Roundoff Noise Reduction in 2-D State-Space Digital Filters Using High-Order Error Feedback and Realization

Takao Hinamoto and Akimitsu Doi
Hiroshima Institute of Technology
Hiroshima 731-5193, Japan
Email: hinamoto@ieee.org doi@cc.it-hiroshima.ac.jp

Wu-Sheng Lu
University of Victoria
Victoria, BC, Canada V8W 3P6
Email: wslu@ece.uvic.ca

***Abstract*— A technique for reducing roundoff noise effects subject to l_2-scaling constraints is developed for two-dimensional state-space digital filters by means of high-order error feedback and state-space realization. First, the roundoff noise gain is mimimized subject to l_2-scaling constraints by choosing the coordinate transformation matrices appropriately. Optimal high-order error-feedback matrices are then determined so as to minimize the roundoff noise gain in the two-dimensional state-space digital filter obtained via the coordinate transformation matrices. Finally a numerical example is presented to illustrate the utility of the proposed technique.**

I. INTRODUCTION

When implementing IIR digital filters with fixed-point arithmetic, it is important to reduce the effects of roundoff noise at the filter's output. Error feedback (EF) is found effective for reducing finite-word-length (FWL) effects in IIR digital filters, and many EF methods have been proposed in the past [1]-[10]. An alternative useful approach is to synthesize the state-space model for the roundoff noise gain to be minimized by applying a linear transformation to state-space coordinates subject to l_2-scaling constraints [11]-[14]. 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 [15]-[17]. Separately and jointly optimized scalar or general EF matrix for state-space filters have been explored [15]. 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 [16]. 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 [17]. However, to date the use of *high-order* EF and its effect on noise-reduction performance for 2-D state-space digital filters have not been explored.

In this paper, the problem of separately optimizing *high-order* EF and state-space realization for minimizing the roundoff noise subject to l_2-scaling constraints is investigated for 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 performance enhancement in noise gain reduction relative to the systems that employ static (i.e. constant) EF.

II. PROBLEM FORMULATION

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 [18]

$$\begin{aligned} \boldsymbol{x}_1(i,j) &= \boldsymbol{A}\boldsymbol{x}_0(i,j) + \boldsymbol{b}u(i,j) \\ y(i,j) &= \boldsymbol{c}\,\boldsymbol{x}_0(i,j) + du(i,j) \end{aligned} \tag{1}$$

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.

Assuming that the quantization is performed before matrix-vector multiplication, an actual FWL implementation of model (1) with error feedforward and *high-order* EF is obtained as

$$\begin{aligned} \tilde{\boldsymbol{x}}_1(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}_0(i,j)] + \boldsymbol{b}u(i,j) + \sum_{k=0}^{N-1} \boldsymbol{D}_k \boldsymbol{e}_{-k}(i,j) \\ \tilde{y}(i,j) &= \boldsymbol{c}\,\boldsymbol{Q}[\tilde{\boldsymbol{x}}_0(i,j)] + du(i,j) + \boldsymbol{h}\boldsymbol{e}_0(i,j) \end{aligned} \tag{2}$$

where $\boldsymbol{h}$ is a $1 \times (m+n)$ error-feedforward vector, $\boldsymbol{D}_k$ is a $(m+n) \times (m+n)$ *diagonal high-order* EF matrix 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 (2) 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 the output $\tilde{y}(i,j)$ all have B-bit fractional representations, while the input $u(i,j)$ is a $(B-B_c)$-bit fraction. The quantizer $\boldsymbol{Q}[\cdot]$

in (2) 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 zero-mean stochastic process with covariance $\sigma^2 \boldsymbol{I}_n$. By subtracting (2) from (1), it follows that

$$\begin{aligned}\Delta\boldsymbol{x}_1(i,j) &= \boldsymbol{A}\Delta\boldsymbol{x}_0(i,j) + \boldsymbol{A}\boldsymbol{e}_0(i,j) - \sum_{k=0}^{N-1}\boldsymbol{D}_k\boldsymbol{e}_{-k}(i,j)\\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}_0(i,j) + (\boldsymbol{c}-\boldsymbol{h})\boldsymbol{e}_0(i,j)\end{aligned} \tag{3}$$

where $\Delta\boldsymbol{x}_k(i,j) = \boldsymbol{x}_k(i,j) - \tilde{\boldsymbol{x}}_k(i,j)$ for $k = 0,1$ and $\Delta y(i,j) = y(i,j) - \tilde{y}(i,j)$. By taking the (z_1,z_2)-transform on both sides of (3) and setting the boundary conditions to be null, we obtain

$$\begin{aligned}\Delta Y(z_1,z_2) &= \boldsymbol{H}_e(z_1,z_2)\boldsymbol{E}_0(z_1,z_2)\\ \boldsymbol{H}_e(z_1,z_2) &= \boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}\Big(\boldsymbol{Z} - \sum_{k=0}^{N-1}\boldsymbol{D}_k\boldsymbol{Z}^{-k}\Big) - \boldsymbol{h}\end{aligned} \tag{4}$$

where $\Delta Y(z_1,z_2)$ and $\boldsymbol{E}_0(z_1,z_2)$ are (z_1,z_2)-transforms of $\Delta y(i,j)$ and $\boldsymbol{e}_0(i,j)$, respectively, and $\boldsymbol{Z} = z_1\boldsymbol{I}_m \oplus z_2\boldsymbol{I}_n$. Define the transition matrix $\boldsymbol{A}^{(i,j)}$ by

$$(\boldsymbol{I}_{m+n} - \boldsymbol{Z}^{-1}\boldsymbol{A})^{-1} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{A}^{(i,j)}z_1^{-i}z_2^{-j}. \tag{5}$$

Since $\boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1} = \boldsymbol{c}(\boldsymbol{I}_{m+n} - \boldsymbol{Z}^{-1}\boldsymbol{A})^{-1}\boldsymbol{Z}^{-1}$, the noise transfer function $\boldsymbol{H}_e(z_1,z_2)$ can be written as

$$\boldsymbol{H}_e(z_1,z_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}(i,j)z_1^{-i}z_2^{-j}\Big(\boldsymbol{Z} - \sum_{k=0}^{N-1}\boldsymbol{D}_k\boldsymbol{Z}^{-k}\Big) - \boldsymbol{h} \tag{6}$$

where

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

New we define the normalized noise gain $J_{e1}(\boldsymbol{h},\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D} = [\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}]$ as

$$\begin{aligned}&J_{e1}(\boldsymbol{h},\boldsymbol{D})\\ &= \mathrm{tr}\Big[\frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{H}_e^*(z_1,z_2)\boldsymbol{H}_e(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}\Big]\end{aligned} \tag{7}$$

where σ_{out}^2 is noise variance at the filter output. Substituting (6) into (7), it follows that

$$\begin{aligned}J_{e1}(\boldsymbol{h},\boldsymbol{D}) &= \mathrm{tr}\Big[\boldsymbol{W}_{10} - \boldsymbol{c}_1^T\boldsymbol{c}_1 - 2\sum_{k=0}^{N-1}\boldsymbol{W}_{1,k+1}\boldsymbol{D}_{1k}\\ &\quad + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{W}_{1,|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\\ &\quad + \mathrm{tr}\Big[\boldsymbol{W}_{40} - \boldsymbol{c}_2^T\boldsymbol{c}_2 - 2\sum_{k=0}^{N-1}\boldsymbol{W}_{4,k+1}\boldsymbol{D}_{4k}\\ &\quad + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{W}_{4,|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big] + (\boldsymbol{c}-\boldsymbol{h})(\boldsymbol{c}-\boldsymbol{h})^T\end{aligned} \tag{8}$$

where

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

It should 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 (1) which can be computed by

$$\begin{aligned}\boldsymbol{K}_c &= \frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{F}(z_1,z_2)\boldsymbol{F}^*(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}\\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{f}(i,j)\boldsymbol{f}^T(i,j) = \begin{bmatrix}\boldsymbol{K}_{c1} & \boldsymbol{K}_{c2}\\ \boldsymbol{K}_{c3} & \boldsymbol{K}_{c4}\end{bmatrix}\end{aligned} \tag{9}$$

where

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

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

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{10}$$

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

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1, \quad i = 1,2,\cdots,m+n. \tag{11}$$

The problem being considered here is to design an optimal coordinate transformation matrix $\boldsymbol{T}$ as well as *diagonal high-order* EF matrices $\boldsymbol{D}_0,\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{N-1}$ that separately minimize the noise gain

$$\begin{aligned}J_{e2}(\overline{\boldsymbol{c}},\boldsymbol{D},\boldsymbol{T}) &= \mathrm{tr}\Big[\overline{\boldsymbol{W}}_{10} - \overline{\boldsymbol{c}}_1^T\overline{\boldsymbol{c}}_1 - 2\sum_{k=0}^{N-1}\overline{\boldsymbol{W}}_{1,k+1}\boldsymbol{D}_{1k}\\ &\quad + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\overline{\boldsymbol{W}}_{1,|k-l|}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]\\ &\quad + \mathrm{tr}\Big[\overline{\boldsymbol{W}}_{40} - \overline{\boldsymbol{c}}_2^T\overline{\boldsymbol{c}}_2 - 2\sum_{k=0}^{N-1}\overline{\boldsymbol{W}}_{4,k+1}\boldsymbol{D}_{4k}\\ &\quad + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\overline{\boldsymbol{W}}_{4,|k-l|}\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]\Big]\end{aligned} \tag{12}$$

subject to l_2-scaling constraints in (11) where

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

and the error feedforward vector $\boldsymbol{h}$ is chosen as $\boldsymbol{h} = \overline{\boldsymbol{c}}$.

III. SEPARATE OPTIMIZATION OF HIGH-ORDER ERROR FEEDBACK AND REALIZATION

The separate minimization is carried out in two major steps. First, we fix the EF matrices to $\boldsymbol{D}_i = \boldsymbol{0}$ for $i = 0, 1, \cdots, N-1$ in (12) so that the objective function is reduced to

$$J_{e2}(\overline{\boldsymbol{c}}, \boldsymbol{0}, \boldsymbol{T}) = \mathrm{tr}[\overline{\boldsymbol{W}}_{10} - \overline{\boldsymbol{c}}_1^T \overline{\boldsymbol{c}}_1] + \mathrm{tr}[\overline{\boldsymbol{W}}_{40} - \overline{\boldsymbol{c}}_2^T \overline{\boldsymbol{c}}_2] \tag{13}$$

which is minimized with respect to matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ subject to the l_2-scaling constraints in (11). Second, with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ optimized in the first step, (12) is minimized under the fixed $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ with respect to matrices $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$.

To perform the first step, we define the Lagrange functions

$$\begin{aligned} J_1(\boldsymbol{P}_1, \lambda_1) &= \mathrm{tr}[\boldsymbol{V}_1 \boldsymbol{P}_1] - \lambda_1 \left(\mathrm{tr}[\boldsymbol{K}_{c1} \boldsymbol{P}_1]^{-1} - m\right) \\ J_2(\boldsymbol{P}_4, \lambda_4) &= \mathrm{tr}[\boldsymbol{V}_4 \boldsymbol{P}_4] - \lambda_4 \left(\mathrm{tr}[\boldsymbol{K}_{c4} \boldsymbol{P}_4]^{-1} - n\right) \end{aligned} \tag{14}$$

where

$$\begin{aligned} \boldsymbol{P}_1 &= \boldsymbol{T}_1 \boldsymbol{T}_1^T, \qquad & \boldsymbol{V}_1 &= \boldsymbol{W}_{10} - \boldsymbol{c}_1^T \boldsymbol{c}_1 \\ \boldsymbol{P}_4 &= \boldsymbol{T}_4 \boldsymbol{T}_4^T, \qquad & \boldsymbol{V}_4 &= \boldsymbol{W}_{40} - \boldsymbol{c}_2^T \boldsymbol{c}_2. \end{aligned}$$

The optimal coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ can be determined by solving the equations

$$\begin{aligned} \frac{\partial J_1(\boldsymbol{P}_1, \lambda_1)}{\partial \boldsymbol{P}_1} &= \boldsymbol{0}, \qquad \frac{\partial J_1(\boldsymbol{P}_1, \lambda_1)}{\partial \lambda_1} = 0 \\ \frac{\partial J_2(\boldsymbol{P}_4, \lambda_4)}{\partial \boldsymbol{P}_4} &= \boldsymbol{0}, \qquad \frac{\partial J_2(\boldsymbol{P}_4, \lambda_4)}{\partial \lambda_4} = 0 \end{aligned} \tag{15}$$

which is led to [15]

$$\begin{aligned} \boldsymbol{T}_1 &= \frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} \theta_{1i} \right)^{\frac{1}{2}} \mathbf{V}_1^{-\frac{1}{2}} \left(\mathbf{V}_1^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{V}^{\frac{1}{2}} \right)^{\frac{1}{4}} \boldsymbol{U}_1 \\ \boldsymbol{T}_4 &= \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \theta_{4i} \right)^{\frac{1}{2}} \mathbf{V}_4^{-\frac{1}{2}} \left(\mathbf{V}_4^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{V}_4^{\frac{1}{2}} \right)^{\frac{1}{4}} \boldsymbol{U}_4 \end{aligned} \tag{16}$$

where $\theta_{11}^2, \theta_{12}^2, \cdots, \theta_{1m}^2$ are the eigenvalues of $\boldsymbol{K}_{c1} \boldsymbol{W}_{c1}$, $\theta_{41}^2, \theta_{42}^2, \cdots, \theta_{4n}^2$ are the eigenvalues of $\boldsymbol{K}_{c4} \boldsymbol{W}_{c4}$, $\boldsymbol{U}_1$ is an arbitrary $m \times m$ orthogonal matrix and $\boldsymbol{U}_4$ is an arbitrary $n \times n$ orthogonal matrix.

Suppose that the eigenvalue-eigenvector decompositions of $\left(\boldsymbol{V}_1^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{V}_1^{\frac{1}{2}}\right)^{\frac{1}{2}}$ and $\left(\boldsymbol{V}_4^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{V}_4^{\frac{1}{2}}\right)^{\frac{1}{2}}$ can be written as

$$\begin{aligned} \left(\boldsymbol{V}_1^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{V}_1^{\frac{1}{2}}\right)^{\frac{1}{2}} &= \boldsymbol{R}_1 \, \mathrm{diag}\{\theta_{11}, \theta_{12}, \cdots, \theta_{1m}\} \boldsymbol{R}_1^T \\ \left(\boldsymbol{V}_4^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{V}_4^{\frac{1}{2}}\right)^{\frac{1}{2}} &= \boldsymbol{R}_4 \, \mathrm{diag}\{\theta_{41}, \theta_{42}, \cdots, \theta_{4n}\} \boldsymbol{R}_4^T, \end{aligned} \tag{17}$$

respectively, where $\boldsymbol{R}_1 \boldsymbol{R}_1^T = \boldsymbol{I}_m$ and $\boldsymbol{R}_4 \boldsymbol{R}_4^T = \boldsymbol{I}_n$. By numerical manupilations, we can obtain an $m \times m$ ($n \times n$) orthogonal matrix $\boldsymbol{S}_1$ ($\boldsymbol{S}_4$) such that [13]

$$\boldsymbol{S}_i \boldsymbol{\Lambda}_i \boldsymbol{S}_i^T = \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \quad \text{for} \quad i = 1, 4 \tag{18}$$

where

$$\begin{aligned} \boldsymbol{\Lambda}_1 &= \mathrm{diag}\{\lambda_{11}, \lambda_{12}, \cdots, \lambda_{1m}\} \\ \lambda_{1i} &= \frac{m\theta_{1i}}{\theta_{11} + \theta_{12} + \cdots + \theta_{1m}} \quad \text{for} \quad i = 1, 2, \cdots, m \\ \boldsymbol{\Lambda}_4 &= \mathrm{diag}\{\lambda_{41}, \lambda_{42}, \cdots, \lambda_{4n}\} \\ \lambda_{4i} &= \frac{n\theta_{4i}}{\theta_{41} + \theta_{42} + \cdots + \theta_{4n}} \quad \text{for} \quad i = 1, 2, \cdots, n. \end{aligned}$$

Here, we choose the orthogonal matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_4$ as

$$\boldsymbol{U}_1 = \boldsymbol{R}_1 \boldsymbol{S}_1^T, \qquad \boldsymbol{U}_4 = \boldsymbol{R}_4 \boldsymbol{S}_4^T. \tag{19}$$

Below we consider the case where $\boldsymbol{D}_i$ is a diagonal matrix for $i = 0, 1, \cdots, N-1$.

In this case, matrix $\boldsymbol{D}_i = \boldsymbol{D}_{1i} \oplus \boldsymbol{D}_{4i}$ assumes the form

$$\begin{aligned} \boldsymbol{D}_{1i} &= \mathrm{diag}\{\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{im}\} \\ \boldsymbol{D}_{4i} &= \mathrm{diag}\{\beta_{i1}, \beta_{i2}, \cdots, \beta_{in}\}. \end{aligned} \tag{20}$$

Then it follows from (12) that

$$\begin{aligned} \frac{\partial J_{e2}(\overline{\boldsymbol{c}}, \boldsymbol{D}, \boldsymbol{T})}{\partial \alpha_{il}} &= -2\left(\overline{\boldsymbol{W}}_{1i}\right)_{ll} + 2 \sum_{k=0}^{N-1} \alpha_{kl} \left(\overline{\boldsymbol{W}}_{1,|i-k|}\right)_{ll} \\ &= 0 \quad \text{for} \quad l = 1, 2, \cdots, m \\ \frac{\partial J_{e2}(\overline{\boldsymbol{c}}, \boldsymbol{D}, \boldsymbol{T})}{\partial \beta_{ir}} &= -2\left(\overline{\boldsymbol{W}}_{4i}\right)_{rr} + 2 \sum_{k=0}^{N-1} \beta_{kr} \left(\overline{\boldsymbol{W}}_{4,|i-k|}\right)_{rr} \\ &= 0 \quad \text{for} \quad r = 1, 2, \cdots, n. \end{aligned} \tag{21}$$

As a result, matrix $\boldsymbol{D}_i = \boldsymbol{D}_{1i} \oplus \boldsymbol{D}_{4i}$ can be derived from

$$\begin{aligned} \begin{bmatrix} \alpha_{0l} \\ \alpha_{1l} \\ \vdots \\ \alpha_{N-1,l} \end{bmatrix} &= \begin{bmatrix} p_{0l} & p_{1l} & \cdots & p_{N-1,l} \\ p_{1l} & p_{0l} & \cdots & p_{N-2,l} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N-1,l} & p_{N-2,l} & \cdots & p_{0l} \end{bmatrix}^{-1} \begin{bmatrix} p_{1l} \\ p_{2l} \\ \vdots \\ p_{Nl} \end{bmatrix} \\ \begin{bmatrix} \beta_{0r} \\ \beta_{1r} \\ \vdots \\ \beta_{N-1,r} \end{bmatrix} &= \begin{bmatrix} q_{0r} & q_{1r} & \cdots & q_{N-1,r} \\ q_{1r} & q_{0r} & \cdots & q_{N-2,r} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N-1,r} & q_{N-2,r} & \cdots & q_{0r} \end{bmatrix}^{-1} \begin{bmatrix} q_{1r} \\ q_{2r} \\ \vdots \\ q_{Nr} \end{bmatrix} \end{aligned} \tag{22}$$

where $p_{il} = \left(\overline{\boldsymbol{W}}_{1i}\right)_{ll}$ for $l = 1, 2, \cdots, m$ and $q_{ir} = \left(\overline{\boldsymbol{W}}_{4i}\right)_{rr}$ for $r = 1, 2, \cdots, n$.

IV. A NUMERICAL EXAMPLE

As a numerical example, consider a 2-D state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{2+2}$ in (1) described by

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

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

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

and $d = 0.87021$. From (8) and (9), the local controllability and the local observability Gramians $\boldsymbol{K}_c$ and $\boldsymbol{W}_o = \boldsymbol{W}_{10} \oplus \boldsymbol{W}_{40}$ of the above filter were computed as

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

$$\boldsymbol{W}_o = \begin{bmatrix} 1.13447 & -1.03553 \\ -1.03553 & 0.97348 \end{bmatrix} \oplus \begin{bmatrix} 87.17246 & 85.25727 \\ 85.25727 & 87.17246 \end{bmatrix}.$$

Applying the coordinate transformation specified by $\boldsymbol{T}_o = \boldsymbol{T}_{o1} \oplus \boldsymbol{T}_{o4} = \mathrm{diag}\{9.336620 \;\; 9.336620 \;\; 1.065113 \;\; 0.986652\}$ so that the l_2-scaling constraints $(\boldsymbol{T}_o^{-1} \boldsymbol{K}_c \boldsymbol{T}_o^{-T})_{ii} = 1$ are satisfied for $i = 1, 2, 3, 4$, the noise gain of the filter with no error feedforward and no EF was computed from (8) as

$$J_{e1}(\boldsymbol{0}, \boldsymbol{0}) = \mathrm{tr}[\boldsymbol{T}_{o1}^T \boldsymbol{W}_{10} \boldsymbol{T}_{o1}] + \mathrm{tr}[\boldsymbol{T}_{o4}^T \boldsymbol{W}_{40} \boldsymbol{T}_{o4}] = 367.51003.$$

The optimal coordinate transformation matrix $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ was constructed using (16)-(19) as

$$\boldsymbol{T} = \begin{bmatrix} -0.207923 & 1.032526 \\ -0.414525 & 1.021700 \end{bmatrix} \oplus \begin{bmatrix} 0.888012 & 0.303911 \\ -0.962330 & -0.132655 \end{bmatrix}.$$

With $\boldsymbol{T}$ found and fixed, we now consider two cases: one employs a single EF matrix $\boldsymbol{D}_0$ (i.e. $N = 1$) and the other uses two EF matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ (i.e. $N = 2$). In each case, EF matrix assumes the form of diagonal matrices. The EF matrices were optimized using the method presented in Section III, and the results in terms of noise gain $J_{e2}(\overline{\boldsymbol{c}}, \boldsymbol{D}, \boldsymbol{T})$ are summarized in Table I, where "3-Bit Quantization" refers to value of $J_{e2}(\overline{\boldsymbol{c}}, \boldsymbol{D}, \boldsymbol{T})$ where the fraction part of each component of $\boldsymbol{D}$ was rounded to 3 bits in a power-of-two representation. From the table, it is observed that (i) employing a dynamic EF (i.e. with $N > 1$) matrices helps reduce the noise gain in a significant way, and (ii) compared with their infinite-precision counterparts, the use of quantized EF matrices reduces implementation complexity with only slightly degraded performance.

TABLE I
SEPARATE OPTIMIZATION

N	$\boldsymbol{D}$	Infinite Precision	3-Bit Quantization
1	Diagonal	1.446145	1.483189
2	Diagonal	1.320965	1.378140

V. CONCLUSION

The problem of separate optimization of high-order EF and state-space realization for roundoff noise minimization subject to dynamic range l_2-scaling constraints in 2-D state-space digital filters has been investigated. After choosing the optimal coordinate transformation matrices under the l_2-scaling constraints, high-order EF matrices are determined so as to minimize the roundoff noise gain in the optimal realization where high-order EF matrices are diagonal. A numerical example has been presented to demonstrate the validity and effectiveness of the proposed technique.

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] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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