JOINTLY OPTIMAL ERROR FEEDFORWARD, HIGH-ORDER ERROR FEEDBACK AND REALIZATION FOR ROUNDOFF NOISE MINIMIZATION IN IIR 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*— Joint optimization of error feedforward, high-order error feedback and state-space realization for minimizing roundoff noise at filter output subject to l_2-scaling constraints for state-space digital filters is investigated. Linear algebraic techniques that convert the problems at hand into an unconstrained optimization problem are explored, and an efficient quasi-Newton algorithm is then applied to solve the unconstrained optimization problem iteratively. In this connection, closed-form formulas are derived for fast and accurate gradient evaluation. Finally, a numerical example is presented to demonstrate that the high-order error feedback does offer much improved performance and that the proposed joint optimization is superior relative to a sequentially optimized system where the state-space coordinate transformation as well as error feedforward and high-order error feedback matrices are optimized separately.**

I. INTRODUCTION

In the implementation of IIR digital filters with fixed-point arithmetic, it is of critical importance to reduce the effects of roundoff noise at the filter's output. Error feedback (EF) is found effective for the reduction of finite-word-length (FWL) effects in IIR digital filters, and many EF methods have been proposed in the past [1]-[5]. Another useful approach for the minimization of roundoff noise gain is to synthesize the state-space filter structure by applying a linear transformation to state-space coordinates subject to l_2-scaling constraints [6]-[9]. 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 [10],[11]. Separately and jointly optimized scalar or general EF matrix for state-space filters have been explored [10]. A jointly-optimized iterative algorithm with a general, diagonal, or scalar EF matrix using a quasi-Newton method has been developed for state-space digital filters [11]. However, to date the use of *high-order* EF and its effect on noise-reduction performance for state-space digital filters have not been explored.

In this paper, the problem of jointly optimizing *high-order* EF and state-space realization for minimizing the roundoff noise subject to l_2-scaling constraints is investigated for state-space digital filters. The novelty of the proposed method has largely to do with the introduction of a *dynamic* EF that, as will be demonstrated later in the paper, is responsible for a considerable performance enhancement in noise gain reduction relative to the systems that employ static (i.e. constant) EF. We convert the constrained optimization problem encountered into an unconstrained optimization problem by using linear-algebraic techniques. An efficient quasi-Newton algorithm is utilized to solve the unconstrained optimization problem at hand. The proposed technique is applied to the case where the *high-order* EF has diagonal matrices. Case studies are presented to illustrate the validity and effectiveness of the proposed algorithm and demonstrate its performance.

II. PROBLEM STATEMENT

Consider a stable, nth-order, 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, a finite-word-length implementation of (1) with error feedforward and *high-order* EF can be obtained as

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

where $\boldsymbol{h}$ and $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ are referred to as a $1 \times n$ error-feedforward vector and $n \times n$ *high-order* EF 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 each output $\tilde{y}(k)$ has 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 Gaussian random process with zero mean and covariance $\sigma^2\boldsymbol{I}_n$. Subtracting (2) from (1) yields

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

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

978-1-4799-3432-4/14/$31.00 ©2014 IEEE

$\Delta \boldsymbol{x}(0) = \boldsymbol{0}$, we have $\Delta Y(z) = \boldsymbol{H}_e(z)\boldsymbol{E}(z)$ with

$$\boldsymbol{H}_e(z) = \boldsymbol{c}(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\Big(\boldsymbol{A} - \sum_{i=0}^{N-1} \boldsymbol{D}_i z^{-i}\Big) + \boldsymbol{c} - \boldsymbol{h} \quad (4)$$

where $\Delta Y(z)$ and $\boldsymbol{E}(z)$ represent the z-transforms of $\Delta y(k)$ and $\boldsymbol{e}(k)$, respectively.

Based on the model developed above, we now define the normalized noise gain $J_{e1}(\boldsymbol{h}, \boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$ in terms of $\boldsymbol{H}_e(z)$ as

$$J_{e1}(\boldsymbol{h}, \boldsymbol{D}) = \text{tr}\Big[\frac{1}{2\pi j}\oint_{|z|=1} \boldsymbol{H}_e^*(z)\boldsymbol{H}_e(z)\frac{dz}{z}\Big]. \quad (5)$$

In order to derive an easy-to-use formula to evaluate and optimize the noise gain, we write $\boldsymbol{H}_e(z)$ in (4) as

$$\boldsymbol{H}_e(z) = \sum_{k=1}^{\infty} \boldsymbol{c}\Big(\boldsymbol{A}^k - \sum_{i=0}^{N-1} \boldsymbol{A}^{k-1-i}\boldsymbol{D}_i\Big)z^{-k} + \boldsymbol{c} - \boldsymbol{h} \quad (6)$$

where $\boldsymbol{A}^i = \boldsymbol{0}$ for $i < 0$. By substituting (6) into (5) and making use of Cauchy's integral theorem, we obtain

$$\begin{aligned} &J_{e1}(\boldsymbol{h}, \boldsymbol{D}) \\ &= \text{tr}\Big[\boldsymbol{W}_o - \sum_{i=0}^{N-1}\Big\{(\boldsymbol{A}^T)^{i+1}\boldsymbol{W}_o\boldsymbol{D}_i + \boldsymbol{D}_i^T\boldsymbol{W}_o\boldsymbol{A}^{i+1}\Big\} \\ &\quad + \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \boldsymbol{D}_i^T\Big\{(\boldsymbol{A}^T)^{j-i}\boldsymbol{W}_o + \boldsymbol{W}_o\boldsymbol{A}^{i-j}\Big\}\boldsymbol{D}_j \\ &\quad - \sum_{i=0}^{N-1} \boldsymbol{D}_i^T\boldsymbol{W}_o\boldsymbol{D}_i - 2\boldsymbol{h}^T\boldsymbol{c} + \boldsymbol{h}^T\boldsymbol{h}\Big] \end{aligned} \quad (7)$$

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

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

It is useful to note that if the high-order EF matrices $\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ are diagonal, then the formula for the noise gain can be considerably simplified to

$$\begin{aligned} J_{e1}(\boldsymbol{h}, \boldsymbol{D}) &= \text{tr}\Big[\boldsymbol{W}_o - 2\sum_{i=0}^{N-1} \boldsymbol{W}_o\boldsymbol{A}^{i+1}\boldsymbol{D}_i \\ &\quad + \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \boldsymbol{W}_o\boldsymbol{A}^{|i-j|}\boldsymbol{D}_i\boldsymbol{D}_j - 2\boldsymbol{h}^T\boldsymbol{c} + \boldsymbol{h}^T\boldsymbol{h}\Big]. \end{aligned} \quad (9)$$

It should also 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

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

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

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

With an equivalent state-space realization as specified in (11) and assuming the use of high-order diagonal EF matrices, the normalized noise gain is found to be

$$\begin{aligned} J_{e2}(\boldsymbol{h}, \boldsymbol{D}, \boldsymbol{T}) &= \text{tr}\Big[\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T} - 2\sum_{i=0}^{N-1} \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{A}^{i+1}\boldsymbol{T}\boldsymbol{D}_i \\ &\quad + \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{A}^{|i-j|}\boldsymbol{T}\boldsymbol{D}_i\boldsymbol{D}_j - 2\boldsymbol{h}^T\boldsymbol{c}\boldsymbol{T} + \boldsymbol{h}^T\boldsymbol{h}\Big]. \end{aligned} \quad (14)$$

Formula (14) provides an analytic foundation for the minimization of the noise gain by *jointly* optimizing the error feedforward and EF loops as well as state-space coordinate transformation. Formally, the problem being considered here can be stated as to jointly design the error feedforward and high-order EF matrices $\boldsymbol{h}, \boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ and coordinate transformation matrix $\boldsymbol{T}$ to minimize the noise gain $J_{e2}(\boldsymbol{h}, \boldsymbol{D}, \boldsymbol{T})$ in (14) subject to the l_2-scaling constraints in (13).

III. JOINT OPTIMIZATION

In what follows, it is assumed that the high-order EF matrices are all diagonal so that (14) is a valid objective function to be minimized. Because the constraints in (13) are nonconvex and highly nonlinear with respect to matrix $\boldsymbol{T}$, it is beneficial if these constraints can be eliminated so as to work with an unconstrained problem that admits fast Newton-like algorithms. To this end, we define

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

which leads (13) to

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

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

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

Substituting (15) into (14), we obtain

$$\begin{aligned} J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}}) &= \text{tr}\Big[\boldsymbol{V}_0 - 2\sum_{p=0}^{N-1} \boldsymbol{V}_{p+1}\boldsymbol{D}_p \\ &\quad + \sum_{p=0}^{N-1}\sum_{q=0}^{N-1} \boldsymbol{V}_{|p-q|}\boldsymbol{D}_p\boldsymbol{D}_q - 2\boldsymbol{h}^T\hat{\boldsymbol{c}} + \boldsymbol{h}^T\boldsymbol{h}\Big] \end{aligned} \quad (18)$$

where

$$\boldsymbol{V}_p = \hat{\boldsymbol{T}}\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}_o\boldsymbol{A}^p\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T.$$

Following the foregoing arguments, the problem of obtaining matrices $\boldsymbol{T}$ and $\boldsymbol{h}, \boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ 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 $\boldsymbol{h}, \boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$ that jointly minimize $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ 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 $\boldsymbol{h}, \boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}$. Then $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ 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 the assignment $\hat{\boldsymbol{T}} = \boldsymbol{D}_0 = \boldsymbol{D}_1 = \cdots = \boldsymbol{D}_{N-1} = \boldsymbol{I}_n$ and $\boldsymbol{h} = (1, 1, \cdots, 1)$. In the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [12]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{19}$$

where

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

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

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

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

The implementation efficiency and solution accuracy of the quasi-Newton algorithm greatly depends on how the gradient $\nabla J(\boldsymbol{x})$ is evaluated. With high-order diagonal EF matrices, we have managed to derive closed-form formulas for computing the partial derivatives of $J(\boldsymbol{x})$ with respect to the elements of $\boldsymbol{T}$ as well as those of $\boldsymbol{h}$ and $\boldsymbol{D} = [\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_{N-1}]$. The derivation process is quite involved, below we include the formulas without derivation details.

The gradients of $J(\boldsymbol{x})$ with respect to $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n$ are found to be

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} &= 2\boldsymbol{e}_j^T \Big[\boldsymbol{V}_0 - \sum_{p=0}^{N-1} (\boldsymbol{V}_{p+1} + \boldsymbol{V}_{p+1}^T) \boldsymbol{D}_p \\
&+ \frac{1}{2} \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} (\boldsymbol{V}_{|p-q|} + \boldsymbol{V}_{|p-q|}^T) \boldsymbol{D}_p \boldsymbol{D}_q - \hat{\boldsymbol{c}}^T \boldsymbol{h} \Big] \hat{\boldsymbol{T}} \boldsymbol{g}_{ij} \\
& \qquad i, j = 1, 2, \cdots, n
\end{aligned} \tag{21}$$

where t_{ij} is the ith element of column vector $\boldsymbol{t}_j$ and

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

The gradient of $J(\boldsymbol{x})$ with respect to the vector $\boldsymbol{h}$ is described by

$$\frac{\partial J(\boldsymbol{x})}{\partial \boldsymbol{h}} = 2(\boldsymbol{h} - \hat{\boldsymbol{c}}). \tag{22}$$

Let the high-order EF matrices assume the form

$$\boldsymbol{D}_p = \mathrm{diag}\{d_{p1}, d_{p2}, \cdots, d_{pn}\} \quad \text{for} \quad p = 0, 1, \cdots, N-1. \tag{23}$$

The gradients of $J(\boldsymbol{x})$ with respect to the diagonal $\boldsymbol{D}_p$ are given by

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial d_{pi}} &= -2 (\boldsymbol{V}_{p+1})_{ii} + 2 \sum_{q=0}^{N-1} d_{qi} (\boldsymbol{V}_{|p-q|})_{ii} \\
p &= 0, 1, \cdots, N-1 \quad \text{and} \quad i = 1, 2, \cdots, n.
\end{aligned} \tag{24}$$

IV. AN ILLUSTRATIVE EXAMPLE

Consider a stable 3rd-order low-pass state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_3$ with $d = 6.59592 \times 10^{-2}$ and

$$\begin{aligned}
\boldsymbol{A} &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.339377 & -1.152652 & 1.520167 \end{bmatrix} \\
\boldsymbol{b} &= \begin{bmatrix} 0 & 0 & 0 & 0.437881 \end{bmatrix}^T \\
\boldsymbol{c} &= \begin{bmatrix} 0.212964 & 0.293733 & 0.718718 \end{bmatrix}.
\end{aligned}$$

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

$$\begin{aligned}
\boldsymbol{K}_c &= \begin{bmatrix} 1.000000 & 0.741988 & 0.227107 \\ 0.741988 & 1.000000 & 0.741988 \\ 0.227107 & 0.741988 & 1.000000 \end{bmatrix} \\
\boldsymbol{W}_o &= \begin{bmatrix} 0.720426 & -1.635144 & 1.753511 \\ -1.635144 & 4.551539 & -4.194133 \\ 1.753511 & -4.194133 & 5.861185 \end{bmatrix}.
\end{aligned}$$

The noise gain of the filter with no error feedforward and no EF was then computed from (9) as $J_{e1}(\boldsymbol{0}, \boldsymbol{0}) = 11.133150$.

Below we examine two cases. In the first case a static EF (i.e. $N = 1$) was employed with a diagonal $\boldsymbol{D}_0$. The quasi-Newton algorithm was applied to minimize (18) with tolerance $\varepsilon = 10^{-8}$ in (20). It took the algorithm 68 iterations to converge to the solution

$$\begin{aligned}
\boldsymbol{h} &= \begin{bmatrix} 2.605427 & 3.528156 & 1.763048 \end{bmatrix} \\
\boldsymbol{D}_0 &= \mathrm{diag}\{ \ 0.390395 \quad 0.574988 \quad 0.659620 \ \} \\
\hat{\boldsymbol{T}} &= \begin{bmatrix} 5.899352 & -0.516405 & 1.228902 \\ 5.935705 & -0.493044 & 2.377935 \\ 4.442176 & -1.380506 & 1.478864 \end{bmatrix}
\end{aligned}$$

with the noise gain $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}}) = 0.273955$. The profile of $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 68 iterations of the algorithm is depicted in Fig. 1

In the second case a dynamic EF with $N = 2$ and diagonal $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ was utilized. The quasi-Newton algorithm was applied to minimize (18) with tolerance $\varepsilon = 10^{-8}$ in (20). It took the algorithm 38 iterations to converge to the solution

$$\begin{aligned}
\boldsymbol{h} &= \begin{bmatrix} 0.749742 & 0.611596 & 1.006179 \end{bmatrix} \\
\boldsymbol{D}_0 &= \mathrm{diag}\{ \ 0.478050 \quad 0.923420 \quad 1.187433 \ \} \\
\boldsymbol{D}_1 &= \mathrm{diag}\{ \ -0.060355 \quad -0.622458 \quad -0.584191 \ \} \\
\hat{\boldsymbol{T}} &= \begin{bmatrix} 1.367901 & -0.014680 & 0.386079 \\ -0.849834 & 1.032236 & 0.308017 \\ 0.078719 & 0.047320 & 1.218349 \end{bmatrix}
\end{aligned}$$

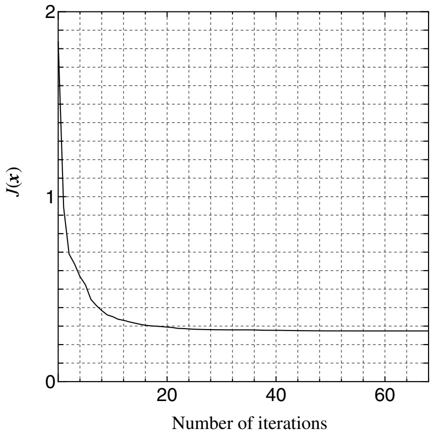

Fig. 1. Profile of $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 68 iterations.

with the noise gain $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}}) = 0.031801$. On comparing with the performance achieved by a static EF as reported in case 1, we observe a much reduced noise gain due to the use of a dynamic EF. The profile of $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 38 iterations of the algorithm is depicted in Fig. 2.

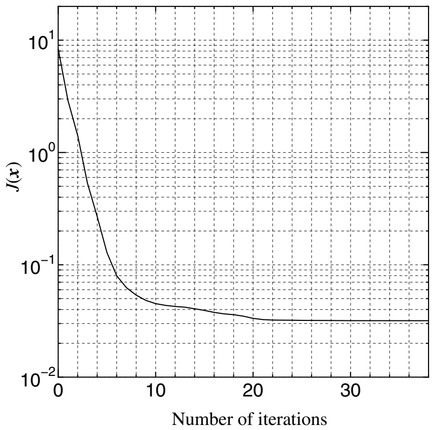

Fig. 2. Profile of $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ during the first 38 iterations.

The other experimental results regarding the noise gain $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ in (18) were summarized in Table I where "Infinite Precision" shows the value of $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ derived from the optimal $\boldsymbol{h}$, $\boldsymbol{D}$ and $\hat{\boldsymbol{T}}$, and "3-Bit Quantization" means that of $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ where only each entry of the optimal $\boldsymbol{h}$ and $\boldsymbol{D}$ was rounded to a power-of-two representation with 3 bits after the binary point, matrix $\hat{\boldsymbol{T}}$ was updated so as to minimize $J_{e3}(\boldsymbol{h}, \boldsymbol{D}, \hat{\boldsymbol{T}})$ with such (rounded) $\boldsymbol{h}$ and $\boldsymbol{D}$ fixed. The separate optimization technique finds an optimal coordinate transformation $\boldsymbol{T}$ first in absence of error feedforward and EF. With $\boldsymbol{T}$ fixed, it then finds the optimal feedforward vector $\boldsymbol{h}$ as well as EF matrix $\boldsymbol{D}$ [10].

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.

TABLE I
PERFORMANCE COMPARISON

N	Optimization	Infinite Precision	3-Bit Quantization
1	Separate	0.644452	0.659679
	Joint	0.273955	0.298340
2	Separate	0.353326	0.373479
	Joint	0.031801	0.054049
3	Separate	0.301305	0.331101
	Joint	1.046941×10^{-7}	0.011724

V. Conclusion

The problem of jointly optimizing error feedforward, *high-order* EF and realization to minimize the effects of roundoff noise at the filter output subject to l_2-scaling constraints for state-space digital filters has been investigated. The problem at hand has been converted into an unconstrained optimization problem by using linear algebraic techniques. Then an efficient quasi-Newton algorithm has been employed to solve the unconstrained optimization problem. The simulation results of a numerical example have demonstrated the validity and effectiveness of the proposed technique.

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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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