Roundoff Noise Minimization of State-Space Digital Filters Using Separate and Joint Error Feedback/Coordinate Transformation Optimization

Takao Hinamoto, *Fellow, IEEE,* Hiroaki Ohnishi, and Wu-Sheng Lu, *Fellow, IEEE*

***Abstract*—This paper investigates the problem of minimizing roundoff noise under l_2-norm dynamic-range scaling constraints in state-space digital filters by means of error feedback as well as joint error feedback/coordinate transformation optimization. First, several techniques for the determination of optimal full-scale, diagonal, and scalar error-feedback matrices for a given state-space digital filter are proposed, where three realization schemes, namely, the general state-space realization, input-balanced realization, and optimal realization in the sense of Hwang–Mullis–Roberts are examined. Furthermore, an iterative approach is developed for jointly optimizing a scalar error-feedback matrix and a coordinate transformation matrix so as to minimize the roundoff noise subject to the l_2-norm dynamic-range scaling constraints. The proposed method may be regarded as an alternative, but much simpler and more general, approach to Hwang's method for synthesizing the optimal filter structure with minimum roundoff noise. A case study is included to illustrate the utility of the proposed techniques.**

***Index Terms*—Optimal coordinate transformation, optimal error feedback, roundoff noise minimization, scaling constraints, state-space digital filters.**

I. Introduction

THE basic arithmetic operations involved in the implementation of an infinite-impulse response (IIR) digital filter are multiplications of input/output samples by the filter coefficients and additions. For arithmetic operations involving fixed-point numbers, the result of a multiplication must be rounded or truncated. This quantization error generates roundoff noise at the filter output. In addition, because the result of an addition can exceed the finite register length, the dynamic range of the digital filter is always a concern in a fixed-point implementation. Error feedback has been known as an effective technique for reducing the roundoff noise at the filter output. This is achieved by extracting the quantization error after multiplication and addition, and then feeding the error signal back through simple circuits. It can be applied to digital filters that are described by either external or internal models for roundoff noise minimization without affecting the filter's input-output characteristics. Many techniques for error feedback have been presented in the past [1]–[10]. Another technique for reducing the roundoff noise is to synthesize an optimal state-space filter structure that minimizes the roundoff noise under l_2-norm dynamic-range scaling constraints on the state-variables by means of coordinate transformation in the state space [11]–[14]. In addition, techniques for reducing the roundoff noise subject to the scaling constraints by combining the coordinate transformation with error feedback have been proposed [15], [16]. It has also been shown that the roundoff noise can be reduced by means of delta operator [16], [17] and that the digital filter in this case can be viewed as a special case of the filter with error feedback [16].

In this paper, several new algorithms for reducing the roundoff noise in state-space digital filters are proposed. First, the problem of roundoff noise reduction for several typical state-space realizations using error feedback is investigated. Closed-form formulas for evaluating the optimal full-scale, diagonal, and scalar error-feedback matrices for a given state-space digital filter are derived, where three realization schemes, namely, the general state-space realization, input-balanced realization (to be defined shortly), and optimal realization (in the sense of [12], [13]) are examined. Furthermore, an iterative noise reduction technique for state-space digital filters that jointly optimizes a scalar error-feedback matrix and a coordinate transformation matrix is proposed. A case study is presented to illustrate the algorithms proposed and to demonstrate their performance as compared with that of the existing methods [15], [16].

Throughout this paper, $\boldsymbol{I}_n$ stands for the identity matrix of dimension $n \times n$, the transpose (conjugate transpose) of a matrix $\boldsymbol{A}$ is indicated by $\boldsymbol{A}^T$ ($\boldsymbol{A}^*$), and the trace, eigenvalue, and ith diagonal element of a square matrix $\boldsymbol{A}$ are denoted by $tr[\boldsymbol{A}]$, $\lambda(\boldsymbol{A})$, and $(\boldsymbol{A})_{ii}$, respectively.

II. State-Space Digital Filters With Error Feedback

Let $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$ be a state-space description of an nth-order IIR digital filter, i.e.

$$\begin{aligned} \boldsymbol{x}(k+1) &= \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{b}u(k) \\ y(k) &= \boldsymbol{c}\boldsymbol{x}(k) + du(k) \end{aligned} \tag{1}$$

where $\boldsymbol{x}(k)$ is an $n \times 1$ state-variable vector, $u(k)$ is a scalar input, $y(k)$ is a scalar output, and $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and d are real constant matrices of appropriate dimensions. The filter described in (1) is assumed to be stable, controllable and observable. Due to finite register sizes, finite-word-length (FWL) constraints are

Manuscript received February 20, 2002; revised September 9, 2002. This paper was recommended by Associate Editor W. P. Zhu.

T. Hinamoto and H. Ohnishi are with the Graduate School of Engineering, Hiroshima University, Higashi-Hiroshima 739-8527, Japan.

W.-S. Lu is with the Department of Electrial and Computer Engineering, University of Victoria, Victoria, BC V8W 3P6, Canada.

Digital Object Identifier 10.1109/TCSI.2002.807512

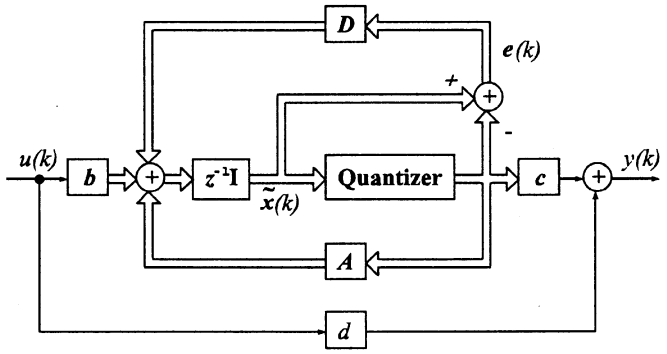

Fig. 1. Error feedback in a state-space digital filter.

imposed on the state variables, input, output, and system parameters as shown in $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$.

By taking the quantization performed before matrix-vector multiplication into account, an FWL implementation of (1) can be expressed as

$$\begin{aligned} \tilde{\boldsymbol{x}}(k+1) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + \boldsymbol{b}u(k) \\ \tilde{y}(k) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + du(k) \end{aligned} \tag{2}$$

where each component of coefficient matrices $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and d assumes an exact fractional B_c-b representation. The FWL state-variable vector $\tilde{\boldsymbol{x}}(k)$ and the output $\tilde{y}(k)$ all have a B-b fractional representation, while the input $u(k)$ is a $(B-B_c)$-b fraction.

The quantizer $\boldsymbol{Q}[\cdot]$ in (2) rounds the B-b fraction $\tilde{\boldsymbol{x}}(k)$ to $(B-B_c)$ b after completing the multiplications and additions, where the sign bit is not counted. In a fixed-point implementation, the quantization is usually carried out by two's complement truncation which discards the lower bits of a double-precision accumulator. Therefore, the quantization error

$$\boldsymbol{e}(k) = \tilde{\boldsymbol{x}}(k) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] \tag{3}$$

is equal to the residue left in the lower part of $\tilde{\boldsymbol{x}}(k)$. The block-diagram representation of a state-space digital filter with error feedback is shown in Fig. 1 where the roundoff error $\boldsymbol{e}(k)$ is modeled as a zero-mean noise process of covariance $\sigma^2\boldsymbol{I}_n$ with

$$\sigma^2 = \frac{1}{12}2^{-2(B-B_c)}$$

and the quantization error $\boldsymbol{e}(k)$ is fed back through an $n \times n$ constant matrix $\boldsymbol{D}$ in the FWL filter (2). From Fig. 1, it is obvious that the filter can be characterized by the state-space model

$$\begin{aligned} \tilde{\boldsymbol{x}}(k+1) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + \boldsymbol{b}u(k) + \boldsymbol{D}\boldsymbol{e}(k) \\ \tilde{y}(k) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + du(k) \end{aligned} \tag{4}$$

where $\boldsymbol{D}$ is referred to as an *error-feedback matrix*.

Subtracting (4) from (1) yields

$$\begin{aligned} \Delta\boldsymbol{x}(k+1) &= \boldsymbol{A}\Delta\boldsymbol{x}(k) + (\boldsymbol{A}-\boldsymbol{D})\boldsymbol{e}(k) \\ \Delta y(k) &= \boldsymbol{c}\Delta\boldsymbol{x}(k) + \boldsymbol{c}\boldsymbol{e}(k) \end{aligned} \tag{5}$$

where

$$\Delta\boldsymbol{x}(k) = \boldsymbol{x}(k) - \tilde{\boldsymbol{x}}(k), \quad \Delta y(k) = y(k) - \tilde{y}(k).$$

Taking the z-transform on both sides of (5) and setting $\Delta\boldsymbol{x}(0) = \boldsymbol{0}$, we have

$$\begin{aligned} \Delta Y(z) &= \boldsymbol{G}_D(z)\boldsymbol{E}(z) \\ \boldsymbol{G}_D(z) &= \boldsymbol{c}(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}(\boldsymbol{A}-\boldsymbol{D}) + \boldsymbol{c} \end{aligned} \tag{6}$$

where $\Delta Y(z)$ and $\boldsymbol{E}(z)$ represent the z-transform of $\Delta y(k)$ and $\boldsymbol{e}(k)$, respectively, and $\boldsymbol{G}_D(z)$ is the transfer function from the quantization error, $\boldsymbol{e}(k)$, to the filter output $\Delta y(k)$.

The noise gain $I(\boldsymbol{D}) = \sigma_{\text{out}}^2/\sigma^2$ is then defined by

$$\begin{aligned} I(\boldsymbol{D}) &= \frac{1}{2\pi j}\int_{|z|=1} \boldsymbol{G}_D(z)\boldsymbol{G}_D^*(z)\frac{dz}{z} \\ &= \text{tr}[\boldsymbol{W}_D] \end{aligned} \tag{7}$$

where

$$\boldsymbol{W}_D = \frac{1}{2\pi j}\int_{|z|=1} \boldsymbol{G}_D^*(z)\boldsymbol{G}_D(z)\frac{dz}{z}.$$

Utilizing the Cauchy integral theorem, the matrix $\boldsymbol{W}_D$ defined in (7) can be expressed in closed-form as

$$\boldsymbol{W}_D = (\boldsymbol{A}-\boldsymbol{D})^T\boldsymbol{W}_o(\boldsymbol{A}-\boldsymbol{D}) + \boldsymbol{c}^T\boldsymbol{c} \tag{8}$$

where $\boldsymbol{W}_o$ is the observability Gramian of the filter that can be obtained by solving the Lyapunov equation

$$\boldsymbol{W}_o = \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}. \tag{9}$$

Matrix $\boldsymbol{W}_o$ is referred to as the *unit noise matrix* for the filter, and $\boldsymbol{W}_D$ can be viewed as the unit noise matrix for the filter with error feedback specified by matrix $\boldsymbol{D}$.

The l_2-norm dynamic-range scaling constraints on the state variables involve the controllability Gramian of the filter, which is defined by

$$\boldsymbol{K}_c = \sum_{k=0}^{\infty} \boldsymbol{A}^k\boldsymbol{b}\boldsymbol{b}^T(\boldsymbol{A}^T)^k \tag{10}$$

and can be computed by solving the Lyapunov equation

$$\boldsymbol{K}_c = \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T. \tag{11}$$

The problem considered here is to design the error-feedback matrix $\boldsymbol{D}$ so as to reduce the noise gain in the sense that

$$\text{tr}[\boldsymbol{W}_D] < \text{tr}[\boldsymbol{W}_o] \tag{12}$$

is satisfied subject to that all the diagonal elements of $\boldsymbol{K}_c$ equal unity. Such constraints on matrix $\boldsymbol{K}_c$ are known as the *l_2-norm dynamic-range scaling constraints* [13].

A different yet equivalent state-space description of (1), $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$, can be obtained via a coordinate transformation $\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k)$ where

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T} \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b} \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{13}$$

Accordingly, the observability and controllability Gramians for $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$ become

$$\overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}, \quad \overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T} \tag{14}$$

respectively. If the l_2-norm dynamic-range scaling constraints are imposed on the state-variable vector $\overline{\boldsymbol{x}}(k)$, i.e.

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1, \quad i = 1, 2, \cdots, n \tag{15}$$

then, it can be shown that [12], [13]

$$\min_{\boldsymbol{T}} \text{tr}[\overline{\boldsymbol{W}}_o] = \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i\right)^2 \tag{16}$$

where σ_i^2 for $i = 1, 2, \cdots, n$ are the eigenvalues of matrix $\boldsymbol{K}_c\boldsymbol{W}_o$. The state-space realization satisfying (15) and (16) is called the *optimal realization* (which is sometimes also referred to as the *optimal filter structure*) with minimum roundoff noise. A method for constructing such a filter structure was proposed in [12], [13]. Obviously, if the realization in (1) is optimal with minimum roundoff noise, then the right-hand side of (12) is minimized.

In the next section, we will derive closed-form formulas for determining the optimal full-scale, diagonal, and scalar error-feedback matrix $\boldsymbol{D}$ that minimizes $\text{tr}[\boldsymbol{W}_D]$ for a given state-space digital filter.

III. Determination of Optimal Error-Feedback Matrices

A. Case 1: $\boldsymbol{D}$ is a General Matrix

Substituting (8) into (7), we obtain

$$\begin{aligned} I(\boldsymbol{D}) &= \text{tr}[\boldsymbol{c}^T\boldsymbol{c} + (\boldsymbol{A}-\boldsymbol{D})^T\boldsymbol{W}_o(\boldsymbol{A}-\boldsymbol{D})] \\ &= \text{tr}[\boldsymbol{W}_o] + \text{tr}[\boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{D}] - 2\text{tr}[\boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{A}]. \end{aligned} \tag{17}$$

Differentiating (17) with respect to the error-feedback matrix $\boldsymbol{D}$ yields

$$\frac{\partial I(\boldsymbol{D})}{\partial \boldsymbol{D}} = 2\boldsymbol{W}_o(\boldsymbol{D}-\boldsymbol{A}). \tag{18}$$

By choosing the error-feedback matrix as $\boldsymbol{D} = \boldsymbol{A}$, the noise gain $I(\boldsymbol{D})$ in (17) achieves its minimum value

$$I_{\min}(\boldsymbol{D}) = \text{tr}[\boldsymbol{W}_o] - \text{tr}[\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A}] = \text{tr}[\boldsymbol{c}^T\boldsymbol{c}]. \tag{19}$$

B. Case 2: $\boldsymbol{D}$ is a Diagonal Matrix

In this case, matrix $\boldsymbol{D}$ assumes the form

$$\boldsymbol{D} = \text{diag}\{\alpha_1, \alpha_2, \cdots, \alpha_n\} \tag{20}$$

which leads (17) to

$$I(\boldsymbol{D}) = \text{tr}[\boldsymbol{W}_o] + \sum_{i=1}^{n}(\boldsymbol{W}_o)_{ii}\alpha_i^2 - 2\sum_{i=1}^{n}(\boldsymbol{W}_o\boldsymbol{A})_{ii}\alpha_i. \tag{21}$$

From (21), it is clear that $I(\boldsymbol{D}) = \text{tr}[\boldsymbol{W}_D] < \text{tr}[\boldsymbol{W}_o]$ holds if α_i's satisfy

$$\alpha_i\left(\alpha_i - 2\frac{(\boldsymbol{W}_o\boldsymbol{A})_{ii}}{(\boldsymbol{W}_o)_{ii}}\right) < 0, \qquad i = 1, 2, \cdots, n. \tag{22}$$

If we compute the gradient of $I(\boldsymbol{D})$ in (21) and set it to zero, i.e.,

$$\frac{\partial I(\boldsymbol{D})}{\partial \alpha_i} = 2\left[(\boldsymbol{W}_o)_{ii}\alpha_i - (\boldsymbol{W}_o\boldsymbol{A})_{ii}\right] = 0, \qquad i = 1, 2, \cdots, n \tag{23}$$

then, we obtain the minimizer $\boldsymbol{D} = \text{diag}\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ with

$$\alpha_i = \frac{(\boldsymbol{W}_o\boldsymbol{A})_{ii}}{(\boldsymbol{W}_o)_{ii}}, \qquad i = 1, 2, \cdots, n \tag{24}$$

at which $I(\boldsymbol{D})$ achieves its minimum as

$$I_{\min}(\boldsymbol{D}) = \text{tr}[\boldsymbol{W}_o] - \sum_{i=1}^{n}\frac{(\boldsymbol{W}_o\boldsymbol{A})_{ii}^2}{(\boldsymbol{W}_o)_{ii}}. \tag{25}$$

In the rest of the paper, the filter in (1) is said to be *input-balanced (internally balanced)* if $\boldsymbol{K}_c = \boldsymbol{I}_n$ and $\boldsymbol{W}_o = \boldsymbol{\Sigma}^2 (\boldsymbol{K}_c = \boldsymbol{W}_o = \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \text{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_n\}$. The theorem below characterizes the diagonal error-feedback matrix that minimizes the noise gain $I(\boldsymbol{D})$ for the input-balanced realization of an IIR digital filter.

Theorem 1: If the filter in (1) is input-balanced, and the diagonal error-feedback matrix $\boldsymbol{D} = \text{diag}\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ satisfies

$$\alpha_i(\alpha_i - 2a_{ii}) < 0, \qquad i = 1, 2, \cdots, n \tag{26}$$

then $I(\boldsymbol{D}) = \text{tr}[\boldsymbol{W}_D] < \text{tr}[\boldsymbol{W}_o]$ where a_{ij} denotes the (i, j)th element of matrix $\boldsymbol{A}$. Moreover, if

$$\alpha_i = a_{ii}, \qquad i = 1, 2, \cdots, n \tag{27}$$

then $I(\boldsymbol{D})$ achieves the minimum value

$$I_{\min}(\boldsymbol{D}) = \sum_{i=1}^{n}(1 - a_{ii}^2)\sigma_i^2. \tag{28}$$

Proof: $\boldsymbol{W}_o = \boldsymbol{\Sigma}^2$ implies that $w_{ii}^o = \sigma_i^2$ and $w_{ij}^o = 0$ for $i \neq j$. Under these circumstances, (22), (24) and (25) are reduced to (26), (27) and (28), respectively. This completes the proof.

C. Case 3: $\boldsymbol{D}$ is a Scalar Matrix $\alpha\boldsymbol{I}_n$

If $\boldsymbol{D} = \alpha\boldsymbol{I}_n$ with a scalar α, then (17) becomes

$$I(\boldsymbol{D}) = \text{tr}[\boldsymbol{W}_o] + \alpha\left(\text{tr}[\boldsymbol{W}_o]\alpha - 2\text{tr}[\boldsymbol{W}_o\boldsymbol{A}]\right). \tag{29}$$

Hence, $I(\boldsymbol{D}) = \text{tr}[\boldsymbol{W}_D] < \text{tr}[\boldsymbol{W}_o]$ holds if α satisfies

$$\alpha\left(\alpha - 2\frac{\text{tr}[\boldsymbol{W}_o\boldsymbol{A}]}{\text{tr}[\boldsymbol{W}_o]}\right) < 0. \tag{30}$$

Moreover, from $\partial I(\boldsymbol{D})/\partial\alpha = 0$ it follows that the value of α that minimizes $I(\boldsymbol{D})$ is given by

$$\alpha = \frac{\text{tr}[\boldsymbol{W}_o\boldsymbol{A}]}{\text{tr}[\boldsymbol{W}_o]} \tag{31}$$

which leads (29) to

$$I_{\min}(\boldsymbol{D}) = \text{tr}[\boldsymbol{W}_o]\left[1 - \left(\frac{\text{tr}[\boldsymbol{W}_o\boldsymbol{A}]}{\text{tr}[\boldsymbol{W}_o]}\right)^2\right]. \tag{32}$$

Therem 2 concerns the optimal realization satisfying (15) and (16).

Theorem 2: If the state-space realization (1) has an optimal realization with minimum roundoff noise, then (30) and (31) can be expressed as

$$\alpha\left(\alpha-2\frac{a_{11}^{(b)}\sigma_1+\cdots+a_{nn}^{(b)}\sigma_n}{\sigma_1+\sigma_2+\cdots+\sigma_n}\right)<0 \tag{33}$$

and

$$\alpha=\frac{a_{11}^{(b)}\sigma_1+\cdots+a_{nn}^{(b)}\sigma_n}{\sigma_1+\sigma_2+\cdots+\sigma_n} \tag{34}$$

respectively, where $a_{ii}^{(b)}$ for $i=1,2,\cdots,n$ are the diagonal elements of the matrix $\boldsymbol{A}$ in the input-balanced realization. Moreover, (32) becomes

$$I_{\min}(\boldsymbol{D})=\frac{1}{n}\left[\left(\sum_{i=1}^{n}\sigma_i\right)^2-\left(\sum_{i=1}^{n}a_{ii}^{(b)}\sigma_i\right)^2\right]. \tag{35}$$

Proof: See Appendix I.

Corollary 1: If the state-space realization (1) is input-balanced, then (30) and (31) can be expressed as

$$\alpha\left(\alpha-2\frac{a_{11}\sigma_1^2+\cdots+a_{nn}\sigma_n^2}{\sigma_1^2+\sigma_2^2+\cdots+\sigma_n^2}\right)<0 \tag{36}$$

and

$$\alpha=\frac{a_{11}\sigma_1^2+\cdots+a_{nn}\sigma_n^2}{\sigma_1^2+\cdots+\sigma_n^2} \tag{37}$$

respectively, and (32) becomes

$$I_{\min}(\boldsymbol{D})=\left(\sum_{i=1}^{n}\sigma_i^2\right)-\frac{(a_{11}\sigma_1^2+\cdots+a_{nn}\sigma_n^2)^2}{\sigma_1^2+\sigma_2^2+\cdots+\sigma_n^2}. \tag{38}$$

Lemma 1: If the l_2-norm dynamic-range scaling constraints $(\overline{\boldsymbol{K}}_c)_{ii}=(\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii}=1$ for $i=1,2,\cdots,n$ are imposed on the state variables, then

$$\min_{\boldsymbol{T}}\ \mathrm{tr}[\overline{\boldsymbol{W}}_D]=\frac{1}{n}\left(\sum_{i=1}^{n}\nu_i\right)^2 \tag{39}$$

where $\overline{\boldsymbol{W}}_D=\boldsymbol{T}^T\boldsymbol{W}_D\boldsymbol{T}$ with $\boldsymbol{D}=\alpha\boldsymbol{I}_n$, and ν_i^2 for $i=1,2,\cdots,n$ are the eigenvalues of $\boldsymbol{K}_c\boldsymbol{W}_D$.

The proof of Lemma 1 is similar to that in [13] and is therefore omitted.

Remark 1: Under the l_2-norm dynamic-range scaling constraints, Williamson [15] obtained a suboptimal realization with an integer error-feedback matrix $\boldsymbol{D}=\boldsymbol{I}_n$ $(\boldsymbol{D}=-\boldsymbol{I}_n)$ for low-pass (high-pass) narrow-band filters by choosing an appropriate coordinate transformation matrix $\boldsymbol{T}$. A similar argument was also made in [16] by restricting $\boldsymbol{D}$ to $\boldsymbol{I}_n$.

Theorem 3: Let the state-space realization in (1) be input-balanced. If there exists a real number α such that

$$\alpha(\alpha-2a_{ii})<0 \tag{40}$$

then

$$\sum_{i=1}^{n}\nu_i<\sum_{i=1}^{n}\sigma_i \tag{41}$$

or equivalently

$$\min_{\boldsymbol{T}}\ \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{W}_D\boldsymbol{T}]<\min_{\boldsymbol{T}}\ \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}]. \tag{42}$$

Moreover, if the diagonal elements of matrix $\boldsymbol{A}$, a_{ii} for $i=1,2,\cdots,n$, are equal, then the upper bound of $\sum_{i=1}^{n}\nu_i$ is minimized by choosing $\alpha=a_{ii}$.

Proof: See Appendix 2.

Remark 2: It is noted that if $\alpha=1$ then (40) is changed to $a_{ii}>1/2$ for all i which is identical to the condition shown in Theorem 4.1 in [16, p. 632].

Remark 3: As a simple example, we note that when a second-order digital filter has complex conjugate poles, the diagonal elements of a $n\times n$ matrix $\boldsymbol{A}$ are equal, i.e.

$$\boldsymbol{A}=\begin{bmatrix}r\cos\theta & r\sin\theta\\ -r\sin\theta & r\cos\theta\end{bmatrix}$$

where the poles are given by $re^{\pm j\theta}$ with $0<r<1$. In this case, we can choose $\alpha=r\cos\theta$.

It is known [16, p. 632] that for the filter in (1), there exists an internally balanced realization $(\boldsymbol{A},\boldsymbol{b},\boldsymbol{c},d)_n$ with

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ -\boldsymbol{A}_{12}^T & \boldsymbol{A}_{22}\end{bmatrix}\quad \boldsymbol{A}_{11}=\boldsymbol{A}_{11}^T\quad \boldsymbol{A}_{22}=\boldsymbol{A}_{22}^T. \tag{43}$$

The next theorem describes a counterpart of Theorem 3 for an internally balanced realization.

Theorem 4: Let the filter in (1) be realized by an internally balanced realization in (43), and let the union of $\lambda(\boldsymbol{A}_{11})$ and $\lambda(\boldsymbol{A}_{22})$ be denoted by $\{\theta_i\}$. If there exists a scalar α satisfying the inequality

$$0<\frac{\alpha}{2}\le\min_i\{\theta_i\}\quad\left(\max_i\{\theta_i\}\le\frac{\alpha}{2}<0\right) \tag{44}$$

then (41) holds for such an α.

Proof: Since $\boldsymbol{A}_{ii}=\boldsymbol{A}_{ii}^T$ for $i=1,2$ and since $\{\theta_i\}=\lambda(\boldsymbol{A}_{11})\cup\lambda(\boldsymbol{A}_{22})$, if

$$\min_i\{\theta_i\}>0\quad\left(\max_i\{\theta_i\}<0\right) \tag{45}$$

then, matrices $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are positive-definite (negative-definite). In general

$$\lambda_{\min}(\boldsymbol{A})\le\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\le\lambda_{\max}(\boldsymbol{A}) \tag{46}$$

holds for any symmetric matrix $\boldsymbol{A}$ where $\boldsymbol{x}$ is any real vector satisfying $||\boldsymbol{x}||=1$. This yields

$$\min_i\{\theta_i\}\le\min_i\{a_{ii}\}\quad\left(\max_i\{a_{ii}\}\le\max_i\{\theta_i\}\right). \tag{47}$$

From (44) and (47) it follows that

$$0<\frac{\alpha}{2}<a_{ii}\quad\left(a_{ii}<\frac{\alpha}{2}<0\right) \tag{48}$$

which satisfy the condition in (40). It then follows from Theorm 3 that (41) holds.

This completes the proof.

Remark 4: It is noted that the substitution of $\alpha=1$ into (44) yields

$$\frac{1}{2}\le\min_i\{\theta_i\}$$

which corresponds to the condition stated in Theorem 4.2 in [16, p. 632].

Theorem 5: Let λ_i for $i = 1, 2, \cdots, n$ be the eigenvalues of the matrix $\boldsymbol{A}$ in (1). If there exists an $\alpha > 0$ ($\alpha < 0$) satisfying

$$\bar{\lambda} \geq 1 - \frac{2-\alpha}{2n} \quad \left(\bar{\lambda} \leq -\left(1 - \frac{2+\alpha}{2n}\right)\right) \tag{49}$$

then (41) holds for this α, where $\bar{\lambda}$ denotes the mean value of the eigenvalues, i.e., $\bar{\lambda} = (\lambda_1 + \lambda_2 + \cdots + \lambda_n)/n$.

Proof: Without loss of generality, we assume that the filter in (1) is input-balanced. Then the diagonal element a_{ii} of matrix $\boldsymbol{A}$ satisfies

$$|a_{ii}| < 1, \qquad i = 1, 2, \cdots, n. \tag{50}$$

Since

$$\sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i \tag{51}$$

holds in general, it follows from (50) that

$$\min_i\{a_{ii}\} + n - 1 > \sum_{i=1}^{n} \lambda_i \quad \left(\max_i\{a_{ii}\} - (n-1) < \sum_{i=1}^{n} \lambda_i\right) \tag{52}$$

Alternatively, (49) can be written as

$$\sum_{i=1}^{n} \lambda_i \geq n - 1 + \frac{\alpha}{2} \quad \left(\sum_{i=1}^{n} \lambda_i \leq -(n-1) + \frac{\alpha}{2}\right) \tag{53}$$

Together (52) and (53) imply that

$$\min_i\{a_{ii}\} > \frac{\alpha}{2} \quad \left(\max_i\{a_{ii}\} < \frac{\alpha}{2}\right) \tag{54}$$

which satisfy the condition in (40). It then follows from Theorem 3 that (41) holds. This completes the proof.

Remark 5: It is noted that the substitution of $\alpha = 1$ into (49) yields

$$\sum_{i=1}^{n} \lambda_i \geq n - \frac{1}{2}$$

which coincides with the condition stated in [16, p. 632, Th. 4.3].

IV. Noise Reduction by Joint Optimization of Error Feedback and Coordinate Transformation

In this section, we consider the problem of joint optimization of a scalar error-feedback matrix $\alpha \boldsymbol{I}_n$ and a coordinate transformation matrix $\boldsymbol{T}$ for roundoff noise minimization under l_2-norm dynamic-range scaling constraints. The proposed joint optimization will be carried out in an iterative manner. First, a scalar α is obtained by modifying (31) under l_2-norm dynamic-range scaling constraints. In what follows, the unit noise matrix $\boldsymbol{W}_D$ in (8) with $\boldsymbol{D} = \alpha \boldsymbol{I}_n$ is denoted by $\boldsymbol{W}_\alpha$. Under joint application of a scalar error-feedback and a coordinate transformation, the noise gain $I(\boldsymbol{D})$ becomes $\mathrm{tr}[\boldsymbol{T}^T \boldsymbol{W}_\alpha \boldsymbol{T}]$. In order to minimize $\mathrm{tr}[\boldsymbol{T}^T \boldsymbol{W}_\alpha \boldsymbol{T}]$ (with α fixed) over an $n \times n$ nonsingular matrix $\boldsymbol{T}$ subject to the constraints in (15), we define the Lagrange function

$$J(\alpha, \boldsymbol{P}, \lambda) = \mathrm{tr}[\boldsymbol{W}_\alpha \boldsymbol{P}] + \lambda(\mathrm{tr}[\boldsymbol{K}_c \boldsymbol{P}^{-1}] - n) \tag{55}$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ and λ is a Lagrange multiplier. Using the formulas for evaluating matrix gradient [18, p. 275]

$$\begin{aligned} \frac{\partial(\mathrm{tr}[\boldsymbol{M}\boldsymbol{X}])}{\partial \boldsymbol{X}} &= \boldsymbol{M}^T \\ \frac{\partial\left(\mathrm{tr}\left[\boldsymbol{M}\boldsymbol{X}^{-1}\right]\right)}{\partial \boldsymbol{X}} &= -\left[\boldsymbol{X}^{-1}\boldsymbol{M}\boldsymbol{X}^{-1}\right]^T \end{aligned} \tag{56}$$

we compute

$$\begin{aligned} \frac{\partial J(\alpha, \boldsymbol{P}, \lambda)}{\partial \boldsymbol{P}} &= \boldsymbol{W}_\alpha - \lambda \boldsymbol{P}^{-1} \boldsymbol{K}_c \boldsymbol{P}^{-1} \\ \frac{\partial J(\alpha, \boldsymbol{P}, \lambda)}{\partial \lambda} &= \mathrm{tr}\left[\boldsymbol{K}_c \boldsymbol{P}^{-1}\right] - n. \end{aligned} \tag{57}$$

If we let $\partial J(\alpha, \boldsymbol{P}, \lambda)/\partial \boldsymbol{P} = \boldsymbol{0}$ and $\partial J(\alpha, \boldsymbol{P}, \lambda)/\partial \lambda = 0$, then

$$\boldsymbol{P}\boldsymbol{W}_\alpha \boldsymbol{P} = \lambda \boldsymbol{K}_c, \quad \mathrm{tr}\left[\boldsymbol{K}_c \boldsymbol{P}^{-1}\right] = n. \tag{58}$$

Note that if matrices $\boldsymbol{W} > 0$ and $\boldsymbol{M} \geq 0$ are symmetric, then the matrix equation $\boldsymbol{P}\boldsymbol{W}\boldsymbol{P} = \boldsymbol{M}$ has the unique solution [19]

$$\boldsymbol{P} = \boldsymbol{W}^{-\frac{1}{2}} \left[\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{M} \boldsymbol{W}^{\frac{1}{2}}\right]^{\frac{1}{2}} \boldsymbol{W}^{-\frac{1}{2}}. \tag{59}$$

It follows from (58) that

$$\begin{aligned} \boldsymbol{P} &= \sqrt{\lambda} \boldsymbol{W}_\alpha^{-\frac{1}{2}} \left[\boldsymbol{W}_\alpha^{\frac{1}{2}} \boldsymbol{K}_c \boldsymbol{W}_\alpha^{\frac{1}{2}}\right]^{\frac{1}{2}} \boldsymbol{W}_\alpha^{-\frac{1}{2}} \\ \frac{1}{\sqrt{\lambda}} \mathrm{tr}\left[\boldsymbol{K}_c \boldsymbol{W}_\alpha\right]^{\frac{1}{2}} &= \frac{1}{\sqrt{\lambda}} \left(\sum_{i=1}^{n} \theta_i\right) = n \end{aligned} \tag{60}$$

where θ_i^2 for $i = 1, 2, \cdots, n$ are the eigenvalues of $\boldsymbol{K}_c \boldsymbol{W}_\alpha$. Therefore, we obtain

$$\boldsymbol{P} = \frac{1}{n} \left(\sum_{i=1}^{n} \theta_i\right) \boldsymbol{W}_\alpha^{-\frac{1}{2}} \left[\boldsymbol{W}_\alpha^{\frac{1}{2}} \boldsymbol{K}_c \boldsymbol{W}_\alpha^{\frac{1}{2}}\right]^{\frac{1}{2}} \boldsymbol{W}_\alpha^{-\frac{1}{2}}. \tag{61}$$

Substituting (61) into (55) yields the minimum value of $J(\alpha, \boldsymbol{P}, \lambda)$ as

$$\min_{\boldsymbol{P}, \lambda} J(\alpha, \boldsymbol{P}, \lambda) = \frac{1}{n} \left(\sum_{i=1}^{n} \theta_i\right)^2 \tag{62}$$

for a given scalar α.

From (61), the optimal coordinate transformation matrix $\boldsymbol{T}$ that minimizes (55) can be obtained in closed form as

$$\boldsymbol{T} = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \theta_i\right)^{\frac{1}{2}} \boldsymbol{W}_\alpha^{-\frac{1}{2}} \left[\boldsymbol{W}_\alpha^{\frac{1}{2}} \boldsymbol{K}_c \boldsymbol{W}_\alpha^{\frac{1}{2}}\right]^{\frac{1}{4}} \boldsymbol{U} \tag{63}$$

where $\boldsymbol{U}$ is an arbitrary $n \times n$ orthogonal matrix. From (63) it follows that

$$\begin{aligned} \overline{\boldsymbol{K}}_c &= \boldsymbol{T}^{-1} \boldsymbol{K}_c \boldsymbol{T}^{-T} \\ &= n \left(\sum_{i=1}^{n} \theta_i\right)^{-1} \boldsymbol{U}^T \left[\boldsymbol{W}_\alpha^{\frac{1}{2}} \boldsymbol{K}_c \boldsymbol{W}_\alpha^{\frac{1}{2}}\right]^{\frac{1}{2}} \boldsymbol{U}. \end{aligned} \tag{64}$$

Next, we choose the $n \times n$ orthogonal matrix $\boldsymbol{U}$ such that the matrix $\overline{\boldsymbol{K}}$ in (64) satisfies the l_2-norm dynamic-range scaling constraints on the state-variables in (15). To this end, we perform the eigenvalue–eigenvector decomposition

$$\left[\boldsymbol{W}_\alpha^{\frac{1}{2}} \boldsymbol{K}_c \boldsymbol{W}_\alpha^{\frac{1}{2}}\right]^{\frac{1}{2}} = \boldsymbol{R} \boldsymbol{\Theta} \boldsymbol{R}^T \tag{65}$$

where

$$\begin{aligned}\boldsymbol{\Theta} &= \text{diag}\{\theta_1, \theta_2, \cdots, \theta_n\} \\ \boldsymbol{R}\boldsymbol{R}^T &= \boldsymbol{I}_n.\end{aligned}$$

Consequently

$$n\left(\sum_{i=1}^{n}\theta_i\right)^{-1}\left[\boldsymbol{W}_\alpha^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{W}_\alpha^{\frac{1}{2}}\right]^{\frac{1}{2}} = \boldsymbol{R}\boldsymbol{\Lambda}^{-2}\boldsymbol{R}^T \tag{66}$$

where

$$\begin{aligned}\boldsymbol{\Lambda} &= \text{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\} \\ \lambda_i &= \left(\frac{\theta_1+\theta_2+\cdots+\theta_n}{n\theta_i}\right)^{\frac{1}{2}}. \\ i &= 1, 2, \cdots, n.\end{aligned}$$

Now an $n \times n$ orthogonal matrix $\boldsymbol{S}$ such that

$$\boldsymbol{S}\boldsymbol{\Lambda}^{-2}\boldsymbol{S}^T = \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \tag{67}$$

can be obtained by numerical manipulations [13, p. 278]. By choosing $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{S}^T$ in (63), the optimal coordinate transformation matrix $\boldsymbol{T}$ satisfying (15) and (62) simultaneously can now be constructed as

$$\boldsymbol{T} = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\theta_i\right)^{\frac{1}{2}}\boldsymbol{W}_\alpha^{-\frac{1}{2}}\left[\boldsymbol{W}_\alpha^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{W}_\alpha^{\frac{1}{2}}\right]^{\frac{1}{4}}\boldsymbol{R}\boldsymbol{S}^T. \tag{68}$$

This coordinate transformation matrix $\boldsymbol{T}$ is used to minimize $\text{tr}[\overline{\boldsymbol{W}}_\alpha] = \text{tr}[\boldsymbol{T}^T\boldsymbol{W}_\alpha\boldsymbol{T}]$ subject to the constraints in (15). This completes the first round of iteration and, if necessary, this process may continue until both $\boldsymbol{T}$ and α converge. Having obtained transformation matrix $\boldsymbol{T}$, an improved value of scalar α can be obtained using

$$\begin{aligned}\alpha &= \frac{\text{tr}\left[\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{A}\boldsymbol{T}\right]}{\text{tr}\left[\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}\right]} \\ &= \frac{\text{tr}\left[\boldsymbol{W}_o\boldsymbol{A}\boldsymbol{P}\right]}{\text{tr}\left[\boldsymbol{W}_o\boldsymbol{P}\right]}.\end{aligned} \tag{69}$$

This iterative procedure for minimizing the roundoff noise under l_2-norm scaling constraints with respect to a scalar parameter α as well as an $n \times n$ symmetric positive-definite matrix $\boldsymbol{P}$ can be summarized as follows:

1) Set $i = 1$ and

$$\boldsymbol{P}(0) = \text{diag}\{(\boldsymbol{K}_c)_{11}^{-1}, (\boldsymbol{K}_c)_{22}^{-1}, \cdots, (\boldsymbol{K}_c)_{nn}^{-1}\}$$

.

2) Compute a scalar $\alpha(i)$ using

$$\alpha(i) = \frac{\text{tr}\left[\boldsymbol{W}_o\boldsymbol{A}\boldsymbol{P}(i-1)\right]}{\text{tr}\left[\boldsymbol{W}_o\boldsymbol{P}(i-1)\right]}.$$

3) Compute $I_{\min}(\alpha(i)\boldsymbol{I}_n) = (1-\alpha(i)^2)\text{tr}[\boldsymbol{W}_o\boldsymbol{P}(i-1)]$.

4) Replace $\boldsymbol{W}_\alpha$ by $\boldsymbol{W}_{\alpha(i)}$ computed using

$$\boldsymbol{W}_{\alpha(i)} = \left(1+\alpha(i)^2\right)\boldsymbol{W}\boldsymbol{o} - \alpha(i)\left(\boldsymbol{A}^T\boldsymbol{W}_o + \boldsymbol{W}_o\boldsymbol{A}\right).$$

5) Derive matrix $\boldsymbol{P}$ from (61), and take the resulting $\boldsymbol{P}$ as $\boldsymbol{P}(i)$.
6) Compute $\text{tr}[\boldsymbol{W}_{\alpha(i)}\boldsymbol{P}(i)]$.
7) Update i to $i+1$.
8) Repeat from Step 2) until the change in either $I_{\min}[\alpha(i)\boldsymbol{I}_n]$ or $\text{tr}[\boldsymbol{W}_{\alpha(i)}\boldsymbol{P}(i)]$ becomes negligible.
9) Obtain matrix $\boldsymbol{T}$ by using (68).

We have applied the proposed algorithm to quite a number of IIR filters, in all the cases we have tried so far, the noise gain is monotonically decreasing and the algorithm converges to the unique optimal solution $\boldsymbol{P}$ regardless of the initial point chosen. In all the cases that we tried, the number of iterations required has been fairly small. For example, simulation results for a ninth-order IIR low-pass digital filter showed that the algorithm converges after nine iterations. Although this may change for high-order filters, it should not be a concern as the order of IIR filters is usually not very high, and the algorithm is typically applied offline.

Suppose the above algorithm converges after N iterations and the optimal coordinate transformation matrix $\boldsymbol{T}(N)$ has been computed from (65)–(68). Then, according to (13), (14), and (21)–(24), the diagonal error-feedback matrix $\boldsymbol{D} = \text{diag}\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ that minimizes

$$\begin{aligned}I(\boldsymbol{D}) = \text{tr}\left[\boldsymbol{T}^T(N)\boldsymbol{W}_o\boldsymbol{T}(N)\right] + \text{tr}\left[\boldsymbol{T}^T(N)\boldsymbol{W}_o\boldsymbol{T}(N)\boldsymbol{D}^2\right] \\ -2\text{tr}\left[\boldsymbol{T}^T(N)\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{T}(N)\boldsymbol{D}\right]\end{aligned} \tag{70}$$

is given by

$$\alpha_i = \frac{\left(\boldsymbol{T}^T(N)\boldsymbol{W}_o\boldsymbol{A}\boldsymbol{T}(N)\right)_{ii}}{\left(\boldsymbol{T}^T(N)\boldsymbol{W}_o\boldsymbol{T}(N)\right)_{ii}}, \quad i = 1, 2, \cdots, n. \tag{71}$$

This diagonal error-feedback matrix $\boldsymbol{D}$ makes it possible to produce more reduction of the noise gain, i.e.

$$I_{\min}(\boldsymbol{D}) < I_{\min}\left[\alpha(N)\boldsymbol{I}_n\right]. \tag{72}$$

V. Case Study

In this section, we present a case study to illustrate the roundoff noise reduction methods proposed in the preceding sections.

A. Low-Pass Digital Filter

The system considered here is a third-order stable low-pass IIR digital filter realized as $(\boldsymbol{A_o}, \boldsymbol{b_o}, \boldsymbol{c_o}, d)_3$ in controllable canonical form

$$\begin{aligned}\boldsymbol{A}_o &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.339\,377 & -1.152\,652 & 1.520\,167 \end{bmatrix} \\ \boldsymbol{b}_o &= [0 \quad 0 \quad 0.437\,881]^T \\ \boldsymbol{c}_o &= [0.212\,964 \quad 0.293\,733 \quad 0.718\,718] \\ d &= 6.595\,92 \times 10^{-2}\end{aligned}$$

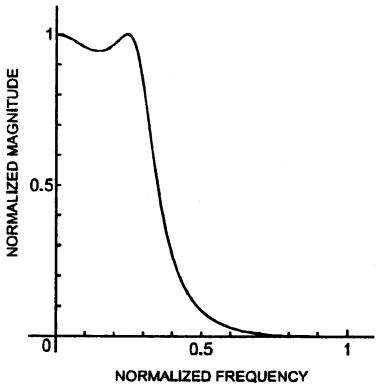

Fig. 2. Magnitude response of the low-pass digital filter.

whose controllability and observability Gramians are given by

$$\boldsymbol{K}_c = \begin{bmatrix} 1.0 & 0.741\,988 & 0.227\,107 \\ 0.741\,988 & 1.0 & 0.741\,988 \\ 0.227\,107 & 0.741\,988 & 1.0 \end{bmatrix}$$

$$\boldsymbol{W}_o = \begin{bmatrix} 0.720\,426 & -1.635\,144 & 1.753\,511 \\ -1.635\,144 & 4.551\,538 & -4.194\,133 \\ 1.753\,511 & -4.194\,133 & 5.861\,185 \end{bmatrix}.$$

The normalized cutoff frequency of the filter is about 0.32 and its magnitude response is shown in Fig. 2.

In the case study, three realization schemes of the above state-space digital filter subject to the l_2-norm dynamic-range scaling constraints will be considered.

- Input-balanced realization $(\boldsymbol{A}_{ib}, \boldsymbol{b}_{ib}, \boldsymbol{c}_{ib}, d)_3$ where

$$\boldsymbol{A}_{ib} = \begin{bmatrix} 0.718\,259 & -0.367\,681 & -0.004\,649 \\ 0.680\,718 & 0.402\,944 & 0.208\,077 \\ -0.032\,955 & -0.796\,728 & 0.398\,964 \end{bmatrix}$$

$$\boldsymbol{b}_{ib} = [-0.590\,672 \quad 0.575\,294 \quad 0.452\,733]^T$$

$$\boldsymbol{c}_{ib} = [-0.933\,934 \quad -0.491\,318 \quad 0.100\,978]$$

with the controllability and observability Gramians

$$\boldsymbol{K}_c^{ib} = \text{diag}\{1.0,\ 1.0,\ 1.0\}$$

$$\boldsymbol{W}_o^{ib} = \text{diag}\{2.499\,998,\ 0.729\,367,\ 0.049\,748\}.$$

- Optimal realization $(\boldsymbol{A}_{\text{opt}}, \boldsymbol{b}_{\text{opt}}, \boldsymbol{c}_{\text{opt}}, d)_3$ where

$$\boldsymbol{A}_{\text{opt}} = \begin{bmatrix} 0.476\,474 & 0.568\,865 & -0.071\,856 \\ -0.715\,925 & 0.476\,474 & 0.154\,270 \\ -0.154\,270 & 0.071\,856 & 0.567\,219 \end{bmatrix}$$

$$\boldsymbol{b}_{\text{opt}} = [0.686\,292 \quad 0.112\,451 \quad -0.713\,811]^T$$

$$\boldsymbol{c}_{\text{opt}} = [-0.099\,639 \quad -0.608\,103 \quad -0.632\,487]$$

with the controllability and observability Gramians

$$\boldsymbol{K}_c^{\text{opt}} = \begin{bmatrix} 1.0 & -0.036\,160 & -0.541\,747 \\ -0.036\,160 & 1.0 & 0.541\,747 \\ -0.541\,747 & 0.541\,747 & 1.0 \end{bmatrix}$$

$$\boldsymbol{W}_o^{\text{opt}} = \begin{bmatrix} 0.785\,120 & -0.028\,390 & -0.425\,337 \\ -0.028\,390 & 0.785\,120 & 0.425\,337 \\ -0.425\,337 & 0.425\,337 & 0.785\,120 \end{bmatrix}.$$

Without using error feedback, the noise gain $I(\boldsymbol{D})$ in (17) becomes $I(\boldsymbol{0}) = \text{tr}[\boldsymbol{W}_o]$. For comparison purposes, the noise gain without using error feedback is computed for each of the three realizations and listed in Table I.

TABLE I
ROUNDOFF NOISE GAIN OF THREE REALIZATIONS WITHOUT ERROR FEEDBACK

Realization	Controllable Canonical	Input-Balanced	Optimal
$I(\boldsymbol{0}) = \text{tr}[\boldsymbol{W}_o]$	11.133150	3.279113	2.355360

TABLE II
ROUNDOFF NOISE GAIN WITH OPTIMAL DIAGONAL ERROR-FEEDBACK MATRICES

Realization	Controllable Canonical	Input-Balanced	Optimal
Infinite Precision	4.598541	1.863033	1.520074
3-Bit Quantization	4.627836	1.866150	1.524145
Integer Quantization	5.245937	2.187819	1.922510

Case A. Noise Reduction Using a Diagonal Error-Feedback Matrix: If a diagonal error-feedback matrix is calculated using (24) for each of the above three realizations, then we obtain

$$\boldsymbol{D} = \text{diag}\{0.826\,041,\ 0.702\,890,\ 0.804\,589\}$$

for the controllable canonical realization

$$\boldsymbol{D} = \text{diag}\{0.718\,259,\ 0.402\,944,\ 0.398\,964\}$$

for the input-balanced realization, and

$$\boldsymbol{D} = \text{diag}\{0.585\,937,\ 0.494\,832,\ 0.689\,722\}$$

for the optimal realization. The noise reduction performance for these three realizations with optimal diagonal error-feedback is given in Table II, where the noise gains with the diagonal elements quantized to 3-bit numbers or integers are also included.

On comparing the above results with that in Table I, it is observed that error feedback with an appropriate diagonal matrix can reduce the roundoff noise considerably even with 3-bit or integer quantization.

Case B. Noise Reduction Using a Scalar Error-Feedback Matrix: We now consider the case where $\boldsymbol{D} = \alpha \boldsymbol{I}_3$ with a scalar α. Applying (31) to the three realizations, we obtain

$$\alpha = \begin{cases} 0.764\,400 & \text{for controllable canonical realization} \\ 0.643\,280 & \text{for input-balanced realization} \\ 0.590\,163 & \text{for optimal realization.} \end{cases}$$

The noise gain performance with the optimal scalar error feedback (with in finite precision or 3-bit or integer rounding) for the three realizations are shown in Table III.

On comparing the results obtained with that in Tables I and II we see that the performance with an optimal scalar error-feedback degrades only slightly in relative to what an optimal diagonal error-feedback matrix can achieve and remains significantly better than their no-error-feedback counterparts.

Case C. Noise Reduction by Joint Optimization of Scalar Error Feedback and Coordinate Transformation: In this case, we apply the iterative optimization procedure described in Section IV to the three realizations. For each realization, the

TABLE III
ROUNDOFF NOISE GAIN WITH OPTIMAL SCALAR ERROR-FEEDBACK MATRICES

Realization	Controllable Canonical	Input-Balanced	Optimal
Infinite Precision	4.627966	1.922186	1.535005
3-Bit Quantization	4.630275	1.923282	1.537864
Integer Quantization	5.245937	2.339451	1.930626

algorithm converges after seven iterations to the same matrix $\boldsymbol{P}^{\circ}$ and a scalar $\alpha = 0.647\,686$. In the case of controllable canonical realization, the coordinate transformation matrix $\boldsymbol{T}^{\circ}$ is given by

$$\boldsymbol{T}^{\circ} = \begin{bmatrix} -1.973\,853 & -0.153\,371 & -2.328\,357 \\ -0.063\,334 & -1.398\,294 & -1.260\,527 \\ 1.402\,772 & -0.676\,604 & -0.969\,851 \end{bmatrix}.$$

As is expected, the noise gain resulted from the algorithm is identical regardless of the realization to which it applies, and is given by $I_{\min}(\boldsymbol{D}) = 1.450\,049$. If $\alpha = 0.647\,686$ is rounded to power-of-two representation with 3 b after binary point, then, the noise gain is founded to be $I(\boldsymbol{D}) = 1.451\,335$.

Next, a refined solution which offers further reduced noise gain is deduced by applying an optimal diagonal error-feedback matrix to the optimized realization, i.e., $(\boldsymbol{T}^{o-1}\boldsymbol{A}\boldsymbol{T}^{o}, \boldsymbol{T}^{o-1}\boldsymbol{b}, \boldsymbol{c}\boldsymbol{T}^{o}, d)_3$. The optimal diagonal error-feedback matrix obtained by using (71) is given by

$$\boldsymbol{D} = \text{diag}\{0.705\,402,\ 0.510\,713,\ 0.683\,277\}$$

which yields $I_{\min}(\boldsymbol{D}) = 1.433\,755$. The above diagonal error-feedback matrix after 3-b quantization (power-of-two representation with 3 b after binary point) gives $I_{\min}(\boldsymbol{D}) = 1.438\,801$, which is less than $I_{\min}(\boldsymbol{D}) = 1.450\,049$ in the optimal scalar error feedback.

To compare the proposed methods with that reported in [15], [16], we choose the error-feedback matrix $\boldsymbol{D} = \boldsymbol{I}_n$ as in [15], [16], and minimize tr$[\boldsymbol{T}^T\boldsymbol{W}_D\boldsymbol{T}]$ with respect to a coordinate transformation matrix $\boldsymbol{T}$. This gives

$$\min_{\boldsymbol{T}} \text{ tr}\left[\boldsymbol{T}^T \text{W}_D \boldsymbol{T}\right] = 1.752\,546$$

which is considerably larger than our results described above.

Case D. Noise Reduction Using a General Error-Feedback Matrix: Finally, we examine the case where $\boldsymbol{D} = \boldsymbol{A}$ or matrix $\boldsymbol{D}$ is given by its approximation in each realization scheme. The performance of optimized general error-feedback matrices after different rounding polices in three state-space realizations as compared with their infinite-precision error-feedback counterparts are summarized in Table IV. More specifically, the table includes (i) the case with infinite precision error feedback; (ii) the case of rounding the optimal infinite-precision coefficients to integers; and (iii) the case of rounding the coefficients to power-of-two representations with 3 b after the binary point.

From the simulations, it is observed that the controllable canonical realization with an infinite-precision full-scale error-feedback matrix $\boldsymbol{D} = \boldsymbol{A}$ yields the smallest noise gain.

TABLE IV
ROUNDOFF NOISE GAIN USING GENERAL ERROR-FEEDBACK MATRICES

Realization	Controllable Canonical	Input-Balanced	Optimal
Infinite Precision	0.648188	1.123823	0.779757
3-Bit Quantization	0.662491	1.130837	0.790560
Integer Quantization	2.809319	1.894676	1.540156

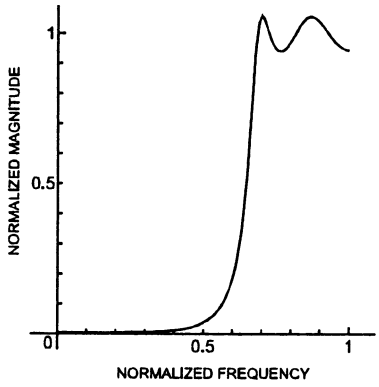

Fig. 3. Magnitude response of the high-pass digital filter.

However, as can be expected, the noise gain with this realization scheme is quite sensitive to the variations in the optimized $\boldsymbol{D}$. On the other hand, the noise gain for the input-balanced and optimal realizations are relatively insensitive to the quantization of $\boldsymbol{D}$ and the best overall performance is obviously offered by the optimal realization scheme.

B. A High-Pass Digital Filter

Now, let us consider a fourth-order stable high-pass filter realized as $(\boldsymbol{A}_o, \boldsymbol{b}_o, \boldsymbol{c}_o, d)_4$ in controllable canonical form

$$\boldsymbol{A}_o^T = \begin{bmatrix} 0 & 0 & 0 & -0.386\,952 \\ 1 & 0 & 0 & -1.467\,381 \\ 0 & 1 & 0 & -2.496\,662 \\ 0 & 0 & 1 & -2.225\,780 \end{bmatrix}$$

$$\boldsymbol{b}_o = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.240\,444 \end{bmatrix}$$

$$\boldsymbol{c}_o^T = \begin{bmatrix} 0.018\,325 \\ -0.219\,056 \\ 0.141\,882 \\ -0.271\,413 \end{bmatrix}.$$

The normalized cutoff frequency of the filter is about 0.66 and its magnitude response is drawn in Fig. 3.

Tables I–IV in the previous example are changed to Tables V–VIII in this example, respectively.

When the iterative optimization procedure described in Section IV is applied to the three realizations, the algorithm converges after seven iterations to a scalar $\alpha = -0.667\,273$ and $I_{\min}(\boldsymbol{D}) = 0.472\,563$ for each realization. If $\alpha = -0.667\,273$ is rounded to power-of-two representation with 3 b after binary point, then the noise gain is founded to be $I(\boldsymbol{D}) = 0.474\,085$.

TABLE V
ROUNDOFF NOISE GAIN OF THREE REALIZATIONS WITHOUT ERROR FEEDBACK

Realization	Controllable Canonical	Input-Balanced	Optimal
$I(\mathbf{0}) = \mathrm{tr}[\boldsymbol{W}_o]$	22.685416	1.125488	0.796551

TABLE VI
ROUNDOFF NOISE GAIN WITH OPTIMAL DIAGONAL ERROR-FEEDBACK MATRICES

Realization	Controllable Canonical	Input-Balanced	Optimal
Infinite Precision	8.803179	0.616628	0.469765
3-Bit Quantization	8.837581	0.617109	0.470395
Integer Quantization	9.892990	0.747868	0.535587

TABLE VII
ROUNDOFF NOISE GAIN WITH OPTIMAL SCALAR ERROR-FEEDBACK MATRICES

Realization	Controllable Canonical	Input-Balanced	Optimal
Infinite Precision	8.81442	0.641995	0.501540
3-Bit Quantization	8.837581	0.643037	0.501755
Integer Quantization	9.892990	0.775624	0.623585

TABLE VIII
ROUNDOFF NOISE GAIN USING GENERAL ERROR-FEEDBACK MATRICES

Realization	Controllable Canonical	Input-Balanced	Optimal
Infinite Precision	0.142117	0.268941	0.190752
3-Bit Quantization	0.150512	0.273714	0.195471
Integer Quantization	3.239478	0.636372	0.526276

Next, a refined solution which offers further reduced noise gain is deduced by applying an optimal diagonal error-feedback matrix to the resulting optimized realization. The optimal diagonal error-feedback matrix obtained by using (71) is given by

$$\boldsymbol{D} = \mathrm{diag}\{-0.712\,866, -0.750\,118, -0.434\,626, -0.661\,687\}$$

which yields $I_{\min}(\boldsymbol{D}) = 0.462\,844$. The above diagonal error-feedback matrix after 3-bit quantization (power-of-two representation with 3 b after binary point) gives $I_{\min}(\boldsymbol{D}) = 0.463\,942$, which is less than $I_{\min}(\boldsymbol{D}) = 0.472\,563$ in the optimal scalar error feedback.

The proposed methods can be compared with that reported in [15], [16] by choosing matrix $\boldsymbol{D} = \boldsymbol{I}_n$ and then minimizing $\mathrm{tr}[\boldsymbol{T}^T\boldsymbol{W}_D\boldsymbol{T}]$ with respect to matrix $\boldsymbol{T}$. Namely,

$$\min_{\boldsymbol{T}} \mathrm{tr}\left[\boldsymbol{T}^T\boldsymbol{W}_D\boldsymbol{T}\right] = 2.546\,985$$

which is much larger than our results described above.

VI. CONCLUSION

The roundoff noise minimization in state-space digital filters in several typical realization schemes by means of error feedback has been investigated, and general, diagonal, and scalar error-feedback matrices that minimize the normalized noise gain in a given state-space digital filter have been derived. Moreover, the noise minimization problem has also been addressed in scenario where a coordinate transformation and a scalar error-feedback matrix is jointly optimized subject to the usual l_2-norm dynamic-range scaling constraints. Simulation results in the form of a case study have been presented to illustrate and support our theoretical analysis and proposed algorithms. The proposed method may be regarded as an alternative, but much simpler and more general, approach to Hwang's method for synthesizing the optimal filter structure with minimum roundoff noise.

The extension of the results obtained in this paper to multidimensional case will appear elsewhere.

APPENDIX I
PROOF OF THEOREM 2

The optimal coordinate transformation matrix $\boldsymbol{T}$ described by (68) can be expressed in the form

$$\boldsymbol{T} = \boldsymbol{T}_b\boldsymbol{T}_o \tag{A1}$$

where α is set to zero and

$$\boldsymbol{T}_b = \boldsymbol{W}_o^{-\frac{1}{2}}\left[\boldsymbol{W}_o^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{W}_o^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{R}$$

$$\boldsymbol{T}_o = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\sigma_i\right)^{\frac{1}{2}}\boldsymbol{R}^T\left[\boldsymbol{W}_o^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{W}_o^{\frac{1}{2}}\right]^{-\frac{1}{4}}\boldsymbol{R}\boldsymbol{S}^T.$$

It follows that

$$\boldsymbol{T}_b^{-1}\boldsymbol{K}_c\boldsymbol{T}_b^{-T} = \boldsymbol{I}_n, \quad \boldsymbol{T}_b^T\boldsymbol{W}_o\boldsymbol{T}_b = \boldsymbol{\Sigma}^2. \tag{A2}$$

Equation (A2) shows that $\boldsymbol{T}_b$ is the coordinate transformation matrix that converts realization $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$ to an input-balanced realization. If $\alpha = 0$, then, (13), (14), (65), (A1), and (A2) imply that

$$\begin{aligned}\mathrm{tr}[\overline{\boldsymbol{W}}_o\overline{\boldsymbol{A}}] &= \mathrm{tr}\left[\boldsymbol{T}_o\boldsymbol{T}_o^T\boldsymbol{\Sigma}^2\boldsymbol{T}_b^{-1}\boldsymbol{A}\boldsymbol{T}_b\right] \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i\right)\mathrm{tr}\left[\boldsymbol{\Sigma}\boldsymbol{T}_b^{-1}\boldsymbol{A}\boldsymbol{T}_b\right] \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i\right)\left[\sum_{i=1}^{n}\left(\boldsymbol{T}_b^{-1}\boldsymbol{A}\boldsymbol{T}_b\right)_{ii}\sigma_i\right]\end{aligned} \tag{A3}$$

since $\boldsymbol{\Theta} = \boldsymbol{\Sigma}$ in this case, where $\boldsymbol{T}_b^{-1}\boldsymbol{A}\boldsymbol{T}_b$ is the system matrix in the input-balanced realization whose ith diagonal element $(\boldsymbol{T}_b^{-1}\boldsymbol{A}\boldsymbol{T}_b)_{ii}$ is replaced by $a_{ii}^{(b)}$ for $i = 1, 2, \cdots, n$ in this theorem. By replacing $\boldsymbol{W}_o$ and $\boldsymbol{A}$ by $\overline{\boldsymbol{W}}_o$ and $\overline{\boldsymbol{A}}$, respectively, and then substituting (16) and (A3) into (30)–(32), we obtain the results in (33)–(35). This completes the proof.

APPENDIX II
PROOF OF THEOREM 3

Lemma 2: Let $\{\rho_i^2, \rho_i \geq 0\}$ and $\{\delta_i^2, \delta_i \geq 0\}$ be the sets of diagonal elements and eigenvalues of a positive semidefinite symmetric matrix, respectively. Then

$$\sum_{i=1}^{n} \rho_i \geq \sum_{i=1}^{n} \delta_i \tag{A4}$$

where equality holds if and only if matrix $\boldsymbol{M}$ is diagonal, i.e., $\rho_i = \delta_i$ for any i [13].

From (8) it is derived that

$$\boldsymbol{W}_D = \boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{D} - \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{D} - \boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{W}_o. \tag{A5}$$

By substituting $\boldsymbol{D} = \alpha\boldsymbol{I}_n$ and $\boldsymbol{W}_o = \boldsymbol{\Sigma}^2$ into (A5), the ith diagonal element of matrix $\boldsymbol{W}_D, w_{ii}$ for $i = 1, 2, \cdots, n$, can be expressed as

$$\begin{aligned} w_{ii} &= (\alpha^2 - 2a_{ii}\alpha + 1)\sigma_i^2 \\ &= \left[(\alpha - a_{ii})^2 + 1 - a_{ii}^2\right]\sigma_i^2. \end{aligned} \tag{A6}$$

Now, noting that $\boldsymbol{K}_c = \boldsymbol{I}_n$, Lemma 2 and (A6) can be applied to yield

$$\sum_{i=1}^{n} \nu_i \leq \sum_{i=1}^{n} w_{ii}^{\frac{1}{2}} = \sum_{i=1}^{n} \sqrt{\alpha^2 - 2a_{ii}\alpha + 1}\,\sigma_i. \tag{A7}$$

Since the filter in (1) is input-balanced, the Lyapunov equation in (11) becomes

$$\boldsymbol{I}_n = \boldsymbol{A}\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T. \tag{A8}$$

This implies that each diagonal element a_{ii} of a stable matrix $\boldsymbol{A}$ satisfies

$$|a_{ii}| < 1, \quad i = 1, 2, \cdots, n. \tag{A9}$$

Therefore, for a scalar α satisfying (40), we have $0 < \alpha^2 - 2a_{ii}\alpha + 1 < 1$. This proves the inequality in (41).

In case the diagonal elements of $\boldsymbol{A}$ are equal, the choice of $\alpha = a_{ii}$ for all i yields

$$\sum_{i=1}^{n} \nu_i \leq \sqrt{1 - \alpha^2} \sum_{i=1}^{n} \sigma_i \tag{A10}$$

and in this case, the upper bound of $\sum_{i=1}^{n} \nu_i$ in (A10) is the tightest. This completes the proof.

REFERENCES

[1] H. A. Spang III and P. M. Shultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun. Syst.*, vol. 10, pp. 373–380, Dec. 1962.

[2] T. Thong and B. Liu, "Error spectrum shaping in narrowband recursive digital filters," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-25, pp. 200–203, Apr. 1977.

[3] T. L. Chang and S. A. White, "An error cancellation digital filter structure and its distributed-arithmetic implementation," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 339–342, Apr. 1981.

[4] D. C. Munson and D. Liu, "Narrowband recursive filters with error spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 160–163, Feb. 1981.

[5] W. E. Higgins and D. C. Munson, "Noise reduction strategies for digital filters: error spectrum shaping versus the optimal linear state-space formulation," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-30, pp. 963–973, Dec. 1982.

[6] M. Renfors, "Roundoff noise in error-feedback state-space filters," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'83)*, Apr. 1983, pp. 619–622.

[7] W. E. Higgins and D. C. Munson, "Optimal and suboptimal error-spectrum shaping for cascade-form digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 429–437, May 1984.

[8] T. I. Laakso and I. O. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," *IEEE Trans. Signal Processing*, vol. 40, pp. 1096–1107, May 1992.

[9] P. P. Vaidyanathan, "On error-spectrum shaping in state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 88–92, Jan. 1985.

[10] T. Hinamoto, S. Karino, N. Kuroda, and T. Kuma, "Error spectrum shaping in two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. 46, pp. 1203–1215, Oct. 1999.

[11] S. Y. Hwang, "Roundoff noise in state-space digital filtering: a general analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 256–262, June 1976.

[12] 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.

[13] 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.

[14] 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.

[15] D. Williamson, "Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using residue feedback," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1210–1220, Oct. 1986.

[16] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Signal Processing*, vol. 41, pp. 629–637, Feb. 1993.

[17] D. Williamson, "Delay replacement in direct form structures," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 453–460, Apr. 1988.

[18] L. L. Scharf, *Statistical Signal Processing*. Reading, MA: Addison-Wesley, 1991.

[19] 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 considerations," *IEEE Trans. Circuits Syst. I*, vol. 39, pp. 365–377, May 1992.

Takao Hinamoto (M'77–SM'84–F'01) received the B.E. degree from Okayama University, Okayama, Japan, in 1969, the M.E. degree from Kobe University, Kobe, Japan, in 1971, and the Dr. Eng. degree from Osaka University, Osaka, Japan, in 1977, all in electrical engineering.

From 1972 to 1988, he was with the Faculty of Engineering, Kobe University. From 1979 to 1981, he was a Visiting Member of Staff in the Department of Electrical Engineering, Queen's University, Kingston, ON, Canada, on leave from Kobe University. During 1988–1991, he was a Professor of electronic circuits in the Faculty of Engineering, Tottori University, Tottori, Japan. Since January 1992, he has been a Professor of electronic control in the Department of Electrical Engineering, Hiroshima University, Hiroshima, Japan. His research interests include digital signal processing, system theory, and control engineering. He has more than 270 published papers in these areas and is the co-editor and co-author of *Two-Dimensional Signal and Image Processing* (Tokyo, Japan: SICE, 1996). He served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II from 1993 to 1995, and presently serves as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I. He was the Guest Editor of the special section of DSP in the August 1998 issue of the *IEICE Transactions on Fundamentals*.

He also served as Chair of the 12th Digital Signal Processing (DSP) Symposium held in Hiroshima in November 1997, sponsored by the DSP Technical Committee of IEICE. Since 1995, he has been a member of the Steering Committee of the IEEE Midwest Symposium on Circuits and Systems, and since 1998, a member of the Digital Signal Processing Technical Committee in the IEEE Circuits and Systems Society. He served as a member of the Technical Program Committee for ISCAS'99. From 1993 to 2000, he served as a senator or member of the Board of Directors in the Society of Instrument and Control Engineers (SICE), and from 1999 to 2001, he was Chair of the Chugoku Chapter of SICE. He played a leading role in establishing the Hiroshima Section of IEEE, and served as the Interim Chair of the section. He is a recipient of the IEEE Third Millennium Medal.

Hiroaki Ohnishi received the B.E. degree in electrical engineering from Hiroshima University, Hiroshima, Japan, in 2001, where he is presently working toward the M.E. degree in electrical engineering.

His research interest is in the fields of digital signal processing.

Wu-Sheng Lu (S'81–M'85–SM'90–F'99) received the undergraduate degree in mathematics from Fudan University, Shanghai, China, in 1964, and the M.S. degree in electrical engineering, and Ph.D. degree in control science from University of Minnesota, Minneapolis, in 1983, and 1984, respectively.

He was a Post-Doctoral Fellow at the University of Victoria, Victoria, BC, Canada, in 1985, and a Visiting Assistant Professor at University of Minnesota in 1986. Since 1987, he has been with University of Victoria where he is currently a Professor. His teaching and research interests are in the areas of digital signal processing and application of optimization methods. He is the co-author, with A. Antoniou, of *Two-Dimensional Digital Filters* (New York, Marcel Dekker, 1992). He was an Associate Editor of the *Canadian Journal of Electrical and Computer Engineering* in 1989, and its Editor from 1990 to 1992. He served as an Associate Editor for IEEE TRANS. ON CIRCUITS AND SYSTEMS II from 1993 to 1995, and for IEEE TRANS. ON CIRCUITS AND SYSTEMS I from 1999 to 2001. He is presently an Associate Editor for the *International Journal of Multidimensional Systems and Signal Processing*.

Dr. Lu is a Fellow of the Engineering Institute of Canada.