Jointly Optimal High-Oder Error Feedback and Realization for Roundoff Noise Minimization in 1-D and 2-D State-Space Digital Filters

Takao Hinamoto, *Life Fellow, IEEE,* Akimitsu Doi and Wu-Sheng Lu, *Life Fellow, IEEE*

***Abstract*—Joint optimization of high-order error feedback and state-space realization for minimizing roundoff noise at filter output subject to l_2-scaling constraints is investigated for one- and two-dimensional state-space digital filters. Linear algebraic techniques that convert the problems at hand into an unconstrained optimization problem are explored, and an efficient quasi-Newton algorithm is then applied to solve the unconstrained optimization problem iteratively. In this connection, closed-form formulas are derived for fast and accurate gradient evaluation. Finally, case studies are presented to demonstrate that the high-order error feedback does offer much improved performance and that the proposed joint optimization is superior relative to a sequentially optimized system where the state-space coordinate transformation and high-order error feedback matrices are optimized separately.**

***Index Terms*—1-D state-space digital filter, 2-D state-space digital filter, Roesser's model, roundoff noise, quantization error, high-order error feedback, realization, l_2-scaling constraints, no overflow oscillation, joint optimization, quasi-Newton method**

I. INTRODUCTION

In the implementation of IIR digital filters with fixed-point arithmetic, it is of critical importance to reduce the effects of roundoff noise at the filter's output. Error feedback (EF) is found effective for the reduction of finite-word-length (FWL) effects in IIR digital filters, and many EF methods have been proposed in the past [1]-[10]. Alternatively, the roundoff noise can also be reduced by introducing a delta operator to IIR digital filters [11]-[13], or by adopting a new structure based on the concept of polynomial operators for digital filter implementation [14]. Another useful approach is to synthesize the state-space filter structure for the roundoff noise gain to be minimized by applying a linear transformation to state-space coordinates subject to l_2-scaling constraints [15]-[18]. As a natural extension of the aforementioned methods, efforts have been made to develop new methods that combine EF and state-space realization, for achieving better performance [19]-[21]. Separately and jointly optimized scalar or general EF matrix for state-space filters have been explored [19]. A jointly-optimized iterative algorithm with a general, diagonal, or scalar EF matrix using a quasi-Newton method has been developed for state-space digital filters [20]. A similar technique employing a quasi-Newton method is extended to two-dimensional (2-D) state-space digital filters with a general, block-diagonal, diagonal, or scalar EF matrix [21]. However, to date the use of *high-order* EF and its effect on noise-reduction performance for 1-D as well as 2-D state-space digital filters have not been explored.

In this paper, the problem of jointly optimizing *high-order* EF and state-space realization for minimizing the roundoff noise subject to l_2-scaling constraints is investigated for 1-D as well as 2-D state-space digital filters. The novelty of the proposed method has largely to do with the introduction of a *dynamic* EF that, as will be demonstrated later in the paper, is responsible for a considerable performance enhancement in noise gain reduction relative to the systems that employ static (i.e. constant) EF. We convert the constrained optimization problem encountered into an unconstrained optimization problem by using linear-algebraic techniques. An efficient quasi-Newton algorithm is utilized to solve the unconstrained optimization problem at hand. The proposed technique is applied to the case where the *high-order* EF has diagonal matrices. Case studies are presented to illustrate the validity and effectiveness of the proposed algorithm and demonstrate its performance.

Throughout, $\boldsymbol{I}_n$ denotes the identity matrix of dimension $n \times n$. The transpose (conjugate transpose) of a matrix $\boldsymbol{A}$ is indicated by $\boldsymbol{A}^T$ ($\boldsymbol{A}^*$). tr$[\boldsymbol{A}]$ is used to denote the trace of a square matrix $\boldsymbol{A}$. Moreover, bold uppercase, bold lowercase and plain lowercase are used to make the distinction between matrices, vectors and scalar values, respectively.

II. ROUNDOFF NOISE MINIMIZATION FOR 1-D STATE-SPACE DIGITAL FILTERS

A. A State-Space Model with High-Order Error Feedback

Consider a stable, controllable and observable state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$ described by

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

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.

By taking the quantizations performed before matrix-vector multiplication into account, an finite-word-length implementation of (1) with error feedforward and *high-order* EF can be

Manuscript received xxxxxx xx, 2013; revised xxxxxx xx, 2013.

T. Hinamoto and A. Doi are with Hiroshima Institute of Technology Hiroshima 731-5193, Japan (e-mails: hinamoto@ieee.org, doi@cc.it-hiroshima.ac.jp, Phone:+81-82-921-4338, Fax:+81-82-921-8978)

W.-S. Lu is with the Depertment of Electrical and Computer Engineering, University of Victoria, Victoria, B.C, Canada, V8W 3P6. (e-mail: wslu@ece.uvic.ca, Phone:+1-250-721-8692, Fax:+1-250-721-6052)

obtained as

$$\begin{aligned}\tilde{\boldsymbol{x}}(k+1) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + \boldsymbol{b}u(k) + \sum_{i=0}^{N-1}\boldsymbol{D}_i\boldsymbol{e}(k-i)\\ \tilde{y}(k) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + du(k) + \boldsymbol{h}\boldsymbol{e}(k)\end{aligned} \tag{2}$$

where $\boldsymbol{h}$ and $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ are referred to as a $1 \times n$ error-feedforward vector and $n \times n$ *high-order* EF matrices, respectively, and

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

The coefficient matrices $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}$, and d in (2) are assumed to have exact fractional B_c-bit representations. The FWL state-variable vector $\tilde{\boldsymbol{x}}(k)$ and each output $\tilde{y}(k)$ has B-bit fractional representations, while the input $u(k)$ is a $(B-B_c)$-bit fraction. The quantizer $\boldsymbol{Q}[\cdot]$ in (2) rounds the B-bit fraction $\tilde{\boldsymbol{x}}(k)$ to $(B-B_c)$-bit after the multiplications and additions, where the sign bit is not counted. It is assumed that the roundoff error $\boldsymbol{e}(k)$ can be modeled as a Gaussian random process with zero mean and covariance $\sigma^2\boldsymbol{I}_n$. Subtracting (2) from (1) yields

$$\begin{aligned}\Delta\boldsymbol{x}(k+1) &= \boldsymbol{A}\Delta\boldsymbol{x}(k) + \boldsymbol{A}\boldsymbol{e}(k) - \sum_{i=0}^{N-1}\boldsymbol{D}_i\boldsymbol{e}(k-i)\\ \Delta y(k) &= \boldsymbol{c}\Delta\boldsymbol{x}(k) + (\boldsymbol{c}-\boldsymbol{h})\boldsymbol{e}(k)\end{aligned} \tag{3}$$

where $\Delta\boldsymbol{x}(k) = \boldsymbol{x}(k) - \tilde{\boldsymbol{x}}(k)$ and $\Delta y(k) = y(k) - \tilde{y}(k)$. By taking the z-transform on both sides of (3) and setting $\Delta\boldsymbol{x}(0) = \boldsymbol{0}$, we have

$$\begin{aligned}\Delta Y(z) &= \boldsymbol{H}_e(z)\boldsymbol{E}(z)\\ \boldsymbol{H}_e(z) &= \boldsymbol{c}(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\Big(\boldsymbol{A} - \sum_{i=0}^{N-1}\boldsymbol{D}_i z^{-i}\Big)\\ &\quad + \boldsymbol{c} - \boldsymbol{h}\end{aligned} \tag{4}$$

where $\Delta Y(z)$ and $\boldsymbol{E}(z)$ represent the z-transforms of $\Delta y(k)$ and $\boldsymbol{e}(k)$, respectively.

B. A Formula for Noise Gain

Based on the model developed above, we now define the normalized noise gain $J_{e1}(\boldsymbol{h}, \boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$ in terms of the transfer function $\boldsymbol{H}_e(z)$ as

$$J_{e1}(\boldsymbol{h}, \boldsymbol{D}) = \mathrm{tr}\Big[\frac{1}{2\pi j}\oint_{|z|=1}\boldsymbol{H}_e^*(z)\boldsymbol{H}_e(z)\frac{dz}{z}\Big]. \tag{5}$$

In order to derive an easy-to-use formula to evaluate and optimize the noise gain, we write $\boldsymbol{H}_e(z)$ in (4) as

$$\boldsymbol{H}_e(z) = \sum_{k=1}^{\infty}\boldsymbol{c}\Big(\boldsymbol{A}^k - \sum_{i=0}^{N-1}\boldsymbol{A}^{k-1-i}\boldsymbol{D}_i\Big)z^{-k} + \boldsymbol{c} - \boldsymbol{h} \tag{6}$$

where $\boldsymbol{A}^i = \boldsymbol{0}$ for $i < 0$. By substituting (6) into (5) and making use of Cauchy's integral theorem, we obtain

$$\begin{aligned}&J_{e1}(\boldsymbol{h}, \boldsymbol{D})\\ &= \mathrm{tr}\Big[\boldsymbol{W}_o - \sum_{i=0}^{N-1}\Big\{(\boldsymbol{A}^T)^{i+1}\boldsymbol{W}_o\boldsymbol{D}_i + \boldsymbol{D}_i^T\boldsymbol{W}_o\boldsymbol{A}^{i+1}\Big\}\\ &\quad + \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\boldsymbol{D}_i^T\Big\{(\boldsymbol{A}^T)^{j-i}\boldsymbol{W}_o + \boldsymbol{W}_o\boldsymbol{A}^{i-j}\Big\}\boldsymbol{D}_j\\ &\quad - \sum_{i=0}^{N-1}\boldsymbol{D}_i^T\boldsymbol{W}_o\boldsymbol{D}_i - 2\boldsymbol{h}^T\boldsymbol{c} + \boldsymbol{h}^T\boldsymbol{h}\Big]\end{aligned} \tag{7}$$

where $\boldsymbol{W}_o$ is the observability Gramian of the filter in (1), that can be obtained by solving the Lyapunov equation

$$\boldsymbol{W}_o = \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}. \tag{8}$$

It is useful to note that if the high-order EF matrices $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ are diagonal, then the formula for the noise gain can be considerably simplified to

$$\begin{aligned}J_{e1}(\boldsymbol{h}, \boldsymbol{D}) &= \mathrm{tr}\Big[\boldsymbol{W}_o - \boldsymbol{c}^T\boldsymbol{c} - 2\sum_{i=0}^{N-1}\boldsymbol{W}_o\boldsymbol{A}^{i+1}\boldsymbol{D}_i\\ &\quad + \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\boldsymbol{W}_o\boldsymbol{A}^{|i-j|}\boldsymbol{D}_i\boldsymbol{D}_j\Big] + (\boldsymbol{c}-\boldsymbol{h})(\boldsymbol{c}-\boldsymbol{h})^T.\end{aligned} \tag{9}$$

It should also be noted that the l_2-scaling constraints on the state-variable vector $\boldsymbol{x}(k)$ involve the controllability Gramian $\boldsymbol{K}_c$ of the filter in (1), which can be computed by solving the Lyapunov equation

$$\boldsymbol{K}_c = \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T. \tag{10}$$

C. Problem Statement

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

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{11}$$

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

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \quad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T} \tag{12}$$

respectively. The l_2-scaling constraints are imposed on the state-variable vector $\overline{\boldsymbol{x}}(k)$ so that

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, n. \tag{13}$$

With an equivalent state-space realization as specified in (11) and assuming the use of high-order diagonal EF matrices, the normalized noise gain is found to be

$$\begin{aligned}J_{e2}(\boldsymbol{D}, \boldsymbol{T}) &= \mathrm{tr}\Big[\boldsymbol{T}^T(\boldsymbol{W}_o - \boldsymbol{c}^T\boldsymbol{c})\boldsymbol{T} - 2\sum_{i=0}^{N-1}\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{A}^{i+1}\boldsymbol{T}\boldsymbol{D}_i\\ &\quad + \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{A}^{|i-j|}\boldsymbol{T}\boldsymbol{D}_i\boldsymbol{D}_j\Big]\end{aligned} \tag{14}$$

where the noise gain is denoted as $J_{e2}(\boldsymbol{D}, \boldsymbol{T})$ to reflect the fact that the noise gain is now dependent on both high-order EF matrices $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ as well as state-space coordinate transformation $\boldsymbol{T}$. Formula (14) provides an analytic foundation for the minimization of the noise gain by *jointly* optimize the EF loop and state-space coordinate transformation. Formally, the problem being considered here can be stated as to jointly deign the high-order EF matrices $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ and coordinate transformation matrix $\boldsymbol{T}$ to minimize the noise gain $J_{e2}(\boldsymbol{D}, \boldsymbol{T})$ in (14) subject to the l_2-scaling constraints in (13), assuming the error feedforward vector $\boldsymbol{h} = \overline{\boldsymbol{c}}$ is chosen so as to eliminate the last term in (9).

D. Joint Optimization of Error Feedback and Realization

In what follows, it is assumed that the high-order EF matrices are all diagonal so that (14) is a valid objective function to be minimized. Because the constraints in (13) are nonconvex and highly nonlinear with respect to matrix $\boldsymbol{T}$, it is beneficial if these constraints can be eliminated so as to work with an unconstrained problem that admits fast Newton-like algorithms. To this end, we define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T \boldsymbol{K}_c^{-\frac{1}{2}} \tag{15}$$

which leads (13) to

$$(\hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{T}}^{-1})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, n. \tag{16}$$

These constraints are always satisfied if $\hat{\boldsymbol{T}}^{-1}$ assumes the form

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||} \right]. \tag{17}$$

Substituting (15) into (14), we obtain

$$\begin{aligned} J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) = \text{tr}\Big[& \boldsymbol{V}_0 - \hat{\boldsymbol{c}}^T \hat{\boldsymbol{c}} - 2 \sum_{p=0}^{N-1} \boldsymbol{V}_{p+1} \boldsymbol{D}_p \\ & + \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \boldsymbol{V}_{|p-q|} \boldsymbol{D}_p \boldsymbol{D}_q \Big] \end{aligned} \tag{18}$$

where

$$\boldsymbol{V}_p = \hat{\boldsymbol{T}} \boldsymbol{K}_c^{\frac{1}{2}} \boldsymbol{W}_o \boldsymbol{A}^p \boldsymbol{K}_c^{\frac{1}{2}} \hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{c}} = \boldsymbol{c} \boldsymbol{K}_c^{\frac{1}{2}} \hat{\boldsymbol{T}}^T.$$

Following the foregoing arguments, the problem of obtaining matrices $\boldsymbol{T}$ and $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ that jointly minimize (14) subject to the l_2-scaling constraints in (13) is now converted into an unconstrained optimization problem of obtaining $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ that jointly minimize $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (18).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$ and $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$. Then $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (18) is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The proposed algorithm starts with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\hat{\boldsymbol{T}} = \boldsymbol{D}_0 = \boldsymbol{D}_1 = \cdots = \boldsymbol{D}_{N-1} = \boldsymbol{I}_n$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [22]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \tag{19}$$

where

$$\begin{aligned} \boldsymbol{d}_k &= -\boldsymbol{S}_k \nabla J(\boldsymbol{x}_k), \quad \alpha_k = arg \left[\min_{\alpha} \ J(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) \right] \\ \boldsymbol{S}_{k+1} &= \boldsymbol{S}_k + \left(1 + \frac{\boldsymbol{\gamma}_k^T \boldsymbol{S}_k \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} \right) \frac{\boldsymbol{\delta}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} - \frac{\boldsymbol{\delta}_k \boldsymbol{\gamma}_k^T \boldsymbol{S}_k + \boldsymbol{S}_k \boldsymbol{\gamma}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} \\ \boldsymbol{S}_0 &= \boldsymbol{I}, \quad \boldsymbol{\delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k, \quad \boldsymbol{\gamma}_k = \nabla J(\boldsymbol{x}_{k+1}) - \nabla J(\boldsymbol{x}_k). \end{aligned}$$

Here, $\nabla J(\boldsymbol{x})$ is the gradient of $J(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, and $\boldsymbol{S}_k$ is a positive-definite approximation of the inverse Hessian matrix of $J(\boldsymbol{x}_k)$. This iteration process continues until

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon \tag{20}$$

where $\varepsilon > 0$ is a prescribed tolerance.

The implementation efficiency and solution accuracy of the quasi-Newton algorithm greatly depends on how the gradient $\nabla J(\boldsymbol{x})$ is evaluated. With high-order diagonal EF matrices, we have managed to derive closed-form formulas for computing the partial derivatives of $J(\boldsymbol{x})$ with respect to the elements of $\boldsymbol{T}$ as well as those of $\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$. The derivation process is quite technical, below we include the formulas without derivation details.

The gradients of $J(\boldsymbol{x})$ with respect to $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n$ are found to be

$$\begin{aligned} \frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} = 2\boldsymbol{e}_j^T \Big[& \boldsymbol{V}_0 - \hat{\boldsymbol{c}}^T \hat{\boldsymbol{c}} - \sum_{p=0}^{N-1} (\boldsymbol{V}_{p+1} + \boldsymbol{V}_{p+1}^T) \boldsymbol{D}_p \\ & + \frac{1}{2} \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} (\boldsymbol{V}_{|p-q|} + \boldsymbol{V}_{|p-q|}^T) \boldsymbol{D}_p \boldsymbol{D}_q \Big] \hat{\boldsymbol{T}} \boldsymbol{g}_{ij} \\ & i, j = 1, 2, \cdots, n \end{aligned} \tag{21}$$

where t_{ij} is the ith element of column vector $\boldsymbol{t}_j$ and

$$\boldsymbol{g}_{ij} = \partial \left\{ \frac{\boldsymbol{t}_j}{||\boldsymbol{t}_j||} \right\} / \partial t_{ij} = \frac{1}{||\boldsymbol{t}_j||^3} (t_{ij} \boldsymbol{t}_j - ||\boldsymbol{t}_j||^2 \boldsymbol{e}_i).$$

Let the high-order EF matrices assume the form

$$\boldsymbol{D}_p = \text{diag}\{d_{p1}, d_{p2}, \cdots, d_{pn}\} \ \text{for} \ p = 0, 1, \cdots, N-1. \tag{22}$$

The gradients of $J(\boldsymbol{x})$ with respect to the diagonal $\boldsymbol{D}_p$ are given by

$$\frac{\partial J(\boldsymbol{x})}{\partial d_{pi}} = -2 \, (\boldsymbol{V}_{p+1})_{ii} + 2 \sum_{q=0}^{N-1} d_{qi} (\boldsymbol{V}_{|p-q|})_{ii} \tag{23}$$

for $p = 0, 1, \cdots, N-1$ and $i = 1, 2, \cdots, n$.

III. ROUNDOFF NOISE MINIMIZATION FOR 2-D STATE-SPACE DIGITAL FILTERS

This section presents a solution of joint optimization of state-space coordinate transformation and higher-order EF for minimizing the noise gain for 2-D state-space digital filters. The development is in spirit parallel to that of Section II, however here one needs to deal with considerably more involved technical issues as expected.

A. A State-Space Model with High-Order Error Feedback

Let us consider a 2-D stable, separately locally controllable and separately locally observable state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m+n}$ described by the Roessor model [23]

$$\begin{aligned} \boldsymbol{x}_1(i,j) &= \boldsymbol{A} \boldsymbol{x}_0(i,j) + \boldsymbol{b} u(i,j) \\ y(i,j) &= \boldsymbol{c} \, \boldsymbol{x}_0(i,j) + d u(i,j) \end{aligned} \tag{24}$$

with

$$\boldsymbol{x}_k(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i+k,j) \\ \boldsymbol{x}^v(i,j+k) \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix}$$
$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}, \qquad \boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{bmatrix}$$

where $\boldsymbol{x}^h(i,j)$ is an $m \times 1$ horizontal state vector, $\boldsymbol{x}^v(i,j)$ is an $n \times 1$ vertical state vector, $u(i,j)$ is a scalar input, $y(i,j)$ is a scalar output, and $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\boldsymbol{A}_4$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ and d are real constant matrices of appropriate dimensions.

Like in the 1-D case, a state-space model that incorporates FWL implementation as well as error feedforward and high-order EF can be established. Assume that the quantization is performed before matrix-vector multiplications, the model is obtained as

$$\begin{aligned} \tilde{\boldsymbol{x}}_1(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}_0(i,j)] + \boldsymbol{b}u(i,j) + \sum_{k=0}^{N-1} \boldsymbol{D}_k \boldsymbol{e}_{-k}(i,j) \\ \tilde{y}(i,j) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}_0(i,j)] + du(i,j) + \boldsymbol{h}\boldsymbol{e}_0(i,j) \end{aligned} \tag{25}$$

where $\boldsymbol{h}$ is a $1 \times (m+n)$ error-feedforward vector, $\boldsymbol{D}_k$'s are $(m+n) \times (m+n)$ *high-order* EF diagonal matrices and

$$\boldsymbol{e}_{-k}(i,j) = \begin{bmatrix} \boldsymbol{e}^h(i-k,j) \\ \boldsymbol{e}^v(i,j-k) \end{bmatrix}, \quad \boldsymbol{D}_k = \begin{bmatrix} \boldsymbol{D}_{1k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{4k} \end{bmatrix}$$
$$\boldsymbol{e}_k(i,j) = \tilde{\boldsymbol{x}}_k(i,j) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}_k(i,j)], \quad \boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 \end{bmatrix}.$$

The coefficient matrices $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and d in (25) are assumed to have exact fractional B_c-bit representations. The FWL horizontal (vertical) state vector $\tilde{\boldsymbol{x}}^h(i,j)$ ($\tilde{\boldsymbol{x}}^v(i,j)$) and each output $\tilde{y}(i,j)$ has B-bit fractional representations, while the input $u(i,j)$ is a $(B-B_c)$-bit fraction. The quantizer $\boldsymbol{Q}[\cdot]$ in (25) rounds the B-bit fraction $\tilde{\boldsymbol{x}}_k(i,j)$ to $(B-B_c)$-bit after the multiplications and additions, where the sign bit is not counted. It is assumed that the roundoff error $\boldsymbol{e}_0(i,j)$ can be modeled as a Gaussian random process with zero mean and covariance $\sigma^2\boldsymbol{I}_n$. By subtracting (25) from (24), it follows that

$$\begin{aligned} \Delta\boldsymbol{x}_1(i,j) &= \boldsymbol{A}\Delta\boldsymbol{x}_0(i,j) + \boldsymbol{A}\boldsymbol{e}_0(i,j) - \sum_{k=0}^{N-1} \boldsymbol{D}_k \boldsymbol{e}_{-k}(i,j) \\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}_0(i,j) + (\boldsymbol{c}-\boldsymbol{h})\boldsymbol{e}_0(i,j) \end{aligned} \tag{26}$$

where $\Delta\boldsymbol{x}_k(i,j) = \boldsymbol{x}_k(i,j) - \tilde{\boldsymbol{x}}_k(i,j)$ for $k=0$ and 1, and $\Delta y(i,j) = y(i,j) - \tilde{y}(i,j)$. By applying the (z_1,z_2)-transform to both sides of (26) and setting the boundary conditions to be null, we obtain

$$\begin{aligned} \Delta\boldsymbol{X}_0(z_1,z_2) &= (\boldsymbol{Z}-\boldsymbol{A})^{-1}\Big(\boldsymbol{A} - \sum_{k=0}^{N-1} \boldsymbol{D}_k \boldsymbol{Z}^{-k}\Big)\boldsymbol{E}_0(z_1,z_2) \\ \Delta Y(z_1,z_2) &= \boldsymbol{c}\boldsymbol{X}_0(z_1,z_2) + (\boldsymbol{c}-\boldsymbol{h})\boldsymbol{E}_0(z_1,z_2) \end{aligned} \tag{27}$$

which yields

$$\begin{aligned} \Delta Y(z_1,z_2) &= \boldsymbol{H}_e(z_1,z_2)\boldsymbol{E}_0(z_1,z_2) \\ \boldsymbol{H}_e(z_1,z_2) &= \boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}\Big(\boldsymbol{Z} - \sum_{k=0}^{N-1} \boldsymbol{D}_k \boldsymbol{Z}^{-k}\Big) - \boldsymbol{h} \end{aligned} \tag{28}$$

where $\Delta\boldsymbol{X}_0(z_1,z_2)$, $\Delta Y(z_1,z_2)$ and $\boldsymbol{E}_0(z_1,z_2)$ are the (z_1,z_2)-transforms of signals $\Delta\boldsymbol{x}_0(i,j)$, $\Delta y(i,j)$ and $\boldsymbol{e}_0(i,j)$, respectively, and $\boldsymbol{Z} = z_1\boldsymbol{I}_m \oplus z_2\boldsymbol{I}_n$.

B. A Formula for 2-D Noise Gain

Based on the model developed above, we define the normalized noise gain $J_{e1}(\boldsymbol{h},\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$ in terms of $\boldsymbol{H}_e(z_1,z_2)$ as

$$\begin{aligned} &J_{e1}(\boldsymbol{h},\boldsymbol{D}) \\ &= \mathrm{tr}\Big[\frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1} \boldsymbol{H}_e^*(z_1,z_2)\boldsymbol{H}_e(z_1,z_2)\frac{dz_1 dz_2}{z_1 z_2}\Big] \end{aligned} \tag{29}$$

where σ_{out}^2 is noise variance at the filter's output. In order to derive an easy-to-use formula to evaluate and optimize the 2-D noise gain, we define the transition matrices $\boldsymbol{A}^{(i,j)}$ by

$$(\boldsymbol{I}_{m+n} - \boldsymbol{Z}^{-1}\boldsymbol{A})^{-1} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{A}^{(i,j)} z_1^{-i} z_2^{-j} \tag{30}$$

and notice that $\boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1} = \boldsymbol{c}(\boldsymbol{I}_{m+n} - \boldsymbol{Z}^{-1}\boldsymbol{A})^{-1}\boldsymbol{Z}^{-1}$ hence $\boldsymbol{H}_e(z_1,z_2)$ can be written as

$$\boldsymbol{H}_e(z_1,z_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{g}(i,j) z_1^{-i} z_2^{-j}\Big(\boldsymbol{Z} - \sum_{k=0}^{N-1} \boldsymbol{D}_k \boldsymbol{Z}^{-k}\Big) - \boldsymbol{h} \tag{31}$$

where

$$\begin{aligned} \boldsymbol{A}^{(0,0)} &= \boldsymbol{I}_{m+n}, \quad \boldsymbol{A}^{(i,j)} = \boldsymbol{0}, \quad i<0 \text{ or } j<0 \\ \boldsymbol{A}^{(1,0)} &= \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{A}^{(0,1)} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix} \\ \boldsymbol{A}^{(i,j)} &= \boldsymbol{A}^{(1,0)}\boldsymbol{A}^{(i-1,j)} + \boldsymbol{A}^{(0,1)}\boldsymbol{A}^{(i,j-1)} \\ &= \boldsymbol{A}^{(i-1,j)}\boldsymbol{A}^{(1,0)} + \boldsymbol{A}^{(i,j-1)}\boldsymbol{A}^{(0,1)}, \quad (i,j) > (0,0) \\ \boldsymbol{g}(i,j) &= \boldsymbol{c}\Big(\boldsymbol{A}^{(i-1,j)}\begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \boldsymbol{A}^{(i,j-1)}\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_n \end{bmatrix}\Big). \end{aligned}$$

By substituting (31) into (29) and making use of 2-D Cauchy's integral theorem, we obtain

$$\begin{aligned} J_{e1}(\boldsymbol{h},\boldsymbol{D}) = \mathrm{tr}\Big[&\boldsymbol{W}_{10} - \boldsymbol{c}_1^T\boldsymbol{c}_1 - 2\sum_{k=0}^{N-1}\boldsymbol{W}_{1,k+1}\boldsymbol{D}_{1k} \\ &+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{W}_{1,|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big] \\ &+\mathrm{tr}\Big[\boldsymbol{W}_{40} - \boldsymbol{c}_2^T\boldsymbol{c}_2 - 2\sum_{k=0}^{N-1}\boldsymbol{W}_{4,k+1}\boldsymbol{D}_{4k} \\ &+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{W}_{4,|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big] + (\boldsymbol{c}-\boldsymbol{h})(\boldsymbol{c}-\boldsymbol{h})^T \end{aligned} \tag{32}$$

where

$$\begin{aligned} \boldsymbol{W}_{1r} &= \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \end{bmatrix}\Big[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}^T(i+r,j)\boldsymbol{g}(i,j)\Big]\begin{bmatrix} \boldsymbol{I}_m \\ \boldsymbol{0} \end{bmatrix} \\ \boldsymbol{W}_{4r} &= \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_n \end{bmatrix}\Big[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}^T(i,j+r)\boldsymbol{g}(i,j)\Big]\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_n \end{bmatrix}. \end{aligned}$$

The above matrices $\boldsymbol{W}_{1r}$ and $\boldsymbol{W}_{4r}$ can be viewed as an extension of the local observability Gramian defined by

$$\begin{aligned}\boldsymbol{W}_o &= \frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{G}^*(z_1,z_2)\boldsymbol{G}(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}\\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}^T(i,j)\boldsymbol{g}(i,j).\end{aligned} \tag{33}$$

It should also be noted that l_2-scaling constraints on the local state vector $\boldsymbol{x}_0(i,j)$ involve the local controllability Gramian $\boldsymbol{K}_c$ of the filter in (24), which can be computed by

$$\begin{aligned}\boldsymbol{K}_c &= \frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{F}(z_1,z_2)\boldsymbol{F}^*(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}\\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{f}(i,j)\boldsymbol{f}^T(i,j) = \begin{bmatrix}\boldsymbol{K}_{c1} & \boldsymbol{K}_{c2}\\ \boldsymbol{K}_{c3} & \boldsymbol{K}_{c4}\end{bmatrix}\end{aligned} \tag{34}$$

where

$$\boldsymbol{f}(i,j) = \boldsymbol{A}^{(i-1,j)}\begin{bmatrix}\boldsymbol{b}_1\\ \boldsymbol{0}\end{bmatrix} + \boldsymbol{A}^{(i,j-1)}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{b}_2\end{bmatrix}.$$

C. Problem Statement

A different yet equivalent local state-space description of (24), $(\overline{\boldsymbol{A}},\overline{\boldsymbol{b}},\overline{\boldsymbol{c}},d)_n$, can be obtained via a coordinate transformation $\overline{\boldsymbol{x}}_0(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}_0(i,j)$ with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ where

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{35}$$

The l_2-scaling constraints are imposed on the local state vector $\overline{\boldsymbol{x}}_0(i,j)$ so that

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1 \quad \text{for} \quad i = 1,2,\cdots,m+n. \tag{36}$$

With an equivalent state-space realization as specified in (35) and assuming the use of high-order diagonal EF, the normalized noise gain is found to be

$$\begin{aligned}J_{e2}(\boldsymbol{D},\boldsymbol{T}) = \text{tr}\Big[&\overline{\boldsymbol{W}}_{10} - \overline{\boldsymbol{c}}_1^T\overline{\boldsymbol{c}}_1 - 2\sum_{k=0}^{N-1}\overline{\boldsymbol{W}}_{1,k+1}\boldsymbol{D}_{1k}\\ &+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\overline{\boldsymbol{W}}_{1,|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\\ +\text{tr}\Big[&\overline{\boldsymbol{W}}_{40} - \overline{\boldsymbol{c}}_2^T\overline{\boldsymbol{c}}_2 - 2\sum_{k=0}^{N-1}\overline{\boldsymbol{W}}_{4,k+1}\boldsymbol{D}_{4k}\\ &+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\overline{\boldsymbol{W}}_{4,|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]\end{aligned} \tag{37}$$

where

$$\overline{\boldsymbol{W}}_{1r} = \boldsymbol{T}_1^T\boldsymbol{W}_{1r}\boldsymbol{T}_1, \qquad \overline{\boldsymbol{W}}_{4r} = \boldsymbol{T}_4^T\boldsymbol{W}_{4r}\boldsymbol{T}_4.$$

Here, the noise gain $J_{e2}(\boldsymbol{D},\boldsymbol{T})$ is a function of both high-order EF matrices $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$ and state-space coordinate transformation $\boldsymbol{T}$. Formula (37) provides an analytic foundation for the minimization of the noise gain by jointly optimize the EF loop and state-space coordinate transformation. Formally, the problem being considered here can be stated as to jointly design the high-order EF matrices $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$ and coordinate transformation matrix $\boldsymbol{T}$ to minimize the noise gain $J_{e2}(\boldsymbol{D},\boldsymbol{T})$ in (37) subject to l_2-scaling constraints in (36), assuming the error feedforward vector $\boldsymbol{h} = \overline{\boldsymbol{c}}$ is chosen so as to eliminate the last term in (32).

D. Joint Optimization of Error Feedback and Realization

In what follows, it is assumed that the high-order EF matrices are all diagonal so that (37) is a valid objective function to be minimized. For the same reason as in the 1-D case, it turns out to be beneficial if the l_2-scaling constraints can be eliminated so as to work with an unconstrained problem that admits fast Newton-like algorithms. To this end, we define

$$\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}_1 \oplus \hat{\boldsymbol{T}}_4 = (\boldsymbol{T}_1 \oplus \boldsymbol{T}_4)^T(\boldsymbol{K}_{c1}\oplus\boldsymbol{K}_{c4})^{-\frac{1}{2}} \tag{38}$$

which leads (36) to

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1 \quad \text{for} \quad i = 1,2,\cdots,m+n. \tag{39}$$

It is noted that these constraints are always satisfied if matrices $\hat{\boldsymbol{T}}_1^{-1}$ and $\hat{\boldsymbol{T}}_4^{-1}$ assumes the form

$$\begin{aligned}\hat{\boldsymbol{T}}_1^{-1} &= \left[\frac{\boldsymbol{t}_{11}}{||\boldsymbol{t}_{11}||},\frac{\boldsymbol{t}_{12}}{||\boldsymbol{t}_{12}||},\cdots,\frac{\boldsymbol{t}_{1m}}{||\boldsymbol{t}_{1m}||}\right]\\ \hat{\boldsymbol{T}}_4^{-1} &= \left[\frac{\boldsymbol{t}_{41}}{||\boldsymbol{t}_{41}||},\frac{\boldsymbol{t}_{42}}{||\boldsymbol{t}_{42}||},\cdots,\frac{\boldsymbol{t}_{4n}}{||\boldsymbol{t}_{4n}||}\right].\end{aligned} \tag{40}$$

Substituting (38) into (37), we obtain

$$\begin{aligned}J_{e3}(\boldsymbol{D},\hat{\boldsymbol{T}}) = \text{tr}\Big[&\boldsymbol{V}_{10} - \hat{\boldsymbol{c}}_1^T\hat{\boldsymbol{c}}_1 - 2\sum_{k=0}^{N-1}\boldsymbol{V}_{1,k+1}\boldsymbol{D}_{1k}\\ &+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{V}_{1,|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\\ +\text{tr}\Big[&\boldsymbol{V}_{40} - \hat{\boldsymbol{c}}_2^T\hat{\boldsymbol{c}}_2 - 2\sum_{k=0}^{N-1}\boldsymbol{V}_{4,k+1}\boldsymbol{D}_{4k}\\ &+ \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{V}_{4,|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]\end{aligned} \tag{41}$$

where

$$\begin{aligned}\boldsymbol{V}_{1r} &= \hat{\boldsymbol{T}}_1\boldsymbol{K}_{c1}^{\frac{1}{2}}\boldsymbol{W}_{1r}\boldsymbol{K}_{c1}^{\frac{1}{2}}\hat{\boldsymbol{T}}_1^T, \qquad \hat{\boldsymbol{c}}_1 = \boldsymbol{c}_1\boldsymbol{K}_{c1}^{\frac{1}{2}}\hat{\boldsymbol{T}}_1^T\\ \boldsymbol{V}_{4r} &= \hat{\boldsymbol{T}}_4\boldsymbol{K}_{c4}^{\frac{1}{2}}\boldsymbol{W}_{4r}\boldsymbol{K}_{c4}^{\frac{1}{2}}\hat{\boldsymbol{T}}_4^T, \qquad \hat{\boldsymbol{c}}_2 = \boldsymbol{c}_2\boldsymbol{K}_{c4}^{\frac{1}{2}}\hat{\boldsymbol{T}}_4^T.\end{aligned}$$

Following the foregoing arguments, the problem of obtaining $\boldsymbol{T}$ and $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$ that jointly minimize (37) subject to the l_2-scaling constraints in (36) is now converted into an unconstrained optimization problem of obtaining $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$ that jointly minimize $J_{e3}(\boldsymbol{D},\hat{\boldsymbol{T}})$ in (41).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1},\boldsymbol{t}_{11},\boldsymbol{t}_{12},\cdots,\boldsymbol{t}_{1m},\boldsymbol{t}_{41},\boldsymbol{t}_{42},\cdots,\boldsymbol{t}_{4n}$. Then $J_{e3}(\boldsymbol{D},\hat{\boldsymbol{T}})$ in (41) is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The proposed algorithm starts with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\boldsymbol{T} = \boldsymbol{D}_0 = \boldsymbol{D}_1 = \cdots = \boldsymbol{D}_{N-1} = \boldsymbol{I}_{m+n}$, updates vector $\boldsymbol{x}_k$ in the kth iteration by using (19) and terminates with a criterion stated in (20).

Relative to its 1-D counterpart, the availability of a closed-form expression for the gradient $\nabla J(\boldsymbol{x})$ is even more critical because of the involvement of considerably more design variables in the 2-D case. With diagonal high-order EF matrices, closed-form formulas for computing the partial derivatives of $J(\boldsymbol{x})$ with respect to the elements of $\boldsymbol{T}$ as well the elements of

$\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$ are given below. Derivation details are omitted due to space limitation.

The gradients of $J(\boldsymbol{x})$ with respect to $\boldsymbol{t}_{11}, \boldsymbol{t}_{12}, \cdots, \boldsymbol{t}_{1m}$, $\boldsymbol{t}_{41}, \boldsymbol{t}_{42}, \cdots, \boldsymbol{t}_{4n}$ are found to be

$$\begin{aligned}\frac{\partial J(\boldsymbol{x})}{\partial t_{1ij}} &= 2\boldsymbol{e}_{1j}^T\Big[\boldsymbol{V}_{10} - \hat{\boldsymbol{c}}_1^T\hat{\boldsymbol{c}}_1 - \sum_{k=0}^{N-1}(\boldsymbol{V}_{1,k+1} + \boldsymbol{V}_{1,k+1}^T)\boldsymbol{D}_{1k} \\ &\quad + \frac{1}{2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}(\boldsymbol{V}_{1,|k-l|} + \boldsymbol{V}_{1,|k-l|}^T)\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\hat{\boldsymbol{T}}_1\boldsymbol{g}_{1ij} \\ \frac{\partial J(\boldsymbol{x})}{\partial t_{4ij}} &= 2\boldsymbol{e}_{4j}^T\Big[\boldsymbol{V}_{40} - \hat{\boldsymbol{c}}_2^T\hat{\boldsymbol{c}}_2 - \sum_{k=0}^{N-1}(\boldsymbol{V}_{4,k+1} + \boldsymbol{V}_{4,k+1}^T)\boldsymbol{D}_{4k} \\ &\quad + \frac{1}{2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}(\boldsymbol{V}_{4,|k-l|} + \boldsymbol{V}_{4,|k-l|}^T)\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]\hat{\boldsymbol{T}}_4\boldsymbol{g}_{4ij}\end{aligned} \tag{42}$$

where t_{1ij} (t_{4ij}) is the ith element of vector $\boldsymbol{t}_{1j}$ ($\boldsymbol{t}_{4j}$) and

$$\begin{aligned}\boldsymbol{g}_{1ij} &= \partial\left\{\frac{\boldsymbol{t}_{1j}}{||\boldsymbol{t}_{1j}||}\right\}/\partial t_{1ij} = \frac{1}{||\boldsymbol{t}_{1j}||^3}(t_{1ij}\boldsymbol{t}_{1j} - ||\boldsymbol{t}_{1j}||^2\boldsymbol{e}_{1i}) \\ \boldsymbol{g}_{4ij} &= \partial\left\{\frac{\boldsymbol{t}_{4j}}{||\boldsymbol{t}_{4j}||}\right\}/\partial t_{4ij} = \frac{1}{||\boldsymbol{t}_{4j}||^3}(t_{4ij}\boldsymbol{t}_{4j} - ||\boldsymbol{t}_{4j}||^2\boldsymbol{e}_{4i})\end{aligned}$$

with $1 \le i, j \le m$ ($1 \le i, j \le n$) for t_{1ij}, $\boldsymbol{g}_{1ij}$ and $\boldsymbol{e}_{1i}$ (t_{4ij}, $\boldsymbol{g}_{4ij}$ and $\boldsymbol{e}_{4i}$).

Let the high-order EF matrices assume the form

$$\begin{aligned}\boldsymbol{D}_{1k} &= \mathrm{diag}\{\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{km}\} \\ \boldsymbol{D}_{4k} &= \mathrm{diag}\{\beta_{k1}, \beta_{k2}, \cdots, \beta_{kn}\}\end{aligned} \tag{43}$$

for $k = 0, 1, \ldots, N-1$. The gradients of $J(\boldsymbol{x})$ with respect to the diagonal $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{4k}$ are given by

$$\begin{aligned}\frac{\partial J(\boldsymbol{x})}{\partial \alpha_{kp}} &= -2\,(\boldsymbol{V}_{1,k+1})_{pp} + 2\sum_{l=0}^{N-1}\alpha_{lp}(\boldsymbol{V}_{1,|k-l|})_{pp} \\ \frac{\partial J(\boldsymbol{x})}{\partial \beta_{kq}} &= -2\,(\boldsymbol{V}_{4,k+1})_{qq} + 2\sum_{l=0}^{N-1}\beta_{lq}(\boldsymbol{V}_{4,|k-l|})_{qq}\end{aligned} \tag{44}$$

for $k = 0, 1, \cdots, N-1$, $p = 1, 2, \cdots, m$ and $q = 1, 2, \cdots, n$.

IV. CASE STUDIES

In this section, we present two sets of simulation results, one for the 1-D case and the other for the 2-D case. The simulations serve the purpose of demonstrating the numerical efficiency of the proposed joint optimization algorithms and effectiveness of using a high-order (i.e. dynamic) EF relative to a constant (i.e. static) FE for performance enhancement.

Example 1: Consider a stable low-pass state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_3$ described by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.339377 & -1.152652 & 1.520167 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} 0 & 0 & 0 & 0.437881 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0.212964 & 0.293733 & 0.718718 \end{bmatrix}$$

$$d = 6.59592 \times 10^{-2}.$$

The controllability and observability Gramians $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$ of the above filter were computed from (10) and (8) as

$$\boldsymbol{K}_c = \begin{bmatrix} 1.000000 & 0.741988 & 0.227107 \\ 0.741988 & 1.000000 & 0.741988 \\ 0.227107 & 0.741988 & 1.000000 \end{bmatrix}$$

$$\boldsymbol{W}_o = \begin{bmatrix} 0.720426 & -1.635144 & 1.753511 \\ -1.635144 & 4.551539 & -4.194133 \\ 1.753511 & -4.194133 & 5.861185 \end{bmatrix}.$$

The noise gain of the filter without feedforward and EF was then computed from (9) as

$$J_{e1}(\boldsymbol{0}, \boldsymbol{0}) = 11.133150.$$

In what follows, we report simulation results for two system set-up's—the first employed a static EF (i.e. $N = 1$) while the second used a dynamic EF with $N = 2$. In both cases the EF matrices involved were all diagonal. In the case of $N = 1$, the quasi-Newton algorithm was applied to minimize (18) with tolerance $\varepsilon = 10^{-8}$ in (20). It took the algorithm 64 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 5.899503 & -0.516441 & 1.228943 \\ 5.935833 & -0.493021 & 2.377971 \\ 4.442394 & -1.380516 & 1.478971 \end{bmatrix}$$

$$\boldsymbol{D}_0 = \mathrm{diag}\{\ 0.390401 \quad 0.574984 \quad 0.659618\ \}.$$

The minimized noise gain was found to be

$$J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 0.273955.$$

The profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 64 iterations of the algorithm is dipicted in Fig. 1.

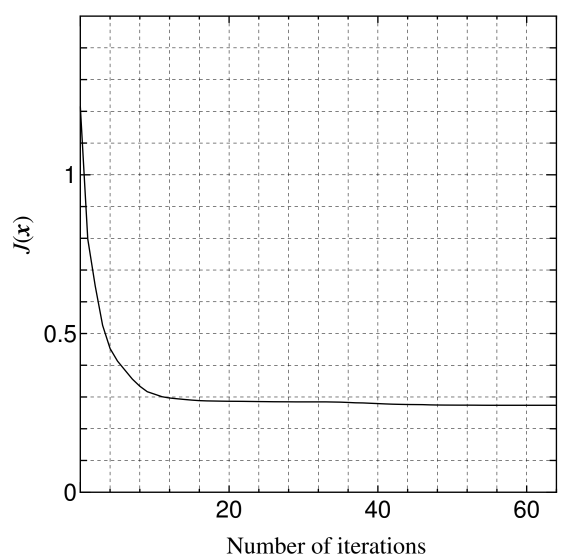

Fig. 1. Profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 64 iterations.

In the case of $N = 2$, the quasi-Newton algorithm was applied to minimize (18) with tolerance $\varepsilon = 10^{-8}$ in (20). It took the algorithm 36 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.367924 & -0.014714 & 0.386079 \\ -0.849858 & 1.032268 & 0.308031 \\ 0.078696 & 0.047338 & 1.218360 \end{bmatrix}$$

$$\boldsymbol{D}_0 = \mathrm{diag}\{\ 0.478043 \quad 0.923417 \quad 1.187428\ \}$$

$$\boldsymbol{D}_1 = \mathrm{diag}\{\ -0.060347 \quad -0.622454 \quad -0.584187\ \}.$$

The minimized noise gain in this case was found to be

$$J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 0.031801$$

which is approximately nine times smaller than what the best static EF can achieve. The profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 36 iterations of the algorithm is dipicted in Fig. 2.

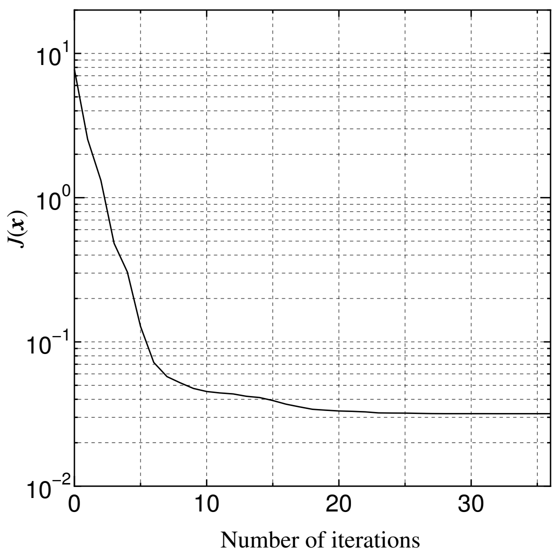

Fig. 2. Profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 36 iterations.

The above and several other simulation results regarding the noise gain $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (18) plus $(\hat{\boldsymbol{c}} - \boldsymbol{h})(\hat{\boldsymbol{c}} - \boldsymbol{h})^T$ were summarized in Table I where the column with "Infinite Precision" shows the value of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ derived from the optimal $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}$. The column with "3-Bit Quantization" means that of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) + (\hat{\boldsymbol{c}} - \boldsymbol{h})(\hat{\boldsymbol{c}} - \boldsymbol{h})^T$ where each entry of the optimal $\boldsymbol{D}$ was rounded to a power-of-two representation with 3 bits after the binary point, matrix $\hat{\boldsymbol{T}}$ was updated so as to minimize $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ with such a $\boldsymbol{D}$ fixed, and each entry of $\hat{\boldsymbol{c}}$ derived from the updated $\hat{\boldsymbol{T}}$ was rounded to a power-of-two representation with 3 bits after the binary point to set a new vector $\boldsymbol{h}$.

In order to evaluate performance improvement by the proposed joint optimization strategy versus conventional separate optimization technique where one finds an optimal coordinate transformation $\boldsymbol{T}$ first, then, with $\boldsymbol{T}$ fixed, finds on optimal EF matrix $\boldsymbol{D}$ as well as a feedforward vector $\boldsymbol{h}$, in each of the cases (with " infinite precision " and ' 3-bit quantization "), Table I includes simulation results obtained using separate optimization and joint optimization, respectively, where each optimization procedure was applied with static ($N = 1$) and dynamic ($N = 2$) EF. On comparing the results given in Table I, it is quite clear that the use of a dynamic EF (i.e. with $N > 1$), when jointly optimized with state-space realization, can improve the performance of round-off noise reduction in a significant manner relative to those achieved by either separate optimization or joint optimization with static EF

TABLE I
PERFORMANCE COMPARISON

N	Optimization	Infinite Precision	3-Bit Quantization
1	Separate	0.644452	0.659679
	Joint	0.273955	0.295692
2	Separate	0.353326	0.373479
	Joint	0.031801	0.046057

Example 2: Consider a stable low-pass 2-D state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{2+2}$ described by

$$\boldsymbol{A} = \begin{bmatrix} 1.88899 & -0.91219 & -1.00000 & 0 \\ 1 & 0 & 0 & 0 \\ 0.02771 & -0.02580 & 1.88899 & 1 \\ -0.02580 & 0.02431 & -0.91219 & 0 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} 0.219089 & 0 & -0.028889 & 0.091219 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0.028889 & -0.091219 & -0.219089 & 0 \end{bmatrix}$$

$$d = 0.87021.$$

The local controllability Gramian $\boldsymbol{K}_c$ and the local observability Gramian $\boldsymbol{W}_o$ of the above filter were computed from (34) and (33) with truncation $(0, 0) \le (i, j) \le (500, 500)$ as

$$\boldsymbol{K}_c = 10 \begin{bmatrix} 8.717246 & 8.525727 & 0.163982 & -0.153908 \\ 8.525727 & 8.717246 & 0.132122 & -0.123319 \\ 0.163982 & 0.132122 & 0.113447 & -0.103553 \\ -0.153908 & -0.123319 & -0.103553 & 0.097348 \end{bmatrix}$$

$$\boldsymbol{W}_o = 10 \begin{bmatrix} 0.113447 & -0.103553 & 0.163982 & 0.132122 \\ -0.103553 & 0.097348 & -0.153908 & -0.123319 \\ 0.163982 & -0.153908 & 8.717246 & 8.525727 \\ 0.132122 & -0.123319 & 8.525727 & 8.717246 \end{bmatrix}$$

By using

$$\boldsymbol{T}_o = \mathrm{diag}\{9.336620 \quad 9.336620 \quad 1.065113 \quad 0.986652\},$$

the l_2-scaling constraints

$$(\boldsymbol{T}_o^{-1} \boldsymbol{K}_c \boldsymbol{T}_o^{-T})_{ii} = 1 \quad \text{for} \quad i = 1, 2, 3, 4,$$

are satisfied and the noise gain of the filter without feedforward and EF was computed from (32) as

$$\overline{J}_{e1}(\boldsymbol{0}, \boldsymbol{0}) = \mathrm{tr}[\boldsymbol{T}_o^T \boldsymbol{W}_o \boldsymbol{T}_o] = 367.51003.$$

Similar to the simulations for the 1-D filter in Example 1, separate and joint optimization were carried out, and in each optimization setting static ($N = 1$) and dynamic ($N = 2$) EF were examined with infinite precision and 3-bit quantization, respectively.

With the EF order set to $N = 1$, the quasi-Newton algorithm was applied to minimize (41) with tolerance $\varepsilon = 10^{-8}$ in (20). It took the algorithm 11 iterations to converge to the solution

$$\hat{\boldsymbol{T}}_1 = \begin{bmatrix} 1.033730 & 0.089870 \\ 0.189249 & 1.020225 \end{bmatrix} \oplus \begin{bmatrix} 1.125495 & 0.413101 \\ 0.276292 & 1.166642 \end{bmatrix}$$

$$\boldsymbol{D}_0 = \text{diag}\{\, 0.978164 \quad 0.912805 \quad 0.949943 \quad 0.895449 \,\}.$$

The minimized noise gain in this case was found to be

$$J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 1.250326.$$

The profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 11 iterations of the algorithm is dipicted in Fig. 3.

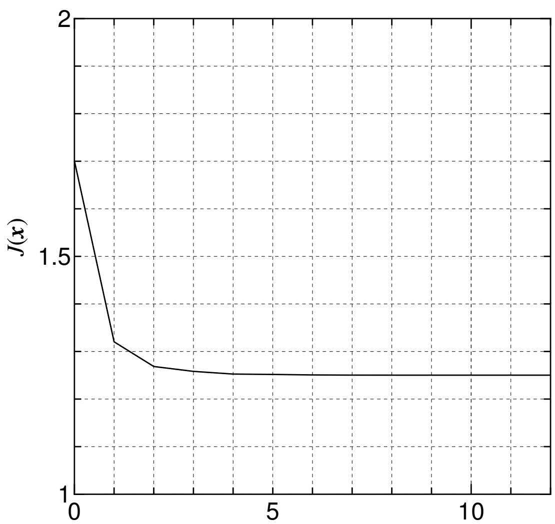

Fig. 3. Profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 11 iterations.

With the EF order set to $N = 2$, the quasi-Newton algorithm was applied to minimize (41) with tolerance $\varepsilon = 10^{-8}$ in (20). It took the algorithm 20 iterations to converge to the solution

$$\hat{\boldsymbol{T}}_1 = \begin{bmatrix} 0.949460 & -0.381269 \\ 0.573481 & 0.847325 \end{bmatrix} \oplus \begin{bmatrix} 1.187712 & 0.543967 \\ 0.347831 & 1.259196 \end{bmatrix}$$

$$\boldsymbol{D}_0 = \text{diag}\{\, 1.904500 \quad 1.055826 \quad 1.348885 \quad 1.056108 \,\}$$

$$\boldsymbol{D}_1 = -\,\text{diag}\{\, 0.928480 \quad 0.141394 \quad 0.408198 \quad 0.167253 \,\}.$$

The minimized noise gain in this case was found to be

$$J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}}) = 1.055408.$$

The profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 20 iterations of the algorithm is dipicted in Fig. 4.

These and other simulation results regarding the noise gain $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ in (41) plus $(\hat{\boldsymbol{c}} - \boldsymbol{h})(\hat{\boldsymbol{c}} - \boldsymbol{h})^T$ were summarized in Table II. From this table, it is observed that joint optimization produces a lower noise gain as compared to separate optimization. Moreover, the minimized noise gain in a dynamic EF with $N = 2$ is considerably smaller than what the system can achieve with a static EF (i.e. $N = 1$).

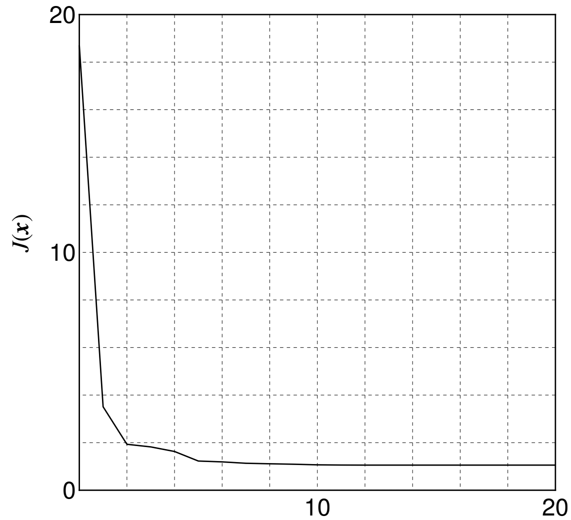

Fig. 4. Profile of $J_{e3}(\boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 20 iterations.

TABLE II
PERFORMANCE COMPARISON

N	Optimization	Infinite Precision	3-Bit Quantization
1	Separate	1.446145	1.483189
	Joint	1.250326	1.274629
2	Separate	1.320965	1.378140
	Joint	1.055408	1.085026

5. CONCLUSION

For 1-D and 2-D state-space digital filters, the problem of jointly optimizing *high-order* EF and realization to minimize the effects of roundoff noise at the filter output subject to l_2-scaling constraints has been investigated. It has been shown that the problem at hand can be converted into an unconstrained optimization problem by using linear algebraic techniques. Closed-form formulas for fast evaluation of the gradient of the objective function have been derived. Then an efficient quasi-Newton algorithm has been employed to solve the unconstrained optimization problem. The proposed techniques have been developed for the case where the *high-order* EF matrices are diagonal. The simulation results of case studies have demonstrated the validity and effectiveness of the proposed techniques.

REFERENCES

[1] H. A. Spang, III and P. M. Shultheiss, “Reduction of quantizing noise by use of feedback,” *IRE Trans. Commun. Syst.*, vol. CS-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. 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. 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. 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. 30, pp. 963-973, Dec. 1982.

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

[7] W. E. Higgins and D. C. Munson, “Optimal and suboptimal error-spectrum shaping for cascade-form digital filters,” *IEEE Trans. Circuits Syst.*, vol. 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. 32, pp. 88-92, Jan. 1985.

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

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

[12] D. Williamson, “Delay replacement in direct form structures”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 453-460, Apr. 1988.

[13] M. M. Ekanayake and K. Premaratne, “Two-dimensional delta-operator formulated discrete-time systems: Analysis and synthesis of minimum roundoff noise realizations,” *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS’96), vol. 2, pp. 213-216, May 1996.

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

[15] S. Y. Hwang, “Roundoff noise in state-space digital filtering: A general analysis,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 256-262, June 1976.

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

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

[18] 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. 26, pp. 149-153, Mar. 1979.

[19] T. Hinamoto, H. Ohnishi and W.-S. Lu, “Roundoff noise minimization of state-space digital filters using separate and joint error feedback/coordinate transformation,” *IEEE Trans. Circuits Syst. I*, vol. 50, pp. 23-33, Jan. 2003.

[20] W.-S. Lu and T. Hinamoto, “Jointly optimized error-feedback and realization for roundoff noise minimization in state-space digital filters,” *IEEE Trans. Signal Processing*, vol. 53, pp. 2135-2145, June 2005.

[21] T. Hinamoto, H. Ohnishi and W.-S. Lu, “Roundoff noise minimization for 2-D state-space digital filters using jointly optimization of error-feedback and realization,” *IEEE Trans. Signal Processing*, vol. 54, pp. 4302-4310, Nov. 2006.

[22] R. Fletcher, Practical Methods of Optimization, 2nd ed. Wiley, New York, 1987.

[23] R. P. Roesser, “A discrete state-space model for linear image processing,” *IEEE Trans. Automat. Contr.*, vol. 20, pp. 1-10, Feb. 1975.

[24] T. Kailath, *Linear Systems*, Prentice Hall, 1980.