ROUNDOFF NOISE MINIMIZATION IN TWO-DIMENSIONAL STATE-SPACE DIGITAL FILTERS USING ERROR FEEDBACK

Takao Hinamoto, Keisuke Higashi and Wu-Sheng Lu†

Graduate School of Engineering, Hiroshima University, Japan
hinamoto@hiroshima-u.ac.jp
†Dept. Elect. Comput. Eng., University of Victoria, Canada
wslu@ece.uvic.ca

ABSTRACT

This paper considers the problem of minimizing roundoff noise in two-dimensional (2-D) state-space digital filters subject to L_2-norm dynamic-range scaling constraints. The minimization will be achieved by using error feedback. Several techniques for the determination of the optimal full-scale, block-diagonal, diagonal, and scalar error-feedback matrices for a given 2-D state-space digital filter are proposed. A numerical example is presented to illustrate the utility of the proposed techniques.

I. INTRODUCTION

One of the primary finite-word-length (FWL) register effects in fixed-point digital filters is the roundoff noise caused by the rounding of products/summations within the realization. One can reduce the roundoff noise at the filter output using error feedback, which is achieved by extracting the quantization error after multiplications and additions, and then feeding the error signal back to a certain point through a simple circuit. Several techniques for error feedback have been presented in the past for 1-D digital filters [1]-[5], and more recently for 2-D digital filters [6]-[9].

This paper proposes several new algorithms for the reduction of roundoff noise in 2-D state-space digital filters. Several closed-form formulas for evaluating the optimal full-scale, block-diagonal, diagonal, and scalar error-feedback matrices for a given 2-D state-space digital filter are derived. A numerical example is presented to illustrate the algorithms proposed and to demonstrate their performance.

II. 2-D STATE-SPACE DIGITAL FILTERS WITH ERROR FEEDBACK

Consider the Roesser local state-space (LSS) model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m,n}$ which is stable, separately locally controllable and separately locally observable:

$$\begin{aligned} \boldsymbol{x}_{11}(i,j) &= \boldsymbol{A}\boldsymbol{x}(i,j) + \boldsymbol{b}u(i,j) \\ y(i,j) &= \boldsymbol{c}\boldsymbol{x}(i,j) + du(i,j) \end{aligned} \tag{1}$$

where

$$\boldsymbol{x}_{11}(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i+1,j) \\ \boldsymbol{x}^v(i,j+1) \end{bmatrix}, \quad \boldsymbol{x}(i,j) = \begin{bmatrix} \boldsymbol{x}^h(i,j) \\ \boldsymbol{x}^v(i,j) \end{bmatrix}$$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{A}_3 & \boldsymbol{A}_4 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{bmatrix}.$$

Here, $\boldsymbol{x}^h(i,j)$ 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.

Carrying out the quantization before matrix-vector multiplication, an FWL implementation of (1) can be expressed as

$$\begin{aligned} \tilde{\boldsymbol{x}}_{11}(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + \boldsymbol{b}u(i,j) \\ \tilde{y}(i,j) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + du(i,j) \end{aligned} \tag{2}$$

where each component of $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}$, and d assumes an exact fractional B_c bit representation. The FWL local state vector $\tilde{\boldsymbol{x}}(i,j)$ and the output $\tilde{y}(i,j)$ all have a B bit fractional representation, 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 $\tilde{\boldsymbol{x}}(i,j)$ to $(B - B_c)$ bits after multiplications and additions, where the sign bit is not counted. The quantization error

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

coincides with the residue left in the lower part of $\tilde{\boldsymbol{x}}(i,j)$. The roundoff error $\boldsymbol{e}(i,j)$ is modeled as a zero-mean noise process of covariance $\sigma^2 \boldsymbol{I}_{m+n}$ with

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

0-7803-7762-1/03/$17.00 ©2003 IEEE

To reduce the filter's roundoff noise, the quantization error $\boldsymbol{e}(i,j)$ is fed back to each input of delay operators through an $(m+n)\times(m+n)$ constant matrix $\boldsymbol{D}$ in the FWL filter (2). The 2-D filter with error feedback can be characterized by

$$\begin{aligned}\tilde{\boldsymbol{x}}_{11}(i,j) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + \boldsymbol{b}u(i,j) + \boldsymbol{D}\boldsymbol{e}(i,j)\\ \tilde{y}(i,j) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,j)] + du(i,j)\end{aligned} \tag{4}$$

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

Subtracting (4) from (1) yields

$$\begin{aligned}\Delta\boldsymbol{x}_{11}(i,j) &= \boldsymbol{A}\Delta\boldsymbol{x}(i,j) + (\boldsymbol{A}-\boldsymbol{D})\boldsymbol{e}(i,j)\\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}(i,j) + \boldsymbol{c}\boldsymbol{e}(i,j)\end{aligned} \tag{5}$$

where

$$\begin{aligned}\Delta\boldsymbol{x}(i,j) &= \boldsymbol{x}(i,j) - \tilde{\boldsymbol{x}}(i,j)\\ \Delta\boldsymbol{x}_{11}(i,j) &= \boldsymbol{x}_{11}(i,j) - \tilde{\boldsymbol{x}}_{11}(i,j)\\ \Delta y(i,j) &= y(i,j) - \tilde{y}(i,j).\end{aligned}$$

Let $\boldsymbol{G}_D(z_1,z_2)$ be the 2-D transfer function from the quantization error, $\boldsymbol{e}(i,j)$, to the filter output, $\Delta y(i,j)$. Then, we obtain

$$\boldsymbol{G}_D(z_1,z_2) = \boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}(\boldsymbol{A}-\boldsymbol{D}) + \boldsymbol{c} \tag{6}$$

where $\boldsymbol{Z} = z_1\boldsymbol{I}_m \oplus z_2\boldsymbol{I}_n$. The noise variance gain $I(\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ is then defined by

$$I(\boldsymbol{D}) = tr[\boldsymbol{W}_D] \tag{7}$$

where σ_{out}^2 denotes noise variance at the output, and

$$\boldsymbol{W}_D = \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}\boldsymbol{G}_D^*(z_1,z_2)\boldsymbol{G}_D(z_1,z_2)\frac{dz_1dz_2}{z_1z_2}$$

with $\Gamma_i = \{z_i : |z_i| = 1\}$ for $i = 1,2$. By applying the 2-D Cauchy integral theorem, we obtain

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

where $\boldsymbol{W}_o$ is called the *local observability Gramian* of the 2-D filter, and is defined by

$$\begin{aligned}\boldsymbol{W}_o &= \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}(\boldsymbol{Z}^*-\boldsymbol{A}^T)^{-1}\boldsymbol{c}^T\boldsymbol{c}(\boldsymbol{Z}-\boldsymbol{A})^{-1}\\ &\quad\cdot\frac{dz_1dz_2}{z_1z_2} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{g}(i,j)^T\boldsymbol{g}(i,j).\end{aligned} \tag{9}$$

If there is no error feedback in the 2-D filter, then the noise variance gain $I(\boldsymbol{D})$ with $\boldsymbol{D} = \boldsymbol{0}$ becomes

$$\begin{aligned}I(\boldsymbol{0}) &= \mathrm{tr}[\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}]\\ &= \mathrm{tr}[\boldsymbol{W}_o].\end{aligned} \tag{10}$$

The l_2-norm dynamic-range scaling constraints on the local state vector involves the *local controllability Gramian* of the 2-D filter, which is defined by

$$\begin{aligned}\boldsymbol{K}_c &= \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}(\boldsymbol{Z}-\boldsymbol{A}^T)^{-1}\boldsymbol{b}\boldsymbol{b}^T(\boldsymbol{Z}^*-\boldsymbol{A}^T)^{-1}\\ &\quad\cdot\frac{dz_1dz_2}{z_1z_2} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{f}(i,j)\boldsymbol{f}(i,j)^T.\end{aligned} \tag{11}$$

The problem considered is to design the error-feedback matrix $\boldsymbol{D}$ that minimizes (7), where matrix $\boldsymbol{W}_D$ is specified by (8), subject to that all the diagonal elements of $\boldsymbol{K}_c$ equal unity.

III. DETERMINATION OF OPTIMAL ERROR FEEDBACK MATRICES

In this section, we derive closed-form formulas for the determination of the optimal full-scale, block-diagonal, diagonal, and scalar error-feedback matrix $\boldsymbol{D}$ to minimize $I(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_D]$ for a given 2-D state-space digital filter.

Case 1: $\boldsymbol{D}$ is a general matrix

Substituting (8) into (7), we obtain

$$\begin{aligned}I(\boldsymbol{D}) &= \mathrm{tr}[\boldsymbol{c}^T\boldsymbol{c} + (\boldsymbol{A}-\boldsymbol{D})^T\boldsymbol{W}_o(\boldsymbol{A}-\boldsymbol{D})]\\ &= \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{D}] - 2\,\mathrm{tr}[\boldsymbol{D}^T\boldsymbol{W}_o\boldsymbol{A}].\end{aligned} \tag{12}$$

Differentiating (12) with respect to the error-feedback matrix $\boldsymbol{D}$ yields

$$\frac{\partial I(\boldsymbol{D})}{\partial \boldsymbol{D}} = 2\boldsymbol{W}_o(\boldsymbol{D}-\boldsymbol{A}). \tag{13}$$

By choosing the error-feedback matrix as $\boldsymbol{D} = \boldsymbol{A}$, the noise gain $I(\boldsymbol{D})$ in (12) achieves its minimum value

$$\begin{aligned}I_{min}(\boldsymbol{D}) &= \mathrm{tr}[\boldsymbol{W}_o] - \mathrm{tr}[\boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A}]\\ &= \mathrm{tr}[\boldsymbol{c}^T\boldsymbol{c}].\end{aligned} \tag{14}$$

Case 2: $\boldsymbol{D}$ is a block-diagonal matrix

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

$$\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4 \tag{15}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_1$ are $m\times m$ and $n\times n$ matrices, respectively, which leads (12) to

$$\begin{aligned}I(\boldsymbol{D}) &= \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{D}_1^T\boldsymbol{W}_{o1}\boldsymbol{D}_1] + \mathrm{tr}[\boldsymbol{D}_4^T\boldsymbol{W}_{o4}\boldsymbol{D}_4]\\ &\quad -2\,\mathrm{tr}[\boldsymbol{D}_1^T(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)]\\ &\quad -2\,\mathrm{tr}[\boldsymbol{D}_4^T(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)]\end{aligned} \tag{16}$$

where

$$\boldsymbol{W}_o = \begin{bmatrix} \boldsymbol{W}_{o1} & \boldsymbol{W}_{o2} \\ \boldsymbol{W}_{o3} & \boldsymbol{W}_{o4} \end{bmatrix}.$$

Letting $\partial I(\boldsymbol{D})/\partial \boldsymbol{D}_i = \boldsymbol{0}$ for $i = 1, 4$ yields

$$\begin{aligned} \boldsymbol{D}_1 &= \boldsymbol{A}_1 + \boldsymbol{W}_{o1}^{-1}\boldsymbol{W}_{o2}\boldsymbol{A}_3 \\ \boldsymbol{D}_4 &= \boldsymbol{A}_4 + \boldsymbol{W}_{o4}^{-1}\boldsymbol{W}_{o3}\boldsymbol{A}_2. \end{aligned} \tag{17}$$

By substituting (17) into (16), we obtain the minimum value of the noise variance gain $I(\boldsymbol{D})$ as

$$\begin{aligned} I_{min}(\boldsymbol{D}) \;=\; & \mathrm{tr}[\boldsymbol{W}_o] - tr[\boldsymbol{D}_1^T(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)] \\ & -tr[\boldsymbol{D}_4^T(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)]. \end{aligned} \tag{18}$$

Case 3: $\boldsymbol{D}$ is a diagonal matrix

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

$$\begin{aligned} \boldsymbol{D}_1 &= \mathrm{diag}\{\alpha_1, \alpha_2, \cdots, \alpha_m\} \\ \boldsymbol{D}_4 &= \mathrm{diag}\{\beta_1, \beta_2, \cdots, \beta_n\} \end{aligned} \tag{19}$$

which leads (16) to

$$\begin{aligned} I(\boldsymbol{D}) = \; & \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{D}_1^2\boldsymbol{W}_{o1}] + \mathrm{tr}[\boldsymbol{D}_4^2\boldsymbol{W}_{o4}] \\ & -2\,\mathrm{tr}[\boldsymbol{D}_1(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)] \\ & -2\,\mathrm{tr}[\boldsymbol{D}_4(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)]. \end{aligned} \tag{20}$$

This implies that if α_i's and β_i's satisfy

$$\begin{aligned} &\alpha_i\left(\alpha_i - 2\frac{(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)_{ii}}{(\boldsymbol{W}_{o1})_{ii}}\right) < 0,\;\; i = 1, 2, \cdots, m \\ &\beta_i\left(\beta_i - 2\frac{(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)_{ii}}{(\boldsymbol{W}_{o4})_{ii}}\right) < 0,\;\; i = 1, 2, \cdots, n \end{aligned} \tag{21}$$

then $I(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_D] < \mathrm{tr}[\boldsymbol{W}_o]$. Letting $\partial I(\boldsymbol{D})/\partial \alpha_i = 0$ for $i = 1, 2, \cdots, m$ and letting $\partial I(\boldsymbol{D})/\partial \beta_i = 0$ for $i = 1, 2, \cdots, n$ gives

$$\begin{aligned} \alpha_i &= \frac{(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)_{ii}}{(\boldsymbol{W}_{o1})_{ii}},\;\; i = 1, 2, \cdots, m \\ \beta_i &= \frac{(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)_{ii}}{(\boldsymbol{W}_{o4})_{ii}},\;\; i = 1, 2, \cdots, n \end{aligned} \tag{22}$$

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

$$\begin{aligned} I_{min}(\boldsymbol{D}) \;=\; & \mathrm{tr}[\boldsymbol{W}_o] - \sum_{i=1}^{m} \frac{(\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3)_{ii}^2}{(\boldsymbol{W}_{o1})_{ii}} \\ & - \sum_{i=1}^{n} \frac{(\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4)_{ii}^2}{(\boldsymbol{W}_{o4})_{ii}} \end{aligned} \tag{23}$$

where $(\boldsymbol{A})_{ii}$ denotes the ith diagonal element of a square matrix $\boldsymbol{A}$.

Case 4: $\boldsymbol{D}_1$ and $\boldsymbol{D}_4$ are scalar matrices $\alpha\boldsymbol{I}_m$ and $\beta\boldsymbol{I}_n$

If $\boldsymbol{D}_1 = \alpha\boldsymbol{I}_m$ and $\boldsymbol{D}_4 = \beta\boldsymbol{I}_n$ with scalars α and β, then (20) becomes

$$\begin{aligned} I(\boldsymbol{D}) = \; & \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{W}_{o1}]\alpha^2 + \mathrm{tr}[\boldsymbol{W}_{o4}]\beta^2 \\ & -2\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3]\alpha \\ & -2\mathrm{tr}[\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4]\beta. \end{aligned} \tag{24}$$

Hence, if α and β satisfy

$$\begin{aligned} &\alpha\left(\alpha - 2\frac{\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3]}{\mathrm{tr}[\boldsymbol{W}_{o1}]}\right) < 0 \\ &\beta\left(\beta - 2\frac{\mathrm{tr}[\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4]}{\mathrm{tr}[\boldsymbol{W}_{o4}]}\right) < 0 \end{aligned} \tag{25}$$

then $I(\boldsymbol{D}) = \mathrm{tr}[\boldsymbol{W}_D] < \mathrm{tr}[\boldsymbol{W}_o]$. Moreover, from $\partial I(\boldsymbol{D})/\partial\alpha = 0$ and $\partial I(\boldsymbol{D})/\partial\beta = 0$, it follows that

$$\begin{aligned} \alpha &= \frac{\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3]}{\mathrm{tr}[\boldsymbol{W}_{o1}]} \\ \beta &= \frac{\mathrm{tr}[\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4]}{\mathrm{tr}[\boldsymbol{W}_{o4}]} \end{aligned} \tag{26}$$

which lead (24) to

$$\begin{aligned} I_{min}(\boldsymbol{D}) \;=\; & \mathrm{tr}[\boldsymbol{W}_o] - \frac{(\mathrm{tr}[\boldsymbol{W}_{o1}\boldsymbol{A}_1 + \boldsymbol{W}_{o2}\boldsymbol{A}_3])^2}{\mathrm{tr}[\boldsymbol{W}_{o1}]} \\ & - \frac{(\mathrm{tr}[\boldsymbol{W}_{o3}\boldsymbol{A}_2 + \boldsymbol{W}_{o4}\boldsymbol{A}_4])^2}{\mathrm{tr}[\boldsymbol{W}_{o4}]}. \end{aligned} \tag{27}$$

IV. A NUMERICAL EXAMPLE

Let a 2-D state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{3,3}$ with $d = 0.0$ be described by

$$\boldsymbol{A}_1 = \begin{bmatrix} 0.621553 & 0.014666 & -0.476979 \\ -0.081625 & 0.621548 & -0.181986 \\ 0.181983 & 0.476990 & 0.663600 \end{bmatrix}$$

$$\boldsymbol{A}_2 = \begin{bmatrix} 0.059369 & -0.004829 & -0.024002 \\ -0.646852 & 0.061969 & 0.227715 \\ -0.229635 & 0.021958 & 0.076674 \end{bmatrix}$$

$$\boldsymbol{A}_3 = \begin{bmatrix} 0.000378 & 0.000763 & 0.001503 \\ -0.000463 & -0.001501 & 0.000812 \\ -0.000021 & -0.000219 & 0.000908 \end{bmatrix}$$

$$\boldsymbol{A}_4 = \begin{bmatrix} 0.620418 & 0.016504 & -0.479313 \\ -0.083124 & 0.620420 & -0.181961 \\ 0.181967 & 0.479315 & 0.661692 \end{bmatrix}$$

$$\boldsymbol{b}_1 = \begin{bmatrix} -0.007708 & 0.081835 & 0.028969 \end{bmatrix}^T$$

$$\boldsymbol{b}_2 = \begin{bmatrix} -0.079883 & 0.846271 & 0.294745 \end{bmatrix}^T$$

$$\boldsymbol{c}_1 = \begin{bmatrix} -0.766526 & 0.072050 & 0.267706 \end{bmatrix}$$

$$\boldsymbol{c}_2 = \begin{bmatrix} -0.074064 & 0.007031 & 0.026238 \end{bmatrix}.$$

which is stable, separately locally controllable and separately locally observable. This corresponds to the *optimal realization* with minimum roundoff noise $I(\mathbf{0}) = 4.927082$, subject to the l_2-norm dynamic-range scaling constraints.

In case $\boldsymbol{D}$ is allowed to be a general matrix, then (13) suggests that we should choose $\boldsymbol{D} = \boldsymbol{A}$ which yields $I_{min}(\boldsymbol{D}) = 0.670643$. Suppose the elements of matrix $\boldsymbol{D}$ are rounded to power-of-two quantization with 3 bits after binary point (integer quantization), then the noise gain is given by

$$I(\boldsymbol{D}) = 0.700468 \qquad (1.726719).$$

If $\boldsymbol{D}$ is constrained to be a block-diagonal matrix, then the optimal $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ is calculated using (17), which gives

$$\boldsymbol{D}_1 = \begin{bmatrix} 0.621550 & 0.014658 & -0.476971 \\ -0.081619 & 0.621565 & -0.181992 \\ 0.181976 & 0.476968 & 0.663613 \end{bmatrix}$$

$$\boldsymbol{D}_4 = \begin{bmatrix} 0.618351 & 0.016703 & -0.478595 \\ -0.082737 & 0.620382 & -0.182106 \\ 0.181250 & 0.479384 & 0.661937 \end{bmatrix}$$

$I_{min}(\boldsymbol{D}_1 \oplus \boldsymbol{D}_4) = 1.331653.$

After 3-bit quantization (integer quantization), this block-diagonal matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ gives

$$I(\boldsymbol{D}_1 \oplus \boldsymbol{D}_4) = 1.351427 \qquad (2.247670).$$

If $\boldsymbol{D}$ is constrained to be a diagonal error-feedback matrix, then it can be calculated using (22) as

$$\boldsymbol{D}_1 = \text{diag}\{0.523405,\ 0.934410,\ 0.859620\}$$

$$\boldsymbol{D}_4 = \text{diag}\{0.521174,\ 0.934021,\ 0.858810\}$$

which yields $I_{min}(\boldsymbol{D}) = 1.833208$. After 3-bit quantization (integer quantization), this diagonal matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ yields

$$I(\boldsymbol{D}_1 \oplus \boldsymbol{D}_4) = 1.840195 \qquad (2.247670).$$

If a scalar error-feedback matrix is calculated using (26), then we obtain

$$\alpha = 0.772479, \qquad \beta = 0.771335$$

which yields $I_{min}(\boldsymbol{D}) = 1.991329$. After 3-bit quantization (integer quantization), this scalar matrix $\boldsymbol{D}$ results in

$$I(\boldsymbol{D}) = 1.993695 \qquad (2.247670).$$

From these results, it is observed that the utilization of an optimal error feedback matrix leads to considerable reduction in roundoff noise, even when a scalar $\boldsymbol{D} = \alpha \boldsymbol{I}_m \oplus \beta \boldsymbol{I}_n$ with quantized α and β.

V. CONCLUSION

In this paper, the problem of minimizing roundoff noise in 2-D state-space digital filters has been investigated by means of error feedback. General, block-diagonal, diagonal, and scalar error-feedback matrices for minimizing the noise variance gain in a given 2-D state-space digital filter have been derived. Simulation results have been presented to illustrate the validity of our theoretical analysis and proposed algorithms.

1. REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] T. Hinamoto,S. Karino and N. Kuroda,"Error spectrum shaping in 2-D digital filters," *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'95), vol. 1, pp. 348-351, May 1995.

[7] P. Agathoklis and C. Xiao, "Low roundoff noise structures for 2-D filters," *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'96), vol. 2, pp. 352-355, May 1996.

[8] T. Hinamoto,S. Karino and N. Kuroda,"2-D state-space digital filters with error spectrum shaping," *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'96), vol. 2, pp. 766-769, May 1996.

[9] 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.