JOINTLY OPTIMIZED ERROR FEEDBACK AND REALIZATION FOR ROUNDOFF NOISE MINIMIZATION IN TWO-DIMENSIONAL STATE-SPACE DIGITAL FILTERS

Takao Hinamoto, Keisuke Higashi and Wu-Sheng Lu†

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

ABSTRACT

The minimization of roundoff noise subject to l_2-norm dynamic-range scaling constraints in two-dimensional (2-D) state-space digital filters is considered by using joint error feedback and coordinate transformation optimization. An iterative approach for minimizing the roundoff noise under l_2-norm dynamic-range scaling constraints is developed by jointly optimizing a scalar error-feedback matrix and a coordinate transformation matrix. A numerical example is presented to illustrate the utility of the proposed technique.

I. INTRODUCTION

When a given transfer function is realized through hardware implementation using fixed-point arithmetic, the internal noise caused by finite-word-length (FWL) registers may be the most serious concern to be dealt with. One of the primary FWL register effects in fixed-point digital filters is the roundoff noise caused by the rounding of products/summations within the realization. The synthesis of state-space digital filter structures with minimum rounfoff noise under l_2-norm dynamic-range scaling constraints has been investigated in two-dimensional (2-D) state-space digital filters [1],[2]. Another technique for the reduction of roundoff noise at the filter output is to use error feedback. Some techniques for error feedback have been presented in the past for 2-D digital filters [3]-[7].

In this paper, an iterative noise reduction technique for 2-D state-space digital filters is developed by jointly optimizing a scalar error feedback matrix and a coordinate transformation matrix. A numerical example is presented to illustrate the algorithm proposed and to demonstrate its performance.

Throughout the paper, the ith diagonal element of a square matrix $\boldsymbol{A}$ is denoted by $(\boldsymbol{A})_{ii}$.

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

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

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

$$\boldsymbol{W}_D = \frac{1}{(2\pi j)^2} \oint_{\Gamma_1} \oint_{\Gamma_2} \boldsymbol{G}_D^*(z_1,z_2)\boldsymbol{G}_D(z_1,z_2) \frac{dz_1 dz_2}{z_1 z_2}$$

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

$$\begin{aligned} \boldsymbol{W}_o \;=\; & \frac{1}{(2\pi j)^2} \oint_{\Gamma_1} \oint_{\Gamma_2} (\boldsymbol{Z}^* - \boldsymbol{A}^T)^{-1} \boldsymbol{c}^T\boldsymbol{c} (\boldsymbol{Z}-\boldsymbol{A})^{-1} \\ & \cdot \frac{dz_1 dz_2}{z_1 z_2} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{g}(i,j)^T \boldsymbol{g}(i,j). \end{aligned} \tag{9}$$

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

$$\begin{aligned} \boldsymbol{K}_c \;=\; & \frac{1}{(2\pi j)^2} \oint_{\Gamma_1} \oint_{\Gamma_2} (\boldsymbol{Z}-\boldsymbol{A}^T)^{-1} \boldsymbol{b}\boldsymbol{b}^T (\boldsymbol{Z}^* - \boldsymbol{A}^T)^{-1} \\ & \cdot \frac{dz_1 dz_2}{z_1 z_2} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{f}(i,j)\boldsymbol{f}(i,j)^T. \end{aligned} \tag{11}$$

III. JOINT OPTIMIZATION OF ERROR FEEDBACK AND COORDINATE TRANSFORMATION

An equivalent description of (1), $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_{m+n}$, obtained via a coordinate transformation $\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k)$ with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ is characterized by

$$\begin{aligned} &\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T} \\ &\overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}, \qquad \overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}. \end{aligned} \tag{12}$$

The purpose of this section is to investigate the joint optimization of scalar error-feedback matrices $\boldsymbol{D}_1 = \alpha\boldsymbol{I}_m$ and $\boldsymbol{D}_4 = \beta\boldsymbol{I}_n$, and coordinate transformation matrices $\boldsymbol{T}_1$ and $\boldsymbol{T}_4$ for roundoff noise minimization under l_2-norm dynamic-range scaling constraints:

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

Let matrix $\boldsymbol{W}_D$ in (8) with $\boldsymbol{D} = \alpha\boldsymbol{I}_m \oplus \beta\boldsymbol{I}_n$ and matrix $\boldsymbol{K}_c$ in (11) be denoted by

$$\boldsymbol{W}_D = \begin{bmatrix} \boldsymbol{W}_{1\alpha} & \boldsymbol{W}_{\alpha\beta}^T \\ \boldsymbol{W}_{\alpha\beta} & \boldsymbol{W}_{4\beta} \end{bmatrix}, \quad \boldsymbol{K}_c = \begin{bmatrix} \boldsymbol{K}_{c1} & \boldsymbol{K}_{c2} \\ \boldsymbol{K}_{c3} & \boldsymbol{K}_{c4} \end{bmatrix},$$

respectively.

Under the joint application of scalar error-feedback and coordinate transformation, we minimize $I(\boldsymbol{D})$ (with α and β temporarily fixed) over an $m \times m$ nonsingular matrix $\boldsymbol{T}_1$ and an $n \times n$ nonsingular matrix $\boldsymbol{T}_4$ subject to the constraints stated in (13). To this end, we define the Lagrange function

$$\begin{aligned} J(\alpha,\beta,\boldsymbol{P}) = \; & \mathrm{tr}[\boldsymbol{W}_{1\alpha}\boldsymbol{P}_1] + \lambda_1(\mathrm{tr}[\boldsymbol{K}_{c1}\boldsymbol{P}_1^{-1}] - m) \\ & + \mathrm{tr}[\boldsymbol{W}_{4\beta}\boldsymbol{P}_4] + \lambda_4(\mathrm{tr}[\boldsymbol{K}_{c4}\boldsymbol{P}_4^{-1}] - n) \end{aligned} \tag{14}$$

where $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$, $\boldsymbol{P}_i = \boldsymbol{T}_i \boldsymbol{T}_i^T$ for $i = 1, 4$, and λ_i for $i = 1, 4$ are Lagrange multipliers. We compute

$$\begin{aligned}
\frac{\partial J(\alpha, \beta, \boldsymbol{P})}{\partial \boldsymbol{P}_1} &= \boldsymbol{W}_{1\alpha} - \lambda_1 \boldsymbol{P}_1^{-1} \boldsymbol{K}_{c1} \boldsymbol{P}_1^{-1} \\
\frac{\partial J(\alpha, \beta, \boldsymbol{P})}{\partial \boldsymbol{P}_4} &= \boldsymbol{W}_{4\beta} - \lambda_4 \boldsymbol{P}_4^{-1} \boldsymbol{K}_{c4} \boldsymbol{P}_4^{-1} \\
\frac{\partial J(\alpha, \beta, \boldsymbol{P})}{\partial \lambda_1} &= \mathrm{tr}[\boldsymbol{K}_{c1} \boldsymbol{P}_1^{-1}] - m \\
\frac{\partial J(\alpha, \beta, \boldsymbol{P})}{\partial \lambda_4} &= \mathrm{tr}[\boldsymbol{K}_{c4} \boldsymbol{P}_4^{-1}] - n.
\end{aligned} \tag{15}$$

Let $\partial J(\alpha, \beta, \boldsymbol{P})/\partial \boldsymbol{P}_i = \boldsymbol{0}$ and $\partial J(\alpha, \beta, \boldsymbol{P})/\partial \lambda_i = 0$ for $i = 1, 4$. Then, it is derived that

$$\begin{aligned}
\boldsymbol{P}_1 &= \sqrt{\lambda_1} \boldsymbol{W}_{1\alpha}^{-\frac{1}{2}} [\boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}^{\frac{1}{2}}]^{\frac{1}{2}} \boldsymbol{W}_{1\alpha}^{-\frac{1}{2}} \\
\boldsymbol{P}_4 &= \sqrt{\lambda_4} \boldsymbol{W}_{4\beta}^{-\frac{1}{2}} [\boldsymbol{W}_{4\beta}^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}^{\frac{1}{2}}]^{\frac{1}{2}} \boldsymbol{W}_{4\beta}^{-\frac{1}{2}} \\
\frac{1}{\sqrt{\lambda_1}} \mathrm{tr}[\boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}]^{\frac{1}{2}} &= \frac{1}{\sqrt{\lambda_1}} \left(\sum_{i=1}^{m} \mu_i \right) = m \\
\frac{1}{\sqrt{\lambda_4}} \mathrm{tr}[\boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}]^{\frac{1}{2}} &= \frac{1}{\sqrt{\lambda_4}} \left(\sum_{i=1}^{n} \nu_i \right) = n
\end{aligned} \tag{16}$$

where μ_i^2 for $i = 1, 2, \cdots, m$ and ν_i^2 for $i = 1, 2, \cdots, n$ are the eigenvalues of $\boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}$ and $\boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}$, respectively. Therefore, we obtain

$$\begin{aligned}
\boldsymbol{P}_1 &= \frac{1}{m} \left(\sum_{i=1}^{m} \mu_1 \right) \boldsymbol{W}_{1\alpha}^{-\frac{1}{2}} [\boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}^{\frac{1}{2}}]^{\frac{1}{2}} \boldsymbol{W}_{1\alpha}^{-\frac{1}{2}} \\
\boldsymbol{P}_4 &= \frac{1}{n} \left(\sum_{i=1}^{n} \nu_i \right) \boldsymbol{W}_{4\beta}^{-\frac{1}{2}} [\boldsymbol{W}_{4\beta}^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}^{\frac{1}{2}}]^{\frac{1}{2}} \boldsymbol{W}_{4\beta}^{-\frac{1}{2}}.
\end{aligned} \tag{17}$$

Substituting (17) into (14) yields the minimum value of $J(\alpha, \beta, \boldsymbol{P})$ for fixed α and β as

$$\min_{\boldsymbol{P}} J(\alpha, \beta, \boldsymbol{P}) = \frac{1}{m} \left(\sum_{i=1}^{m} \mu_i \right)^2 + \frac{1}{n} \left(\sum_{i=1}^{n} \nu_i \right)^2. \tag{18}$$

An iterative procedure for minimizing (14) with respect to scalar parameters λ_i for $i = 1, 4$, α and β as well as an $(m+n) \times (m+n)$ symmetric positive-definite matrix $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ can be summarized as follows:

1) Set $i = 1$ and

$$\boldsymbol{P}(0) = \mathrm{diag}\{(\boldsymbol{K}_c)_{11}, (\boldsymbol{K}_c)_{22}, \cdots, (\boldsymbol{K}_c)_{m+n,m+n}\}.$$

2) Compute scalars $\alpha(i)$ and $\beta(i)$ using

$$\begin{aligned}
\alpha(i) &= \frac{\mathrm{tr}[(\boldsymbol{W}_{o1} \boldsymbol{A}_1 + \boldsymbol{W}_{o2} \boldsymbol{A}_3) \boldsymbol{P}_1(i-1)]}{\mathrm{tr}[\boldsymbol{W}_{o1} \boldsymbol{P}_1(i-1)]} \\
\beta(i) &= \frac{\mathrm{tr}[(\boldsymbol{W}_{o3} \boldsymbol{A}_2 + \boldsymbol{W}_{o4} \boldsymbol{A}_4) \boldsymbol{P}_4(i-1)]}{\mathrm{tr}[\boldsymbol{W}_{o4} \boldsymbol{P}_4(i-1)]}.
\end{aligned}$$

3) Compute $I_{min}(\alpha(i)\boldsymbol{I}_m \oplus \beta(i)\boldsymbol{I}_n) = (1 - \alpha(i)^2) \cdot \mathrm{tr}[\boldsymbol{W}_{o1} \boldsymbol{P}_1(i-1)] + (1 - \beta(i)^2)\, \mathrm{tr}[\boldsymbol{W}_{o4} \boldsymbol{P}_4(i-1)]$.

4) Replace $\boldsymbol{W}_{1\alpha}$ and $\boldsymbol{W}_{4\beta}$ by $\boldsymbol{W}_{1\alpha(i)}$ and $\boldsymbol{W}_{4\beta(i)}$ computed using

$$\begin{aligned}
\boldsymbol{W}_{1\alpha(i)} &= (1 + \alpha(i)^2) \boldsymbol{W}_{o1} - \alpha(i)[(\boldsymbol{W}_{o1} \boldsymbol{A}_1 \\
&\quad + \boldsymbol{W}_{02} \boldsymbol{A}_3)^T + \boldsymbol{W}_{o1} \boldsymbol{A}_1 + \boldsymbol{W}_{02} \boldsymbol{A}_3] \\
\boldsymbol{W}_{4\beta(i)} &= (1 + \beta(i)^2) \boldsymbol{W}_{o4} - \beta(i)[(\boldsymbol{W}_{o4} \boldsymbol{A}_4 \\
&\quad + \boldsymbol{W}_{03} \boldsymbol{A}_2)^T + \boldsymbol{W}_{o4} \boldsymbol{A}_4 + \boldsymbol{W}_{03} \boldsymbol{A}_2],
\end{aligned}$$

respectively.

5) Derive $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ from (17), and take the resulting matrix $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ as $\boldsymbol{P}(i) = \boldsymbol{P}_1(i) \oplus \boldsymbol{P}_4(i)$.

6) Compute $\mathrm{tr}[\boldsymbol{W}_{1\alpha(i)} \boldsymbol{P}_1(i)] + \mathrm{tr}[\boldsymbol{W}_{4\alpha(i)} \boldsymbol{P}_4(i)]$.

7) Update $i := i + 1$ and repeat from Step 2) until the change in either $I(\alpha(i)\boldsymbol{I}_m \oplus \beta(i)\boldsymbol{I}_n)$ or $\mathrm{tr}[\boldsymbol{W}_{1\alpha(i)} \boldsymbol{P}_1(i)] + \mathrm{tr}[\boldsymbol{W}_{4\alpha(i)} \boldsymbol{P}_4(i)]$ becomes insignificant compared to a prescribed tolerance.

From (17), the optimal coordinate transformation matrices $\boldsymbol{T}_1$ and $\boldsymbol{T}_4$ that minimize (14) can be obtained in closed form as

$$\begin{aligned}
\boldsymbol{T}_1 &= \frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} \mu_i \right)^{\frac{1}{2}} \boldsymbol{W}_{1\alpha}^{-\frac{1}{2}} [\boldsymbol{W}_{1\alpha}^{\frac{1}{2}} \boldsymbol{K}_{c1} \boldsymbol{W}_{1\alpha}^{\frac{1}{2}}]^{\frac{1}{4}} \boldsymbol{U}_1 \\
\boldsymbol{T}_4 &= \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \nu_i \right)^{\frac{1}{2}} \boldsymbol{W}_{4\beta}^{-\frac{1}{2}} [\boldsymbol{W}_{4\beta}^{\frac{1}{2}} \boldsymbol{K}_{c4} \boldsymbol{W}_{4\beta}^{\frac{1}{2}}]^{\frac{1}{4}} \boldsymbol{U}_4
\end{aligned} \tag{19}$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_4$ are $m \times m$ and $n \times n$ orthogonal matrices, respectively, which are obtained by applying a method reported in [7].

Once the optimal coordinate transformation matrix $\boldsymbol{T}(N) = \boldsymbol{T}_1(N) \oplus \boldsymbol{T}_4(N)$ is computed after N iterations, the diagonal error-feedback matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ with $\boldsymbol{D}_1 = \mathrm{diag}\{\alpha_1, \alpha_2, \cdots, \alpha_m\}$ and $\boldsymbol{D}_4 = \mathrm{diag}\{\beta_1, \beta_2, \cdots, \beta_n\}$ that minimizes $I(\boldsymbol{D})$, (7), in a new realization characterized by (12) is given by

$$\begin{aligned}
\alpha_i &= \frac{(\boldsymbol{T}_1^T(N)(\boldsymbol{W}_{o1} \boldsymbol{A}_1 + \boldsymbol{W}_{o2} \boldsymbol{A}_3) \boldsymbol{T}_1(N))_{ii}}{(\boldsymbol{T}_1^T(N) \boldsymbol{W}_{o1} \boldsymbol{T}_1(N))_{ii}} \\
\beta_i &= \frac{(\boldsymbol{T}_4^T(N)(\boldsymbol{W}_{o3} \boldsymbol{A}_2 + \boldsymbol{W}_{o4} \boldsymbol{A}_4) \boldsymbol{T}_4(N))_{ii}}{(\boldsymbol{T}_4^T(N) \boldsymbol{W}_{o4} \boldsymbol{T}_4(N))_{ii}}.
\end{aligned} \tag{20}$$

This diagonal error-feedback matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ leads to further reduction of the noise variance gain, i.e.,

$$I_{min}(\boldsymbol{D}) < I_{min}(\alpha(N)\boldsymbol{I}_m \oplus \beta(N)\boldsymbol{I}_n). \tag{21}$$

IV. A NUMERICAL EXAMPLE

Let a 2-D state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{3,3}$ where $d = 0.0$ 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 $I(\boldsymbol{0}) = 4.927082$, subject to the l_2-norm dynamic-range scaling constraints.

Now we apply the iterative optimization procedure to the above *optimal realization*. The proposed algorithm converges after nine iterations to $\alpha = 0.845965$, $\beta = 0.845177$, and a coordinate transformation matrix $\boldsymbol{T}(9) = \boldsymbol{T}_1(9) \oplus \boldsymbol{T}_4(9)$ with

$$\boldsymbol{T}_1(9) = \begin{bmatrix} -0.915091 & 0.162455 & 0.058237 \\ -0.470675 & 0.116512 & 0.581377 \\ -0.094509 & 0.398443 & 0.657685 \end{bmatrix}$$

$$\boldsymbol{T}_4(9) = \begin{bmatrix} 0.157780 & -0.908860 & -0.069259 \\ 0.117342 & -0.463086 & -0.587994 \\ 0.403126 & -0.087756 & -0.657977 \end{bmatrix}$$

which yield the noise variance gain $I(\alpha \boldsymbol{I}_3 \oplus \beta \boldsymbol{I}_3) = 1.665203$. After 3-bit quantization (integer quantization), this scalar error-feedback matrix gives $I_{min}(\alpha \boldsymbol{I}_3 \oplus \beta \boldsymbol{I}_3) = 1.670002$ (1.803185).

Next, a refined solution which offers further reduced noise variance gain is deduced by calculating an optimal diagonal error-feedback matrix for the optimal realization $(\boldsymbol{T}^{o-1}\boldsymbol{A}\boldsymbol{T}^o, \boldsymbol{T}^{o-1}\boldsymbol{b}, \boldsymbol{c}\boldsymbol{T}^o, d)_{3,3}$. In this case, the optimal diagonal error-feedback matrix $\boldsymbol{D} = \boldsymbol{D}_1 \oplus \boldsymbol{D}_4$ is obtained using (20) as

$$\boldsymbol{D}_1 = \text{diag}\{0.310843,\ 0.931481,\ 0.773711\}$$

$$\boldsymbol{D}_4 = \text{diag}\{0.931135,\ 0.307381,\ 0.772593\}$$

which yields $I_{min}(\boldsymbol{D}) = 1.497455$. After 3-bit quantization (integer quantization), the above diagonal error-feedback matrix gives $I(\boldsymbol{D}) = 1.512761$ (1.629652).

VI. CONCLUSION

The minimization of roundoff noise in 2-D state-space digital filters has been investigated by means of joint optimization of error feedback/coordinate transformation. An iterative procedure for minimizing the roundoff noise in a 2-D digital filter has also been developed by jointly optimizing a scalar error-feedback matrix and a coordinate transformation matrix subject to the usual l_2-norm dynamic-range scaling constraints. Simulation results have been presented to illustrate the validity of our proposed algorithm.

1. REFERENCES

[1] M. Kawamata and T. Higuchi, "A unified study on the roundoff noise in 2-D state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 724-730, July 1986.

[2] W.-S. Lu and A. Antoniou, "Synthesis of 2-D state-space fixed-point digital filter structures with minimum roundoff noise," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 965-973, Oct. 1986.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] T. Hinamoto and H. Ohnishi, "Minimization of roundoff noise in state-space digital filters using error feedback and coordinate transformation, " *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'02), vol. 5, pp. 709-712, May 2002.