Denoting $\Delta y(t) = y^*(t) - y(t)$ and $E(t) = x^*(t) - x(t)$, it follows from (43) and (61) that

$$\begin{aligned} E(t+1) &= A_{\mathrm{in}}E(t) + \epsilon(t) \\ \Delta y(t) &= C_{\mathrm{in}}E(t) + \eta(t). \end{aligned} \tag{63}$$

Keeping in mind the assumption that all roundoff noises are independent and white, it turns out from (63) that

$$\sigma_{\Delta y}^2 \triangleq E[(\Delta y(t))^2] = C_{\mathrm{in}}PC_{\mathrm{in}}^{\mathcal{T}} + R_\eta \tag{64}$$

where

$$\begin{aligned} P &= A_{\mathrm{in}}PA_{\mathrm{in}}^{\mathcal{T}} + R_\epsilon \\ R_\epsilon &\triangleq E[\epsilon(t)\epsilon^2(t)] \\ &= A^{(N)}\left[A_2R_{e_1}A_2^{\mathcal{T}} + R_{e_2}\right]\left(A^{(N)}\right)^{\mathcal{T}} + R_{e_3} \\ &= \left\{A^{(N)}\left(A_2D_1A_2^{\mathcal{T}} + D_2\right)\left(A^{(N)}\right)^{\mathcal{T}} + D_3\right\}\sigma_0^2 \\ &\triangleq R_0\sigma_0^2. \end{aligned} \tag{65}$$

It can then be shown that

$$\sigma_{\Delta y}^2 = \mathrm{tr}\left(W_o^{(\mathrm{in})}R_\epsilon\right) + R_\eta. \tag{66}$$

$$\begin{aligned} &\frac{\partial A^{(N)}}{\partial \beta_k} = 0, \quad \frac{\partial A_1}{\partial \beta_k} = 0 \\ &\frac{\partial A_2}{\partial \beta_k} = \begin{cases} -\prod_{i=N+1-k}^{(N-1)} A^{(i)}e_ke_{k+1}^{\mathcal{T}}\prod_{i=1}^{(N-1-k)} A^{(i)} & 2 \le k \le n-2 \\ -\prod_{i=2(n-1)+2}^{(N-1)} A^{(i)}e_{n-1}e_n^{\mathcal{T}} & k = n-1 \end{cases} \end{aligned} \tag{48}$$

$$\begin{aligned} &\frac{\partial A^{(N)}}{\partial \gamma_k} = 0, \quad \frac{\partial A_2}{\partial \gamma_k} = 0 \\ &\frac{\partial A_1}{\partial \gamma_k} = \begin{cases} \prod_{i=2k+1}^{2(n-1)} A^{(i)}e_{k+1}e_{k+1}^{\mathcal{T}}\prod_{i=1}^{2k-1} A^{(i)} & 1 \le k \le n-2 \\ e_ne_n^{\mathcal{T}}\prod_{i=1}^{2(n-1)-1} A^{(i)} & k = n-1. \end{cases} \end{aligned} \tag{49}$$

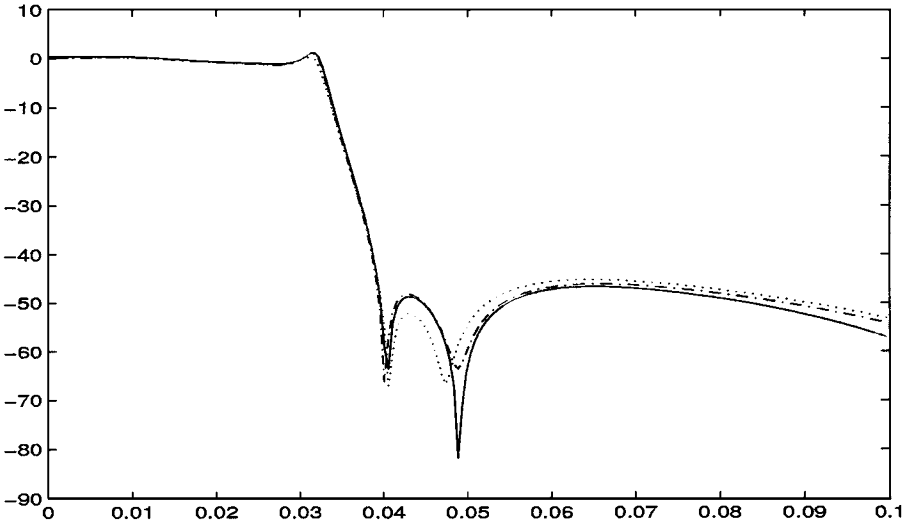


Fig. 1. Magnitude responses—solid line: ideal; dotted line: R_G with $B_c = 10$; and dashdot line: $R_{\rm LGS}$ with $B_c = 10$.

TABLE I

Structure	R_c	R_{L_1/L_2}	R_{L_2}	R_G	R_{DJSS}	R_{LGS}
M_{L_2}	1.9072×10^{10}	184.9349	163.9790	168.0311	221.1138	270.8383
G	-	-	-	16.3306	21.6053	18.2649
N_p	13	49	49	49	49	29

The roundoff noise gain for the proposed structure (43) is therefore

$$G = \frac{\sigma_{\Delta y}^2}{\sigma_0^2} = \mathrm{tr}\left(W_o^{(\mathrm{in})} R_0\right) + (n+1) \tag{67}$$

where R_0 is defined in (65).

VI. A Numerical Example and Simulation Results

In this section, we present a design example to illustrate the performance of the proposed structure and compare it with five other structures: the controllable realization $R_c, R_{\rm DJSS}$, and fully parameterized optimal realizations $R_{L_1/L_2}, R_{L_2}$ and R_G.

Example: This is a sixth-order *narrow* band low-pass filter used in [6]. The normalized passband frequency f_p is 0.031 25, the stopband frequency f_s is 0.039 062 5, and the passband ripple is 1 dB. The attenuation in the stopband is greater than 46.68 dB. The corresponding magnitude response is depicted in Fig. 1 with solid line.

Table I shows the statistics on the sensitivity measure M_{L_2}, the roundoff noise gain[5] G, and the number of nontrivial parameters N_p involved in implementation for the six structures.

In this example, one can see that R_c is very sensitive to the parameter errors. This is due to the very narrow bandwidth of the filter. The other five structures have a much smaller sensitivity measure. The proposed $R_{\rm LGS}$ yields a slightly larger M_{L_2} value than R_G and $R_{\rm DJSS}$. In fact, all five structures yield almost the same frequency response for a given parameter perturbation. Let us truncate all nontrivial parameters of a structure into B_c bits in their fractional part. The magnitude responses of R_G and $R_{\rm LGS}$, both truncated with $B_c = 10$, are depicted in Fig. 1 with the dotted and dashdot lines, respectively. It clearly shows that both structures yield almost the same response, which is very close to the desired one, especially in the passband. To achieve the same response, R_c requires at least $B_c = 22$ bits. In terms of roundoff noise gain, $R_{\rm LGS}$ is quite close to R_G and better than $R_{\rm DJSS}$.

Comment 6.1: The relationship between the G values of R_G and $R_{\rm LGS}$ depends on examples. In fact, $R_{\rm LGS}$ may yield a smaller roundoff noise gain than R_G. This is due to the fact that R_G is optimal in the set of the *fully* parameterized *state-space realizations,* while $R_{\rm LGS}$ does *not* belong to the state-space realizations S_H and in $R_{\rm LGS}$ there is a smaller number of rounding operations. Other examples also show that the M_{L_2} value for $R_{\rm LGS}$ can be smaller than that for $R_{\rm DJSS}$.

VII. Conclusion

In this paper, the digital filter structure problem in FWL implementation has been discussed. Our contribution is three-fold. Firstly, a new stability property of the JSS-structure has been revealed. Secondly, based on the DJSS-structure a new structure has been developed, which is sparse and yields a very nice performance. The performance of this proposed structure has been analyzed by deriving the corresponding expressions for sensitivity measure and roundoff noise gain. A design example has been given, with which it is shown that the proposed structure is not only simpler than the DJSS-structure but also generally yields a better performance which is very close to that of the fully parameterized optimal roundoff noise realizations.

[5] The roundoff noise gain G is defined with l_2-scaling. It is meaningless to present the G value for $R_c, R_{L_1/L_2}$ and R_{L_2} since they are not l_2-scaled.

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, Sept. 1976.

[2] 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.

[3] D. V. B. Rao, “Analysis of coefficient quantization errors in state-space digital filters,” *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-34, pp. 131–139, Feb. 1986.

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

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

[6] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects*, ser. Communication and Control Engineering Series. London, U.K.: Springer Verlag, 1993.

[7] G. Amit and U. Shaked, “Small roundoff realization of fixed-point digital filters and controllers,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 880–891, Jun. 1988.

[8] C. Xiao, “Improved L_2-sensitivity for state-space digital system,” *IEEE Trans. Signal Processing*, vol. 45, pp. 837–840, Apr. 1997.

[9] G. Li, “On the structure of digital controllers with finite word length consideration,” *IEEE Trans. Automat. Contr.*, vol. 43, pp. 689–693, May 1998.

[10] 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. 1989.

[11] G. Li, B. D. O. Anderson, M. Gevers, and J. E. Perkins, “Optimal FWL design of state-space digital systems with weighted sensitivity minimization and sparseness consideration,” *IEEE Trans. Circuits Syst.*, vol. CAS-39, pp. 365–377, May 1992.

[12] 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, 1984.

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

[14] G. Li, “Two dimensional system optimal realizations with L_2-sensitivity minimization,” *IEEE Trans. Signal Processing*, vol. 46, pp. 809–693, Mar. 1998.

[15] W. Y. Yan and J. B. Moore, “On L^2-sensitivity minimization of linear state-space systems,” *IEEE Trans. Circuits Syst. I*, vol. 39, pp. 641–648, Aug. 1992.

[16] D. S. Humpherys, *The Analysis, Design, and Synthesis of Electrical Filters*. Englewood Cliffs, NJ: Prentice-Hall, 1970.

[17] D. E. Johnson, *Introduction to Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1976.

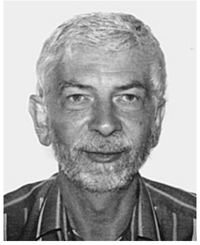
[18] J. B. Knowles and E. M. Olcayto, “Coefficient accuracy and digital filter response,” *IEEE Trans. Circuits Theory*, vol. CT-15, pp. 31–41, Mar. 1968.

[19] R. E. Crochiere, “A new statistical approach to the coefficient word length problem for digital filters,” *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 190–196, Mar. 1975.

[20] W. J. Lutz and S. L. Hakimi, “Design of multi-input multi-output systems with minimum sensitivity,” *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1114–1122, Sept. 1988.

Gang Li (M’93) received the B. Engr. degree in electrical engineering from Beijing Institute of Technology, Beijing, China, in 1982, and the M.Engr. and the Ph.D. degrees both from Louvain University, Belgium, in 1988 and 1990, respectively.

He was with the Control Group at Louvain University as a Postdoctoral Researcher until April 1992. Since May 1992, he has been with the School of EEE, Nanyang Technological University. His research interests include digital filter and controller design, numerical problems in estimation and control applications, adaptive signal processing, signal transforms, and speech modeling/coding. Professor Li is a coauthor (with M. Gevers) of the *Parametrizations in Control, Estimation, and Adaptive Filtering Problems: Accuracy Aspects* (London, U.K.: Springer-Verlag, 1993).



Michel Gevers (M’66–SM’86–F’90) received the E.E. degree from the Université Catholique de Louvain, Belgium, in 1968, and the Ph.D. degree from Stanford University, Stanford, CA, in 1972.

He is now Professor and President of CESAME (Center of Engineering Systems and Applied Mechanics) at the Université Catholique de Louvain in Louvain la Neuve, Belgium. He is the coordinator of the Belgian Inter University Pole on Modeling, Identification, Simulation and Control of Complex Systems, funded by the Federal Ministry of Science. His present research interests include system identification and its interconnection with robust control design. He has published 170 papers and conference papers. He is coauthor (with R. R. Bitmead and V. Wertz) of *Adaptive Optimal Control—The Thinking Man's GPC* (Englewood Cliffs, NJ: Prentice Hall, 1990) and (with G. Li) of *Parametrizations in Control, Estimation and Filtering Problems: Acuracy Aspects* (New York: Springer-Verlag, 1993).

Dr. Gevers is a Distinguished Member of the IEEE Control Systems Society. He is President of the European Union Control Association (EUCA), a member of the Board of Governors of the IEEE Control Systems Society, and Chairman of the International Committee of this Society, and Vice-President of the IEEE Control Systems Society.

Youxian Sun received the Diploma from the Department of Chemical Engineering, Zhejiang University, China, in 1964.

He joined the Department of Chemical Engineering, Zhejiang University, in 1964. He is now Professor and Dean of the Institute of Control Science and Engineering, Zhejiang University. His research interests include robust control, nonlinear control, optimal control, and system identification.

Professor Sun is a Fellow of the Chinese Academy of Engineering.