Roundoff Noise Minimization for a Class of 2-D State-Space Digital Filters Using Separate/Joint Optimization of Error Feedback and Realization

Takao Hinamoto
Graduate School of Engineering
Hiroshima University
Higashi-Hiroshima 739-8527, Japan
Email: hinamoto@hiroshima-u.ac.jp

Toru Oumi
Graduate School of Engineering
Hiroshima University
Higashi-Hiroshima 739-8527, Japan
Email: oumi@hiroshima-u.ac.jp

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

***Abstract*— Techniques for the separate/joint optimization of error-feedback and realization are developed to minimize the roundoff noise subject to L_2-norm dynamic-range scaling constraints for a class of 2-D state-space digital filters. In the joint optimization, the problem at hand is converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained problem obtained is then solved by applying an efficient quasi-Newton algorithm. A numerical example is presented to illustrate the utility of the proposed techniques.**

I. INTRODUCTION

When implementing IIR digital filters in fixed-point arithmetic, the problem of reducing the effects of roundoff noise (RN) at the filter output is of critical importance. Error feedback (EF) is a useful tool for reducing finite-word-length (FWL) effects in IIR digital filters. Many EF techniques have been proposed for 2-D IIR digital filters [1]-[5]. Another useful approach is to construct the 2-D state-space filter structure for the RN gain to be minimized by applying a linear transformation to the state-space coordinates subject to L_2-norm dynamic-range scaling constraints [6],[7]. As a natural extension of the fore-mentioned methods, efforts have been made to develop new methods that combine EF and coordinate transformation for better performance in the RN reduction. In [8], separately/jointly-optimized iterative algorithms have been developed for 2-D filters with EF matrix.

This paper investigates the problems of separately/jointly optimizing EF and realization subject to L_2-norm dynamic-range scaling constraints for 2-D state-space digital filters described by the Fornasini-Marchesini second model. The former is solved analytically and the latter iteratively by applying an efficient quasi-Newton algorithm [9]. Computer simulation results demonstrate the validity of the proposed techniques.

II. ROUNDOFF NOISE ANALYSIS AND SCALING

Consider a 2-D IIR digital filter that is described by the Fornasini-Marchesini second local state-space (LSS) model $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_n$:

$$\begin{aligned} \boldsymbol{x}(i,\,j) &= \boldsymbol{A}_1\boldsymbol{x}(i-1,\,j) + \boldsymbol{A}_2\boldsymbol{x}(i,\,j-1) \\ &\quad + \boldsymbol{b}_1 u(i-1,\,j) + \boldsymbol{b}_2 u(i,\,j-1) \\ y(i,\,j) &= \boldsymbol{c}\boldsymbol{x}(i,\,j) + du(i,\,j) \end{aligned} \tag{1}$$

where $\boldsymbol{x}(i,j)$ is an $n\times 1$ local state vector, $u(i,j)$ is a scalar input, $y(i,j)$ is a scalar output, and $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}$, and d are real matrices of appropriate dimensions. The LSS model in (1) is assumed to be stable, locally controllable and locally observable.

Due to finite register sizes, FWL constraints are imposed on the local state vector, input, output, and coefficients in the realization $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_n$. Assuming that the quantization is carried out before matrix-vector multiplication, the actual 2-D FWL filter of (1) with EF and error feedforward can be implemented as

$$\begin{aligned} \tilde{\boldsymbol{x}}(i,\,j) &= \boldsymbol{A}_1\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i-1,\,j)] + \boldsymbol{A}_2\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,\,j-1)] \\ &\quad + \boldsymbol{b}_1 u(i-1,\,j) + \boldsymbol{b}_2 u(i,\,j-1) \\ &\quad + \boldsymbol{D}_1\boldsymbol{e}(i-1,\,j) + \boldsymbol{D}_2\boldsymbol{e}(i,\,j-1) \\ \tilde{y}(i,\,j) &= \boldsymbol{c}\,\boldsymbol{Q}[\tilde{\boldsymbol{x}}(i,\,j)] + du(i,\,j) + \boldsymbol{h}\boldsymbol{e}(i,\,j) \end{aligned} \tag{2}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are $n\times n$ EF matrices, $\boldsymbol{h}$ is a $1\times n$ error-feedforward vector,

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

and each component of matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}$ and d assumes an exact fractional B_c-bit representation. The FWL local state vector $\tilde{\boldsymbol{x}}(i,j)$ and 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 the multiplications and additions, where the sign bit is not counted. The quantization error $\boldsymbol{e}(i,j)$ is modeled as a zero-mean noise process of covariance $\sigma^2\boldsymbol{I}_n$ with

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

Subtracting (2) from (1) yields

$$\begin{aligned} \Delta\boldsymbol{x}(i,\,j) &= \boldsymbol{A}_1\Delta\boldsymbol{x}(i-1,\,j) + \boldsymbol{A}_2\Delta\boldsymbol{x}(i,\,j-1) \\ &\quad + (\boldsymbol{A}_1 - \boldsymbol{D}_1)\boldsymbol{e}(i-1,\,j) \\ &\quad + (\boldsymbol{A}_2 - \boldsymbol{D}_2)\boldsymbol{e}(i,\,j-1) \\ \Delta y(i,\,j) &= \boldsymbol{c}\Delta\boldsymbol{x}(i,\,j) + (\boldsymbol{c}-\boldsymbol{h})\boldsymbol{e}(i,\,j) \end{aligned} \tag{3}$$

where

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

1-4244-0173-9/06/$20.00 ©2006 IEEE.

The 2-D transfer function from the quantization error $\boldsymbol{e}(i,j)$ to the filter output $\Delta y(i,j)$ is given by

$$\begin{aligned}\boldsymbol{G}(z_1,z_2) &= \boldsymbol{c}\left(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2\right)^{-1} \\ &\quad \cdot \left[z_1^{-1}(\boldsymbol{A}_1-\boldsymbol{D}_1) + z_2^{-1}(\boldsymbol{A}_2-\boldsymbol{D}_2)\right] + \boldsymbol{c} - \boldsymbol{h}.\end{aligned} \tag{4}$$

For the 2-D filter in (2), the noise gain $I(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{h}) = \sigma_{out}^2/\sigma^2$ can be evaluated by

$$\begin{aligned}&I(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{h}) \\ &\quad = \mathrm{tr}\left[\frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}\boldsymbol{G}^*(z_1,\,z_2)\boldsymbol{G}(z_1,\,z_2)\frac{dz_1\,dz_2}{z_1 z_2}\right]\end{aligned} \tag{5}$$

where σ_{out}^2 denotes noise variance at the output, and $\Gamma_i = \{z_i : |z_i| = 1\}$ for $i = 1, 2$.

Let the transition matrix $\boldsymbol{A}^{(i,\,j)}$ be defined by

$$\left(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2\right)^{-1} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{A}^{(i,\,j)}z_1^{-i}z_2^{-j} \tag{6}$$

where $i, j \geq 0$. Then the following properties holds:

$$\begin{aligned}\boldsymbol{A}^{(0,\,0)} &= \boldsymbol{I}_n, \qquad \boldsymbol{A}^{(i,\,j)} = \boldsymbol{0} \quad \text{for } i<0 \text{ or } j<0 \\ \boldsymbol{A}^{(i,\,j)} &= \boldsymbol{A}_1\boldsymbol{A}^{(i-1,\,j)} + \boldsymbol{A}_2\boldsymbol{A}^{(i,\,j-1)} \\ &= \boldsymbol{A}^{(i-1,\,j)}\boldsymbol{A}_1 + \boldsymbol{A}^{(i,\,j-1)}\boldsymbol{A}_2 \quad \text{for } i,j>0.\end{aligned} \tag{7}$$

Substituting (6) into (4) yields

$$\begin{aligned}\boldsymbol{G}(z_1,z_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}&\big[\boldsymbol{c}\boldsymbol{A}^{(i-1,\,j)}(\boldsymbol{A}_1-\boldsymbol{D}_1) \\ &+\boldsymbol{c}\boldsymbol{A}^{(i,\,j-1)}(\boldsymbol{A}_2-\boldsymbol{D}_2)\big]z_1^{-i}z_2^{-j} \\ &+\boldsymbol{c}-\boldsymbol{h}.\end{aligned} \tag{8}$$

Substituting (8) into (5), it follows that

$$I(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{h}) = J_1(\boldsymbol{D}_1,\boldsymbol{D}_2) + \mathrm{tr}[(\boldsymbol{c}-\boldsymbol{h})^T(\boldsymbol{c}-\boldsymbol{h})] \tag{9}$$

where

$$\begin{aligned}&J_1(\boldsymbol{D}_1,\boldsymbol{D}_2) \\ &\quad = \mathrm{tr}\Big[\,[(\boldsymbol{A}_1-\boldsymbol{D}_1)^T,\,(\boldsymbol{A}_2-\boldsymbol{D}_2)^T]\boldsymbol{W}'\begin{bmatrix}\boldsymbol{A}_1-\boldsymbol{D}_1\\ \boldsymbol{A}_2-\boldsymbol{D}_2\end{bmatrix}\Big] \\ &\quad = \mathrm{tr}\Big[(\boldsymbol{A}_1-\boldsymbol{D}_1)^T\boldsymbol{W}_o(\boldsymbol{A}_1-\boldsymbol{D}_1) \\ &\qquad +(\boldsymbol{A}_2-\boldsymbol{D}_2)^T\boldsymbol{W}^T(\boldsymbol{A}_1-\boldsymbol{D}_1) \\ &\qquad +(\boldsymbol{A}_1-\boldsymbol{D}_1)^T\boldsymbol{W}(\boldsymbol{A}_2-\boldsymbol{D}_2) \\ &\qquad +(\boldsymbol{A}_2-\boldsymbol{D}_2)^T\boldsymbol{W}_o(\boldsymbol{A}_2-\boldsymbol{D}_2)\Big].\end{aligned}$$

Here, the $2n \times 2n$ matrix $\boldsymbol{W}'$ is defined by

$$\begin{aligned}\boldsymbol{W}' &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\begin{bmatrix}(\boldsymbol{c}\boldsymbol{A}^{(i-1,\,j)})^T\\ (\boldsymbol{c}\boldsymbol{A}^{(i,\,j-1)})^T\end{bmatrix}\Big[\boldsymbol{c}\boldsymbol{A}^{(i-1,\,j)},\ \boldsymbol{c}\boldsymbol{A}^{(i,\,j-1)}\Big] \\ &= \begin{bmatrix}\boldsymbol{W}_o & \boldsymbol{W}\\ \boldsymbol{W}^T & \boldsymbol{W}_o\end{bmatrix}\end{aligned} \tag{10}$$

where matrix $\boldsymbol{W}_o$ is the local observability Gramian of the LSS model in (1). In the case when there is no EF but error feedforward exists, it follows from (9) that

$$\begin{aligned}I(\boldsymbol{0},\boldsymbol{0},\boldsymbol{c}) &= \mathrm{tr}\Big[\,[\boldsymbol{A}_1^T,\,\boldsymbol{A}_2^T]\boldsymbol{W}'\begin{bmatrix}\boldsymbol{A}_1\\ \boldsymbol{A}_2\end{bmatrix}\Big] \\ &= \mathrm{tr}\Big[\boldsymbol{A}_1^T\boldsymbol{W}_o\boldsymbol{A}_1 + \boldsymbol{A}_2^T\boldsymbol{W}^T\boldsymbol{A}_1 \\ &\qquad + \boldsymbol{A}_1^T\boldsymbol{W}\boldsymbol{A}_2 + \boldsymbol{A}_2^T\boldsymbol{W}_o\boldsymbol{A}_2\Big].\end{aligned} \tag{11}$$

The local controllability Gramian $\boldsymbol{K}_c$ is defined by

$$\boldsymbol{K}_c = \sum_{k=1}^{\infty}\sum_{i=0}^{k}\boldsymbol{f}(i,\,k-i)\boldsymbol{f}^T(i,\,k-i) \tag{12}$$

where

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

The LSS model in (1) is said to satisfy L_2-norm dynamic-range scaling constraints provided that

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

where $(\boldsymbol{K}_c)_{ii}$ denotes the ith entry of matrix $\boldsymbol{K}_c$.

III. SEPARATE OPTIMIZATION OF REALIZATION AND ERROR FEEDBACK

Applying a coordinate transformation defined by

$$\overline{\boldsymbol{x}}(i,\,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,\,j) \tag{14}$$

to the LSS model $(\boldsymbol{A}_1,\boldsymbol{A}_2,\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{c},d)_n$ in (1), we obtain a new realization $(\overline{\boldsymbol{A}}_1,\overline{\boldsymbol{A}}_2,\overline{\boldsymbol{b}}_1,\overline{\boldsymbol{b}}_2,\overline{\boldsymbol{c}},d)_n$ characterized by

$$\begin{aligned}\overline{\boldsymbol{A}}_1 &= \boldsymbol{T}^{-1}\boldsymbol{A}_1\boldsymbol{T}, \qquad \overline{\boldsymbol{A}}_2 = \boldsymbol{T}^{-1}\boldsymbol{A}_2\boldsymbol{T} \\ \overline{\boldsymbol{b}}_1 &= \boldsymbol{T}^{-1}\boldsymbol{b}_1, \quad \overline{\boldsymbol{b}}_2 = \boldsymbol{T}^{-1}\boldsymbol{b}_2, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}\end{aligned} \tag{15}$$

where $\boldsymbol{T}$ is an $n \times n$ nonsingular matrix. The Gramians $\overline{\boldsymbol{K}}_c$, $\overline{\boldsymbol{W}}_o$ and $\overline{\boldsymbol{W}}$ in the new realization can then be written as

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \quad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}, \quad \overline{\boldsymbol{W}} = \boldsymbol{T}^T\boldsymbol{W}\boldsymbol{T} \tag{16}$$

respectively. If the L_2-norm dynamic-range scaling constraints are imposed on the new realization, then we have

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, n. \tag{17}$$

For the new realization with no EF and error feedforward, we consider the problem of minimizing a measure

$$M(\boldsymbol{P},\lambda) = \mathrm{tr}[\boldsymbol{V}\boldsymbol{P}] + \lambda(\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n) \tag{18}$$

with respect to $n \times n$ nonsingular matrix $\boldsymbol{P}$ and scalar λ where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$, λ is a Lagrange multiplier, and

$$\boldsymbol{V} = \boldsymbol{A}_1^T\boldsymbol{W}_o\boldsymbol{A}_1 + \boldsymbol{A}_2^T\boldsymbol{W}^T\boldsymbol{A}_1 + \boldsymbol{A}_1^T\boldsymbol{W}\boldsymbol{A}_2 + \boldsymbol{A}_2^T\boldsymbol{W}_o\boldsymbol{A}_2.$$

We compute

$$\begin{aligned}\frac{\partial J(\boldsymbol{P},\lambda)}{\partial \boldsymbol{P}} &= \boldsymbol{V} - \lambda\boldsymbol{P}^{-1}\boldsymbol{K}_c\boldsymbol{P}^{-1} \\ \frac{\partial J(\boldsymbol{P},\lambda)}{\partial \lambda} &= \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n.\end{aligned} \tag{19}$$

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

$$\boldsymbol{P}\,\boldsymbol{V}\,\boldsymbol{P}=\lambda \boldsymbol{K}_c, \qquad \text{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}]=n. \tag{20}$$

It follows from (20) that

$$\begin{aligned}\boldsymbol{P}&=\sqrt{\lambda}\boldsymbol{V}^{-\frac{1}{2}}[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}\\ \frac{1}{\sqrt{\lambda}}\text{tr}[\boldsymbol{K}_c\boldsymbol{V}]^{\frac{1}{2}}&=\frac{1}{\sqrt{\lambda}}\left(\sum_{i=1}^{n}\theta_i\right)=n\end{aligned} \tag{21}$$

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

$$\boldsymbol{P}=\frac{1}{n}\left(\sum_{i=1}^{n}\theta_i\right)\boldsymbol{V}^{-\frac{1}{2}}[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}]^{\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}. \tag{22}$$

Substituting (22) into (18) yields the minimum value of $M(\boldsymbol{P},\lambda)$ as

$$\min_{\boldsymbol{P},\lambda} M(\boldsymbol{P},\lambda)=\frac{1}{n}\left(\sum_{i=1}^{n}\theta_i\right)^2. \tag{23}$$

Note that matrix $\boldsymbol{T}$ assumes the form

$$\boldsymbol{T}=\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{U} \tag{24}$$

where $\boldsymbol{P}^{1/2}$ is the square root of $\boldsymbol{P}$ obtained above, and $\boldsymbol{U}$ is the $n\times n$ orthogonal matrix such that (17) is satisfied.

If the coordinate transformation in (14) is applied to the LSS model in (1), then (9) is changed to

$$\overline{I}(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{h},\boldsymbol{T})=J_2(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{T})+\text{tr}[(\overline{\boldsymbol{c}}-\boldsymbol{h})^T(\overline{\boldsymbol{c}}-\boldsymbol{h})] \tag{25}$$

where

$$\begin{aligned}J_2(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{T})=\text{tr}\Big[&(\overline{\boldsymbol{A}}_1-\boldsymbol{D}_1)^T\overline{\boldsymbol{W}}_o(\overline{\boldsymbol{A}}_1-\boldsymbol{D}_1)\\ &+(\overline{\boldsymbol{A}}_2-\boldsymbol{D}_2)^T\overline{\boldsymbol{W}}^T(\overline{\boldsymbol{A}}_1-\boldsymbol{D}_1)\\ &+(\overline{\boldsymbol{A}}_1-\boldsymbol{D}_1)^T\overline{\boldsymbol{W}}(\overline{\boldsymbol{A}}_2-\boldsymbol{D}_2)\\ &+(\overline{\boldsymbol{A}}_2-\boldsymbol{D}_2)^T\overline{\boldsymbol{W}}_o(\overline{\boldsymbol{A}}_2-\boldsymbol{D}_2)\Big].\end{aligned}$$

Case 1: $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are general matrices

In this case, we select the matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ as

$$\boldsymbol{D}_1=\overline{\boldsymbol{A}}_1, \qquad \boldsymbol{D}_2=\overline{\boldsymbol{A}}_2. \tag{26}$$

Case 2: $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices

We define

$$\begin{aligned}\boldsymbol{D}_1&=\text{diag}\{\alpha_1,\alpha_2,\cdots,\alpha_n\}\\ \boldsymbol{D}_2&=\text{diag}\{\beta_1,\beta_2,\cdots,\beta_n\}.\end{aligned} \tag{27}$$

From $\partial J_2(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{T})/\partial\alpha_i=0$ and $\partial J_2(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{T})/\partial\beta_i=0$, it is derived that for $i=1,2,\cdots,n$

$$\begin{aligned}\alpha_i&=\frac{\overline{\boldsymbol{W}}_o(i,i)\boldsymbol{M}_1(i,i)-\overline{\boldsymbol{W}}(i,i)\boldsymbol{M}_2(i,i)}{\overline{\boldsymbol{W}}_o(i,i)^2-\overline{\boldsymbol{W}}(i,i)^2}\\ \beta_i&=\frac{\overline{\boldsymbol{W}}_o(i,i)\boldsymbol{M}_2(i,i)-\overline{\boldsymbol{W}}(i,i)\boldsymbol{M}_1(i,i)}{\overline{\boldsymbol{W}}_o(i,i)^2-\overline{\boldsymbol{W}}(i,i)^2}\end{aligned} \tag{28}$$

where $\boldsymbol{X}(i,j)$ denotes the ijth element of matrix $\boldsymbol{X}$ and

$$\begin{aligned}\boldsymbol{M}_1&=\overline{\boldsymbol{W}}_o\overline{\boldsymbol{A}}_1+\overline{\boldsymbol{W}}\;\overline{\boldsymbol{A}}_2\\ \boldsymbol{M}_2&=\overline{\boldsymbol{W}}_o\overline{\boldsymbol{A}}_2+\overline{\boldsymbol{W}}^T\overline{\boldsymbol{A}}_1.\end{aligned}$$

Case 3: $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are scalar matrices

With scalars α and β, we define

$$\boldsymbol{D}_1=\alpha\boldsymbol{I}_n, \qquad \boldsymbol{D}_2=\beta\boldsymbol{I}_n. \tag{29}$$

From $\partial J_2(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{T})/\partial\alpha=0$ and $\partial J_2(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{T})/\partial\beta=0$, we obtain

$$\begin{aligned}\alpha&=\frac{\text{tr}[\overline{\boldsymbol{W}}_o]\text{tr}[\boldsymbol{M}_1]-\text{tr}[\overline{\boldsymbol{W}}]\text{tr}[\boldsymbol{M}_2]}{(\text{tr}[\overline{\boldsymbol{W}}_o])^2-(\text{tr}[\overline{\boldsymbol{W}}])^2}\\ \beta&=\frac{\text{tr}[\overline{\boldsymbol{W}}_o]\text{tr}[\boldsymbol{M}_2]-\text{tr}[\overline{\boldsymbol{W}}]\text{tr}[\boldsymbol{M}_1]}{(\text{tr}[\overline{\boldsymbol{W}}_o])^2-(\text{tr}[\overline{\boldsymbol{W}}])^2}.\end{aligned} \tag{30}$$

IV. JOINT OPTIMIZATION OF ERROR FEEDBACK AND REALIZATION

Define

$$\hat{\boldsymbol{T}}=\boldsymbol{T}^T\boldsymbol{K}_c^{-\frac{1}{2}}. \tag{31}$$

Then (17) can be written as

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii}=1 \;\text{ for }\; i=1,2,\cdots,n. \tag{32}$$

The constraints in (32) simply state that each column in matrix $\hat{\boldsymbol{T}}^{-1}$ must be a unity vector. These are satisfied if $\hat{\boldsymbol{T}}^{-1}$ assumes the form

$$\hat{\boldsymbol{T}}^{-1}=\left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||},\frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||},\cdots,\frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\right] \tag{33}$$

where $\boldsymbol{t}_i$'s for $i=1,2,\cdots,n$ are $n\times 1$ real vectors. In such a case, (25) can be expressed as

$$\hat{I}(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{h},\hat{\boldsymbol{T}})=J_3(\boldsymbol{D}_1,\boldsymbol{D}_2,\hat{\boldsymbol{T}})+\text{tr}[(\hat{\boldsymbol{c}}-\boldsymbol{h})^T(\hat{\boldsymbol{c}}-\boldsymbol{h})] \tag{34}$$

where

$$\begin{aligned}J_3(\boldsymbol{D}_1,\boldsymbol{D}_2,\hat{\boldsymbol{T}})=\text{tr}\Big[&(\hat{\boldsymbol{A}}_1-\boldsymbol{D}_1)^T\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{A}}_1-\boldsymbol{D}_1)\\ &+(\hat{\boldsymbol{A}}_2-\boldsymbol{D}_2)^T\hat{\boldsymbol{W}}^T(\hat{\boldsymbol{A}}_1-\boldsymbol{D}_1)\\ &+(\hat{\boldsymbol{A}}_1-\boldsymbol{D}_1)^T\hat{\boldsymbol{W}}(\hat{\boldsymbol{A}}_2-\boldsymbol{D}_2)\\ &+(\hat{\boldsymbol{A}}_2-\boldsymbol{D}_2)^T\hat{\boldsymbol{W}}_o(\hat{\boldsymbol{A}}_2-\boldsymbol{D}_2)\Big]\end{aligned}$$

with

$$\begin{aligned}\hat{\boldsymbol{A}}_1&=\hat{\boldsymbol{T}}^{-T}(\boldsymbol{K}_c^{-\frac{1}{2}}\boldsymbol{A}_1\boldsymbol{K}_c^{\frac{1}{2}})\,\hat{\boldsymbol{T}}^T\\ \hat{\boldsymbol{A}}_2&=\hat{\boldsymbol{T}}^{-T}(\boldsymbol{K}_c^{-\frac{1}{2}}\boldsymbol{A}_2\boldsymbol{K}_c^{\frac{1}{2}})\,\hat{\boldsymbol{T}}^T, \quad \hat{\boldsymbol{c}}=(\boldsymbol{c}\boldsymbol{K}_c^{\frac{1}{2}})\hat{\boldsymbol{T}}^T\\ \hat{\boldsymbol{W}}_o&=\hat{\boldsymbol{T}}(\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}_o\boldsymbol{K}_c^{\frac{1}{2}})\hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{W}}=\hat{\boldsymbol{T}}(\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{K}_c^{\frac{1}{2}})\hat{\boldsymbol{T}}^T.\end{aligned}$$

When selecting vector $\boldsymbol{h}$ as $\boldsymbol{h}=\hat{\boldsymbol{c}}$, the problem of obtaining matrices $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\boldsymbol{T}$ that minimize $J_2(\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{T})$ in (25) subject to the scaling constraints in (17) can be converted into an unconstrained optimization problem of obtaining matrices $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\hat{\boldsymbol{T}}$ that minimize $J_3(\boldsymbol{D}_1,\boldsymbol{D}_2,\hat{\boldsymbol{T}})$ in (34).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\hat{\boldsymbol{T}}$. Then, $J_3(\boldsymbol{D}_1,\boldsymbol{D}_2,\hat{\boldsymbol{T}})$ is a function

of $\boldsymbol{x}$, denoted by $J_3(\boldsymbol{x})$. The algorithm starts with a trivial initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\boldsymbol{D}_1 = \boldsymbol{D}_2 = \hat{\boldsymbol{T}} = \boldsymbol{I}_n$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [9]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{35}$$

where

$$\boldsymbol{d}_k = -\boldsymbol{S}_k \nabla J_3(\boldsymbol{x}_k), \qquad \alpha_k = arg \min_{\alpha} J_3(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$$

$$\boldsymbol{S}_{k+1} = \boldsymbol{S}_k + \left(1 + \frac{\boldsymbol{\gamma}_k^T \boldsymbol{S}_k \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k}\right) \frac{\boldsymbol{\delta}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} - \frac{\boldsymbol{\delta}_k \boldsymbol{\gamma}_k^T \boldsymbol{S}_k + \boldsymbol{S}_k \boldsymbol{\gamma}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k}$$

$$\boldsymbol{S}_0 = \boldsymbol{I},\ \boldsymbol{\delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k,\ \boldsymbol{\gamma}_k = \nabla J_3(\boldsymbol{x}_{k+1}) - \nabla J_3(\boldsymbol{x}_k).$$

Here, $\nabla J_3(\boldsymbol{x})$ is the gradients of $J_3(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, and $\boldsymbol{S}_k$ is a positive-definite approximation of the inverse Hessian matrix of $J_3(\boldsymbol{x})$. This iteration process continues until

$$|J_3(\boldsymbol{x}_{k+1}) - J_3(\boldsymbol{x}_k)| < \varepsilon \tag{36}$$

where $\varepsilon > 0$ is a prescribed tolerance. If the iteration is terminated at step k, $\boldsymbol{x}_k$ is viewed as a solution point.

When $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are general matrices, vector $\boldsymbol{x}$ consists of matrix $\hat{\boldsymbol{T}}$ only. After obtaing matrix $\hat{\boldsymbol{T}}$, we select matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ as

$$\boldsymbol{D}_1 = \hat{\boldsymbol{A}}_1, \qquad \boldsymbol{D}_2 = \hat{\boldsymbol{A}}_2. \tag{37}$$

The gradient of $J(\boldsymbol{x})$ with respect to the ijth element of $\hat{\boldsymbol{T}}$ is found to be

$$\frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} = \lim_{\Delta \to 0} \frac{J(\hat{\boldsymbol{T}}_{ij}) - J(\hat{\boldsymbol{T}})}{\Delta} \tag{38}$$

where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i,j)th component, and is given by [10]

$$\hat{\boldsymbol{T}}_{ij} = \hat{\boldsymbol{T}} + \frac{\Delta \hat{\boldsymbol{T}} \boldsymbol{g}_{ij} \boldsymbol{e}_j^T \hat{\boldsymbol{T}}}{1 - \Delta \boldsymbol{e}_j^T \hat{\boldsymbol{T}} \boldsymbol{g}_{ij}}$$

and $\boldsymbol{g}_{ij}$ is computed using

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

V. A NUMERICAL EXAMPLE

Consider a stable, locally controllable, and locally observable 2-D state-space digital filter with order $n = 4$ specified by

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & 0 & 0 & -0.00411 \\ 1 & 0 & 0 & 0.08007 \\ 0 & 1 & 0 & -0.42458 \\ 0 & 0 & 1 & 1.04460 \end{bmatrix}, \qquad \boldsymbol{b}_1 = \begin{bmatrix} -0.01452 \\ 0.01234 \\ 0.02054 \\ 0.04762 \end{bmatrix}$$

$$\boldsymbol{A}_2 = \begin{bmatrix} -0.22608 & -0.40594 & -0.30955 & -0.14469 \\ 1.61428 & 1.61040 & 1.02336 & 0.43872 \\ 0.10054 & -0.60615 & -0.45322 & -0.31019 \\ -0.00723 & 0.24580 & 0.38668 & 0.56289 \end{bmatrix}$$

$$\boldsymbol{b}_2 = \begin{bmatrix} 0.01189 & 0.02351 & -0.00637 & 0.02094 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}, \qquad d = 0.00943.$$

After carrying out the L_2-scaling for the above LSS model with a diagonal coordinate transformation matrix, the noise gain of the scaled LSS model with error feedforward was computed as $I(\boldsymbol{0}, \boldsymbol{0}, \mathbf{c}) = 76.641884$. Next, the matrix $\boldsymbol{P}$ was derived from (22) and substituting it into (18) produced $M(\boldsymbol{P}, \lambda) = 3.230958$. The other results obtained by applying the proposed technique are summarized in Tables I and II.

VI. CONCLUSION

The separate/joint optimization of EF and realization has been investigated to minimize RN subject to L_2-scaling constraints for a class of 2-D state-space digital filters. It has been shown that the problem in the joint optimization can be converted into an unconstrained optimization problem by using linear algebraic techniques. An efficient quasi-Newton algorithm has then been employed to solve the unconstrained optimization problem iteratively. Our computer simulation results have demonstrated the effectiveness of the proposed techniques.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] T. Hinamoto, N. Kuroda and T. Kuma, "Error feedback for noise reduction in 2-D digital filers with quadrantally symmetric or antisymmetric coefficients," *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS'97), vol. 4, pp. 2461-2464, June 1997.

[5] 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.

[6] 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.

[7] 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.

[8] T. Hinamoto, K. Higashi and W.-S. Lu, "Separate/joint optimization of error feedback and coordinate transformation for roundoff noise minimization in two-dimensional state-space digital filters," *IEEE Trans. Signal Processing*, vol. 51, pp.2436-2445, Sept. 2003.

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

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

TABLE I
ROUNDOFF NOISE GAIN IN *SEPARATE OPTIMIZATION*

Error Feedback Matrices	General	Diagonal	Scalar
Infinite Precision	0	0.454620	0.469953
3-Bit Quantization	0.049936	0.461004	0.481157
Integer Quantization	3.100355	2.217879	1.586120

TABLE II
ROUNDOFF NOISE GAIN IN *JOINT OPTIMIZATION*

Error Feedback Matrices	General	Diagonal	Scalar
Infinite Precision	0	0.186190	0.233562
3-Bit Quantization	0.049936	0.216724	0.243361
Integer Quantization	3.100355	1.736732	1.808645