JOINTLY OPTIMIZED HIGH-ORDER ERROR FEEDBACK AND REALIZATION FOR ROUNDOFF NOISE MINIMIZAION IN STATE-SPACE DIGITAL FILTERS

Takao Hinamoto, Akimitsu Doi
Hiroshima Institute of Technology
Hiroshima 731-5193, Japan
Emails: 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*— The joint optimization problem of high-order error feedback and realization for state-space digital filters to minimize the effects of roundoff noise at the filter output subject to l_2-scaling constraints is investigated. The problem is converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained optimization problem at hand is then solved iteratively by applying an efficient quasi-Newton algorithm with closed-form formulas for key gradient evaluation. Finally, a numerical example is presented to illustrate the utility of the proposed technique.**

1. INTRODUCTION

The problem of reducing the effects of roundoff noise at the filter output is of critical importance in the implementation of IIR digital filters in fixed-point arithmetic. Error feedback (EF) is found useful for the reduction of finite-word-length (FWL) effects in IIR digital filters, and many EF techniques have been reported in the past [1]-[10]. Alternatively, the roundoff noise can also be reduced by introducing a delta operator to IIR digital filters [11]-[13], or by applying a new structure based on the concept of polynomial operators for digital filter implementation [14]. Another useful approach is to construct the state-space filter structures for the roundoff noise gain to be minimized by applying a linear transformation to state-space coordinates subject to l_2-scaling constraints [15]-[18]. As a natural extension of the aforementioned methods, efforts have been made to develop new methods that combine EF and state-space realization, for achieving better performance [19]-[20]. Separately-optimized analytical algorithms have been proposed for state-space digital filters [19]. Jointly-optimized iterative algorithms have also been considered for filters with a general or scalar EF matrix [19]. In [20], a jointly-optimized iterative algorithm has been developed for state-space digital filters with a general, diagonal, or scalar EF matrix using a quasi-Newton method.

This paper investigates the problem of jointly optimizing *high-order* EF and realization for state-space digital filters to minimize the roundoff noise subject to l_2-scaling constraints. An iterative technique which relies on an efficient quasi-Newton algorithm [21] is developed. To this end, the constrained optimization problem encountered is converted into an unconstrained optimization problem by using linear-algebraic techniques. The proposed technique is applied to the cases where the high-order EF has diagonal or scalar matrices. A numerical example is presented to illustrate the proposed algorithm and demonstrate its performance.

2. PROBLEM STATEMENT

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

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

where $\boldsymbol{x}(k)$ is an $n \times 1$ state-variable vector, $u(k)$ is a scalar input, $y(k)$ is a scalar output, and $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}$ and d are real constant matrices of appropriate dimensions.

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

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

where $\boldsymbol{h}$ and $\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N$ are referred to as a $1 \times n$ *error-feedforward vector* and $n \times n$ *high-order error-feedback matrices*, respectively, and $\boldsymbol{e}(k) = \tilde{\boldsymbol{x}}(k) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)]$.

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

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

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

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

where $\Delta Y(z)$ and $\boldsymbol{E}(z)$ represent the z-transforms of $\Delta y(k)$ and $\boldsymbol{e}(k)$, respectively. $\boldsymbol{H}_e(z)$ in (4) is written as

$$\boldsymbol{H}_e(z) = \sum_{k=1}^{\infty} \boldsymbol{c}\left(\boldsymbol{A}^k - \sum_{i=1}^{N} \boldsymbol{A}^{k-i}\boldsymbol{D}_i\right) z^{-k} + \boldsymbol{c} - \boldsymbol{h} \quad (5)$$

where $\boldsymbol{A}^i = \boldsymbol{0}$ for $i < 0$. Next, we define the normalized noise gain $J_{e1}(\boldsymbol{h}, \boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N]$ as

$$J_{e1}(\boldsymbol{h}, \boldsymbol{D}) = \text{tr}\left[\frac{1}{2\pi j}\oint_{|z|=1} \boldsymbol{H}_e^*(z)\boldsymbol{H}_e(z)\frac{dz}{z}\right] \quad (6)$$

Substituting (5) into (6) yields

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

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

$$\boldsymbol{W}_o = \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}. \quad (8)$$

Assuming that $\boldsymbol{D}_1, D_2, \cdots, D_N$ are diagonal, (7) can be written as

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

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

$$\boldsymbol{K}_c = \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T. \quad (10)$$

A different yet equivalent state-space description of (1), $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$, can be obtained via a coordinate transformation $\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k)$ where

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

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

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

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

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

The problem being considered is to design the *high-order* EF diagonal matrices $\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N$ as well as the optimal coordinate transformation matrix $\boldsymbol{T}$ that jointly minimize

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

subject to l_2-scaling constraints in (13) where the error feedforward vector $\boldsymbol{h}$ is assumed to be chosen as $\boldsymbol{h} = \overline{\boldsymbol{c}}$.

3. JOINT OPTIMIZATION OF HIGH-ORDER ERROR FEEDBACK AND REALIZATION

To deal with (13), we define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}_c^{-\frac{1}{2}}. \quad (15)$$

The l_2-scaling constraints in (13) can then be written as

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

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

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

Substituting (15) into (14), we obtain

$$\begin{aligned} J(\hat{\boldsymbol{T}}, \boldsymbol{D}) = \text{tr}\Bigg[&\hat{\boldsymbol{A}}^T\hat{\boldsymbol{W}}_o\hat{\boldsymbol{A}} - 2\sum_{p=1}^{N}\hat{\boldsymbol{W}}_o\hat{\boldsymbol{A}}^p\boldsymbol{D}_p \\ &+ \sum_{p=1}^{N}\sum_{q=1}^{N}\hat{\boldsymbol{W}}_o\hat{\boldsymbol{A}}^{|p-q|}\boldsymbol{D}_p\boldsymbol{D}_q\Bigg] \end{aligned} \quad (18)$$

where

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

From the foregoing arguments, the problem of obtaining matrices $\boldsymbol{T}$ and $\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N$ that jointly minimize (14) subject to the l_2-scaling constraints in (13) is now converted into an unconstrained optimization problem of obtaining $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N$ that jointly minimize $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ in (18).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$ and $\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N$. Then $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ in (18) is a function of $\boldsymbol{x}$, denoted by $J(\boldsymbol{x})$. The proposed algorithm starts with an initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\hat{\boldsymbol{T}} = \boldsymbol{D}_1 = \boldsymbol{D}_2 = \cdots = \boldsymbol{D}_N = \boldsymbol{I}_n$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [21]

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

where

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

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

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

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

In what follows, we derive closed-form expressions of $\nabla J(\boldsymbol{x})$ for the cases where $\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N$ assume the form of diagonal or scalar matrices.

Case 1: $\boldsymbol{D}_p = \text{diag}\{d_{p1}, d_{p2}, \cdots, d_{pn}\}$ for $p = 1, 2, \cdots, N$

In this case, it follows that

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

where

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

Moreover,

$$\begin{aligned}\frac{\partial J(\boldsymbol{x})}{\partial d_{pi}} &= -2(\hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}}^p)_{ii} + 2 \sum_{q=1}^{N} d_{qi} (\hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}}^{|p-q|})_{ii} \\ & p = 1, 2, \cdots, N; \quad i = 1, 2, \cdots, n.\end{aligned} \tag{22}$$

Case 2: $\boldsymbol{D}_p = \alpha_p \boldsymbol{I}_n$ for $p = 1, 2, \cdots, N$

The gradient of $J(\boldsymbol{x})$ can be calculated as

$$\begin{aligned}\frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= 2\boldsymbol{e}_j^T \Bigg[\hat{\boldsymbol{A}}^T \hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}} - \sum_{p=1}^{N} \alpha_p (\hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}}^p + (\hat{\boldsymbol{A}}^T)^p \hat{\boldsymbol{W}}_o) \\ &+ \frac{1}{2} \sum_{p=1}^{N} \sum_{q=1}^{N} \alpha_p \alpha_q (\hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}}^{|p-q|} + (\hat{\boldsymbol{A}}^T)^{|p-q|} \hat{\boldsymbol{W}}_o) \Bigg] \hat{\boldsymbol{T}} \boldsymbol{g}_{ij} \\ \frac{\partial J(\boldsymbol{x})}{\partial \alpha_p} &= -2 \text{ tr}(\hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}}^p) + 2 \sum_{q=1}^{N} \alpha_q \text{tr}(\hat{\boldsymbol{W}}_o \hat{\boldsymbol{A}}^{|p-q|}) \\ & i, j = 1, 2, \cdots, n; \quad p = 1, 2, \cdots, N.\end{aligned} \tag{23}$$

4. AN ILLUSTRATIVE EXAMPLE

As a numerical example, consider a state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_3$ described by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.453770 & -1.556160 & 1.974860 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} 0 & 0 & 0.2420961 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0.095706 & 0.095086 & 0.327556 \end{bmatrix}$$

$$d = 0.015940.$$

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

$$\boldsymbol{K}_c = \begin{bmatrix} 1.000000 & 0.872501 & 0.562822 \\ 0.872501 & 1.000000 & 0.872501 \\ 0.562822 & 0.872501 & 1.000000 \end{bmatrix}$$

$$\boldsymbol{W}_o = \begin{bmatrix} 0.820742 & -2.035323 & 1.628159 \\ -2.035323 & 5.307270 & -4.264912 \\ 1.628159 & -4.264912 & 3.941488 \end{bmatrix}.$$

The noise gain of the filter with no error feedforward and no EF was then computed from (9) as $J_{e1}(\boldsymbol{0}, \boldsymbol{0}) = 10.069510$.

Case 1: $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices where $N = 2$

The quasi-Newton algorithm was applied to minimize (18) with tolerance $\varepsilon = 10^{-8}$ in (20). It took the algorithm 44 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.444866 & -0.214988 & 0.437178 \\ -1.309716 & 1.306211 & 0.465756 \\ -0.577484 & 0.237097 & 1.275346 \end{bmatrix}$$

$$\boldsymbol{D}_1 = \text{diag}\{\ 0.6405 \quad 1.2864 \quad 1.44310\ \}$$

$$\boldsymbol{D}_2 = \text{diag}\{\ -0.0242 \quad -0.6992 \quad -0.6984\ \}$$

with the noise gain $J(\hat{\boldsymbol{T}}, \boldsymbol{D}) = 0.003552$. The profile of $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ during the first 44 iterations of the algorithm is dipicted in Fig. 1.

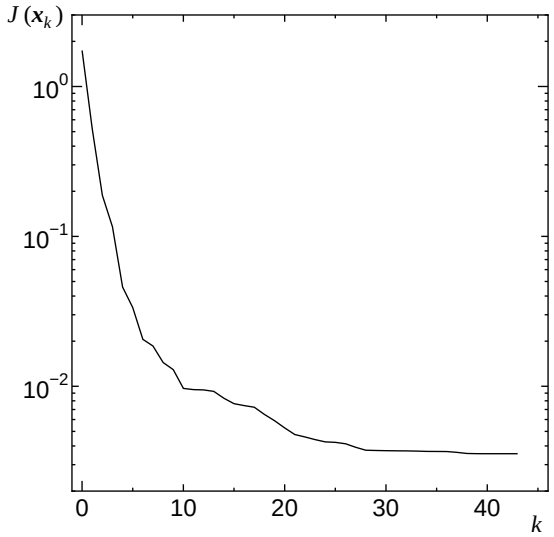

Fig. 1. Profile of $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ during the first 44 iterations.

Case 2: $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are scalar matrices where $N = 2$

The quasi-Newton algorithm with $\varepsilon = 10^{-8}$ was applied to minimize (18) for scalar high-order EF matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. It took the algorithm 19 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 1.219422 & -0.555940 & -0.042678 \\ -0.434261 & 1.235725 & 0.697171 \\ -0.079476 & -0.077024 & 1.153868 \end{bmatrix}$$

$$\boldsymbol{D}_1 = 1.1973\, \boldsymbol{I}_3, \qquad \boldsymbol{D}_2 = -0.5099\, \boldsymbol{I}_3$$

with the noise gain $J(\hat{\boldsymbol{T}}, \boldsymbol{D}) = 0.050379$. The profile of $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ during the first 19 iterations of the algorithm is shown in Fig. 2.

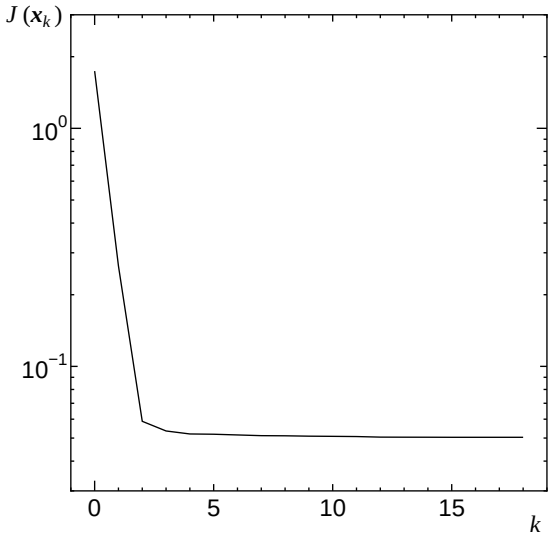

Fig. 2. Profile of $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ during the first 19 iterations.

The other results regarding the noise gain $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ in (18) were summarized in Table I where "Infinite Precision" shows the value of $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ derived from the optimal $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}$, and "3-Bit Quantization" means that of $J(\hat{\boldsymbol{T}}, \boldsymbol{D})$ where only each entry of the optimal matrix $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2]$ was rounded to a power-of-two representation with 3 bits after the binary point.

TABLE I
PERFORMANCE COMPARISON

N	$\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N$	Infinite Precision	3-Bit Quantization
1	Diagonal	0.049981	0.083396
	Scalar	0.092729	0.105829
2	Diagonal	0.003552	0.008444
	Scalar	0.050379	0.060035

The results reported above have clearly demonstrated that the use of high-order (i.e. $N > 1$) EF, when jointly optimized with state-space realization, can improve the performance of roundoff noise reduction in a significant manner.

5. CONCLUSION

The joint optimization problem of high-order EF and realization to minimize the effects of roundoff noise of state-space digital filters subject to l_2-scaling constraints has been investigated. It has been shown that the problem at hand can be converted into an unconstrained optimization problem by using linear algebraic techniques. Closed-form formulas for fast evaluation of the gradient of the objective function have been derived. An efficient quasi-Newton algorithm has been employed to solve the unconstrained optimization problem. The proposed technique has been applied to the cases where the high-order EF has diagonal or scalar matrices. A numerical example has demonstrated the 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] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Signal Processing*, vol. 41, pp. 629-637, Feb. 1993.

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

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

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

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

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

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

[18] L. B. Jackson, A. G. Lindgren and Y. Kim, "Optimal synthesis of second-order state-space structures for digital filters," *IEEE Trans. Circuits Syst.* , vol. 26, pp. 149-153, Mar. 1979.

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

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

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