IEICE TRANS. FUNDAMENTALS, VOL.E97–A, NO.9 SEPTEMBER 2014

PAPER

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

Akimitsu DOI[†a)], ***Member***, **Takao HINAMOTO**[†b)], ***Fellow***, ***and*** **Wu-Sheng LU**[††c)], ***Nonmember***

SUMMARY For two-dimensional IIR digital filters described by the Fornasini-Marchesini second model, the problem of jointly optimizing high-order error feedback and realization to minimize the effects of roundoff noise at the filter output subject to l_2-scaling constraints is investigated. The problem at hand is converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained optimization problem 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 validity and effectiveness of the proposed technique.

key words: *Fornasini-Marchesini's second model, high-order error feedback and realization, joint optimization, roundoff noise minimization, 2-D filters*

1. Introduction

When IIR digital filters are implemented in fixed-point arithmetic, reduction of roundoff noise effects at the filter output becomes critically important. Error feedback (EF) is an effective tool to reduce finite-word-length (FWL) effects in IIR digital filters, and many EF methods have been proposed in the past [1]–[9]. Another useful approach is to synthesize 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 [10]–[14]. 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 [15]–[18]. Separate and joint optimizations of scalar as well as general EF matrices for state-space filters have been explored in [15]. A quasi-Newton method for joint optimization of general, diagonal, or scalar EF matrix for state-space digital filters is proposed in [16]. The similar technique has been applied to two-dimensional (2-D) state-space digital filters [17]. More recently, the use of high-order EF is found to offer improved noise-reduction performance for 1-D and 2-D state-space digital filters [18]. We stressed that the 2-D IIR filters in the references cited above are all described by the Roesser model [19]. Studies of 2-D IIR filters characterized by the Fornasini-Marchesini second model [20] in terms of stability analysis, low sensitivity realization, and roundoff noise minimization via EF and state-space realization have been reported [21]–[27]. However, studies on the use of high-order EF for the Fornasini-Marchesini second model are not available to date.

This paper investigates the joint optimization problem of high-order EF and realization of 2-D state-space digital filters described by the Fornasini-Marchesini second model [20] for roundoff noise minimization subject to l_2-scaling constraints. The novelty of the proposed method has largely to do with the introduction of high-order EF to 2-D IIR digital filters. The high-order EF, as will be demonstrated later in the paper, is responsible for a considerable performance improvement in noise gain reduction relative to those that employ constant EF. Alongside with this introduction, a state-space model for quantized 2-D discrete-time systems with high-order EF is established, a tractable formula for roundoff noise gain is derived, and our objective–the joint optimization of high-order EF and state-space realization for roundoff noise minimization subject to l_2-scaling constraints is formulated as a constrained optimization problem. The constrained optimization problem at hand is converted into an unconstrained optimization problem by using linear algebraic techniques. An efficient quasi-Newton algorithm [28] is utilized to solve the unconstrained optimization problem at hand. The proposed technique is applied to the case where the high-order EF matrices are diagonal. The present technique can be viewed as an extension of constant EF in [27] to high-order EF. A numerical example is presented to illustrate the proposed algorithm and demonstrate its performance.

Throughout this paper, $\boldsymbol{I}_n$ stands for the identity matrix of dimension $n \times n$, the transpose (conjugate transpose) of a matrix $\boldsymbol{A}$ is indicated by $\boldsymbol{A}^T$ ($\boldsymbol{A}^*$), and the trace and ith diagonal element of a square matrix $\boldsymbol{A}$ are denoted by tr$[\boldsymbol{A}]$ and $(\boldsymbol{A})_{ii}$, respectively.

2. System Model and Problem Description

2.1 System Model

Consider a stable, locally controllable and locally observ-

Manuscript received February 21, 2014.
Manuscript revised April 23, 2014.

[†]The authors are with Hiroshima Institute of Technology, Hiroshima-shi, 731-5193 Japan.
[††]The author is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, V8W 3P6 Canada.
a) E-mail: doi@cc.it-hiroshima.ac.jp
b) E-mail: hinamoto@ieee.org
c) E-mail: wslu@ece.uvic.ca

DOI: 10.1587/transfun.E97.A.1918

Copyright © 2014 The Institute of Electronics, Information and Communication Engineers

able 2-D digital filter $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_n$ described by the Fornasini-Marchesini second model [20]

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

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 constant matrices of appropriate dimensions.

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

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

where

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

$\boldsymbol{h}$ is a $1\times n$ error-feedforward vector, and $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{2k}$ for $k = 1, 2, \cdots, N$ are referred to as $n \times n$ *high-order* EF matrices.

The coefficient matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}$ and d in (2) are assumed to have exact fractional B_c-bit representations. The FWL local state vector $\tilde{\boldsymbol{x}}(i,j)$ and the output $\tilde{y}(i,j)$ all have B-bit fractional representations, 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 $(B-B_c)$-bit after the multiplications and additions, where the sign bit is not counted. It is assumed that the roundoff error $\boldsymbol{e}(i,j)$ can be modeled as a zero-mean noise process with covariance $\sigma^2\boldsymbol{I}_n$. A block diagram of the Fornasini-Marchesini second model with first-order EF and error feedforward path is illustrated in Fig. 1.

By subtracting (2) from (1), we obtain

$$\begin{aligned} \Delta\boldsymbol{x}(i,j) &= \boldsymbol{A}_1\Delta\boldsymbol{x}(i-1,j) + \boldsymbol{A}_2\Delta\boldsymbol{x}(i,j-1) \\ &\quad +\boldsymbol{A}_1\boldsymbol{e}(i-1,j) + \boldsymbol{A}_2\boldsymbol{e}(i,j-1) \\ &\quad -\sum_{k=1}^{N}\{\boldsymbol{D}_{1k}\boldsymbol{e}(i-k,j) + \boldsymbol{D}_{2k}\boldsymbol{e}(i,j-k)\} \\ \Delta y(i,j) &= \boldsymbol{c}\Delta\boldsymbol{x}(i,j) + (\boldsymbol{c}-\boldsymbol{h})\boldsymbol{e}(i,j) \end{aligned} \tag{3}$$

where

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

By taking the (z_1,z_2)-transform on both sides of (3) and setting the boundary conditions $\Delta\boldsymbol{x}(i,0) = \Delta\boldsymbol{x}(0,j) = \boldsymbol{0}$ for $i, j = 1, 2, \cdots$, we have

$$\begin{aligned} \Delta Y(z_1,z_2) &= \boldsymbol{H}_e(z_1,z_2)\boldsymbol{E}(z_1,z_2) \\ \boldsymbol{H}_e(z_1,z_2) &= \boldsymbol{c}(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2)^{-1} \\ &\quad \cdot\left(\boldsymbol{I}_n - \sum_{k=1}^{N}\{z_1^{-k}\boldsymbol{D}_{1k} + z_2^{-k}\boldsymbol{D}_{2k}\}\right) - \boldsymbol{h} \end{aligned} \tag{4}$$

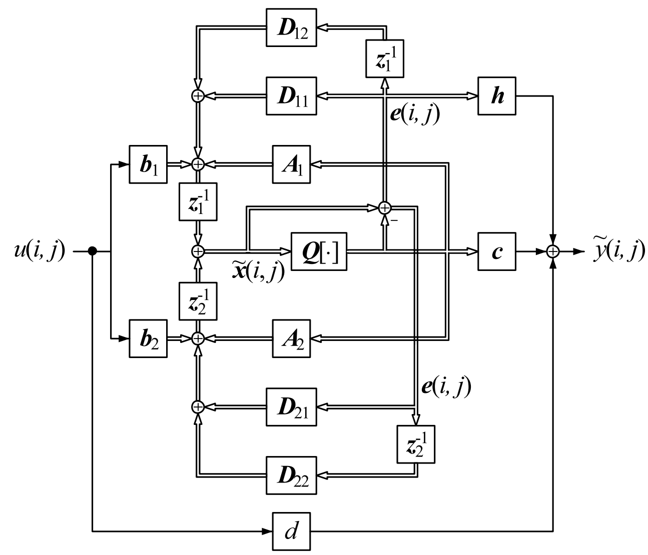

Fig. 1 Block diagram of Fornasini-Marchesini's second model with first-order EF and error feedforward.

where $\Delta Y(z_1,z_2)$ and $\boldsymbol{E}(z_1,z_2)$ are the (z_1,z_2)-transforms of $\Delta y(i,j)$ and $\boldsymbol{e}(i,j)$, respectively. Defining the transition matrix $\boldsymbol{A}^{(i,j)}$, the noise transfer function $\boldsymbol{H}_e(z_1,z_2)$ can be written as

$$\begin{aligned} \boldsymbol{H}_e(z_1,z_2) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{c}\boldsymbol{A}^{(i,j)}z_1^{-i}z_2^{-j} \\ &\quad \cdot\left(\boldsymbol{I}_n - \sum_{k=1}^{N}\{z_1^{-k}\boldsymbol{D}_{1k} + z_2^{-k}\boldsymbol{D}_{2k}\}\right) - \boldsymbol{h} \end{aligned} \tag{5}$$

where

$$\begin{aligned} (\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2)^{-1} &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{A}^{(i,j)}z_1^{-i}z_2^{-j} \\ \boldsymbol{A}^{(i,j)} &= \boldsymbol{A}_1\boldsymbol{A}^{(i-1,j)} + \boldsymbol{A}_2\boldsymbol{A}^{(i,j-1)} \\ &= \boldsymbol{A}^{(i-1,j)}\boldsymbol{A}_1 + \boldsymbol{A}^{(i,j-1)}\boldsymbol{A}_2 \\ \boldsymbol{A}^{(0,0)} = \boldsymbol{I}_n, \qquad \boldsymbol{A}^{(i,j)} &= \boldsymbol{0} \ \text{ for } \ i<0 \ \text{ or } \ j<0. \end{aligned}$$

2.2 Problem Description

We define the normalized noise gain $J_{e1}(\boldsymbol{h},\boldsymbol{D}) = \sigma_{out}^2/\sigma^2$ with $\boldsymbol{D} = [\boldsymbol{D}_{11}, \boldsymbol{D}_{12}, \cdots, \boldsymbol{D}_{1N}, \boldsymbol{D}_{21}, \cdots, \boldsymbol{D}_{2N}]$ as

$$\begin{aligned} J_{e1}(\boldsymbol{h},\boldsymbol{D}) = \mathrm{tr}\Big[&\frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1} \\ &\boldsymbol{H}_e^*(z_1,z_2)\boldsymbol{H}_e(z_1,z_2)\frac{dz_1}{z_1}\frac{dz_2}{z_2}\Big] \end{aligned} \tag{6}$$

By substituting (5) into (6) and making use of 2-D Cauchy's integral theorem, we obtain

$$\begin{aligned} &J_{e1}(\boldsymbol{h},\boldsymbol{D}) \\ &= \mathrm{tr}\Big[\boldsymbol{W}_{00} - 2\sum_{k=1}^{N}\Big\{\boldsymbol{W}_{k0}\boldsymbol{D}_{1k} + \boldsymbol{W}_{0k}\boldsymbol{D}_{2k}\Big\} \\ &\quad + \sum_{k=1}^{N}\sum_{l=1}^{N}\Big\{\boldsymbol{W}_{k-l,0}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}^T + \boldsymbol{W}_{0,k-l}\boldsymbol{D}_{2k}\boldsymbol{D}_{2l}^T \\ &\qquad + \boldsymbol{V}_{kl}\boldsymbol{D}_{1k}\boldsymbol{D}_{2l}^T + \boldsymbol{V}_{lk}\boldsymbol{D}_{1l}\boldsymbol{D}_{2k}^T\Big\} \\ &\qquad - \boldsymbol{c}^T\boldsymbol{h} - \boldsymbol{h}^T\boldsymbol{c} + \boldsymbol{h}^T\boldsymbol{h}\Big] \end{aligned} \tag{7}$$

where

$$\boldsymbol{W}_{kl} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(\boldsymbol{c}\boldsymbol{A}^{(i+k,j+l)})^T\boldsymbol{c}\boldsymbol{A}^{(i,j)}$$
$$\boldsymbol{V}_{kl} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(\boldsymbol{c}\boldsymbol{A}^{(i+k,j)})^T\boldsymbol{c}\boldsymbol{A}^{(i,j+l)}.$$

The above matrices correspond to generalizations of the local observability Gramian $\boldsymbol{W}_o$ of the filter in (1) that is defined by

$$\begin{aligned} \boldsymbol{W}_o &= \frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{G}^*(z_1,z_2)\boldsymbol{G}(z_1,z_2)\frac{dz_1}{z_1}\frac{dz_2}{z_2} \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(\boldsymbol{c}\boldsymbol{A}^{(i,j)})^T\boldsymbol{c}\boldsymbol{A}^{(i,j)} \end{aligned} \tag{8}$$

where

$$\boldsymbol{G}(z_1,z_2) = \boldsymbol{c}(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2)^{-1}.$$

It should be noted that l_2-scaling constraints on the local state vector $\boldsymbol{x}(i,j)$ involve the local controllability Gramian $\boldsymbol{K}_c$ of the filter in (1) which can be defined by

$$\begin{aligned} \boldsymbol{K}_c &= \frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{F}(z_1,z_2)\boldsymbol{F}^*(z_1,z_2)\frac{dz_1}{z_1}\frac{dz_2}{z_2} \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\boldsymbol{f}(i,j)\boldsymbol{f}^T(i,j) \end{aligned} \tag{9}$$

where

$$\boldsymbol{F}(z_1,z_2) = (\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2)^{-1}(z_1^{-1}\boldsymbol{b}_1 + z_2^{-1}\boldsymbol{b}_2)$$
$$\boldsymbol{f}(i,j) = \boldsymbol{A}^{(i-1,j)}\boldsymbol{b}_1 + \boldsymbol{A}^{(i,j-1)}\boldsymbol{b}_2.$$

A different yet equivalent local state-space description of (1), $(\overline{\boldsymbol{A}}_1,\overline{\boldsymbol{A}}_2,\overline{\boldsymbol{b}}_1,\overline{\boldsymbol{b}}_2,\overline{\boldsymbol{c}},d)_n$, can be obtained via a coordinate transformation $\overline{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j)$ where

$$\begin{aligned} &\overline{\boldsymbol{A}}_1 = \boldsymbol{T}^{-1}\boldsymbol{A}_1\boldsymbol{T}, \quad \overline{\boldsymbol{A}}_2 = \boldsymbol{T}^{-1}\boldsymbol{A}_2\boldsymbol{T} \\ &\overline{\boldsymbol{b}}_1 = \boldsymbol{T}^{-1}\boldsymbol{b}_1, \quad \overline{\boldsymbol{b}}_2 = \boldsymbol{T}^{-1}\boldsymbol{b}_2, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \end{aligned} \tag{10}$$

The l_2-scaling constraints are imposed on the local state vector $\overline{\boldsymbol{x}}(i,j)$ so that

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1, \ i = 1,2,\cdots,n. \tag{11}$$

The problem being considered here is to design an optimal coordinate transformation matrix $\boldsymbol{T}$ as well as *high-order* EF diagonal matrices $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{2k}$ for $k = 1,2,\cdots,N$ that jointly minimize the noise gain

$$\begin{aligned} &J_{e2}(\boldsymbol{T},\boldsymbol{D}) \\ &= \mathrm{tr}\Big[\overline{\boldsymbol{W}}_{00} - \overline{\boldsymbol{c}}^T\overline{\boldsymbol{c}} - 2\sum_{k=1}^{N}\Big\{\overline{\boldsymbol{W}}_{k0}\boldsymbol{D}_{1k} + \overline{\boldsymbol{W}}_{0k}\boldsymbol{D}_{2k}\Big\} \\ &\quad + \sum_{k=1}^{N}\sum_{l=1}^{N}\Big\{\overline{\boldsymbol{W}}_{|k-l|,0}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l} + \overline{\boldsymbol{W}}_{0,|k-l|}\boldsymbol{D}_{2k}\boldsymbol{D}_{2l} \\ &\qquad + 2\overline{\boldsymbol{V}}_{kl}\boldsymbol{D}_{1k}\boldsymbol{D}_{2l}\Big\}\Big] \end{aligned} \tag{12}$$

subject to l_2-scaling constraints in (11) where the error feed-forward vector $\boldsymbol{h}$ is chosen as $\boldsymbol{h} = \overline{\boldsymbol{c}}$ and

$$\overline{\boldsymbol{W}}_{kl} = \boldsymbol{T}^T\boldsymbol{W}_{kl}\boldsymbol{T}, \qquad \overline{\boldsymbol{V}}_{kl} = \boldsymbol{T}^T\boldsymbol{V}_{kl}\boldsymbol{T}.$$

3. Joint Optimization of High-Order Error Feedback and Realization

To deal with the l_2-scaling constraints in (11), we define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}_c^{-\frac{1}{2}}. \tag{13}$$

Then (11) becomes

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1, \quad i = 1,2,\cdots,n. \tag{14}$$

It is noted that these constraints are always satisfied if $\hat{\boldsymbol{T}}^{-1}$ assumes the form

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\right]. \tag{15}$$

Substituting (13) into (12), we obtain

$$\begin{aligned} &J_{e3}(\hat{\boldsymbol{T}},\boldsymbol{D}) \\ &= \mathrm{tr}\Big[\hat{\boldsymbol{W}}_{00} - \hat{\boldsymbol{c}}^T\hat{\boldsymbol{c}} - 2\sum_{k=1}^{N}\Big\{\hat{\boldsymbol{W}}_{k0}\boldsymbol{D}_{1k} + \hat{\boldsymbol{W}}_{0k}\boldsymbol{D}_{2k}\Big\} \\ &\quad + \sum_{k=1}^{N}\sum_{l=1}^{N}\Big\{\hat{\boldsymbol{W}}_{|k-l|,0}\boldsymbol{D}_{1k}\boldsymbol{D}_{1l} + \hat{\boldsymbol{W}}_{0,|k-l|}\boldsymbol{D}_{2k}\boldsymbol{D}_{2l} \\ &\qquad + 2\hat{\boldsymbol{V}}_{kl}\boldsymbol{D}_{1k}\boldsymbol{D}_{2l}\Big\}\Big] \end{aligned} \tag{16}$$

where

$$\hat{\boldsymbol{W}}_{kl} = \hat{\boldsymbol{T}}\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}_{kl}\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T$$
$$\hat{\boldsymbol{V}}_{kl} = \hat{\boldsymbol{T}}\boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{V}_{kl}\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T, \qquad \hat{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T.$$

From the foregoing arguments, the problem of obtaining matrices $\boldsymbol{T}$, $\boldsymbol{D}_{1l}$ and $\boldsymbol{D}_{2l}$ for $l = 1,2,\cdots,N$ that jointly minimize (12) subject to the l_2-scaling constraints in (11) is now converted into an unconstrained optimization problem of obtaining $\hat{\boldsymbol{T}}$, $\boldsymbol{D}_{1l}$ and $\boldsymbol{D}_{2l}$ for $l = 1,2,\cdots,N$ that jointly minimize $J_{e3}(\hat{\boldsymbol{T}},\boldsymbol{D})$ in (16).

Let $\boldsymbol{x}$ be the column vector that collects the variables in matrices $[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n]$, $\boldsymbol{D}_{1l}$ and $\boldsymbol{D}_{2l}$ for $l = 1,2,\cdots,N$. Then $J_{e3}(\hat{\boldsymbol{T}},\boldsymbol{D})$ in (16) 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 $\boldsymbol{T} = \boldsymbol{D}_{1l} = \boldsymbol{D}_{2l} = \boldsymbol{I}_n$

for $l = 1, 2, \cdots, N$. In the kth iteration, a quasi-Newton algorithm, known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [28]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{17}$$

where

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

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

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon \tag{18}$$

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

The BFGS algorithm is a well-known descent algorithm, meaning that a twice continuously differentiable objective function at the iterates generated by the BFGS algorithm is monotonically decreasing. Although the objective function involved in the joint optimization is not convex in general, in theory it was shown [29] that if the starting point is sufficiently close to a local minimizer and the initial Hessian approximation is sufficiently close to the Hessian at that minimizer, then the BFGS iterates will converge to the minimizer. In practice, the above iterative algorithm was applied to quite a number of simulation examples, and fast convergence was observed in each and every instance where the BFGS algorithm converged to an identical solution point for a variety of initial state-space realizations. In the next section, the effectiveness of the proposed algorithm is illustrated using a representative example from our simulation studies.

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 EF diagonal 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 $\boldsymbol{D}_{1l}$ and $\boldsymbol{D}_{2l}$ for $l = 1, 2, \cdots, N$. The derivation process is quite involved, below we include the formulas without derivation details.

In what follows, we define for $k = 1, 2, \ldots, N$

$$\begin{aligned}
\boldsymbol{D}_{1k} &= \mathrm{diag}\{\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{kn}\} \\
\boldsymbol{D}_{2k} &= \mathrm{diag}\{\beta_{k1}, \beta_{k2}, \cdots, \beta_{kn}\}.
\end{aligned} \tag{19}$$

Then the gradient of $J(\boldsymbol{x})$ with respect to $\boldsymbol{T}$ is found to be

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial t_{ij}} = 2\boldsymbol{e}_j^T \Bigg[& \hat{\boldsymbol{W}}_{00} - \hat{\boldsymbol{c}}^T \hat{\boldsymbol{c}} - \sum_{k=1}^{N} \Big\{ (\hat{\boldsymbol{W}}_{k0} + \hat{\boldsymbol{W}}_{k0}^T) \boldsymbol{D}_{1k} \\
& + (\hat{\boldsymbol{W}}_{0k} + \hat{\boldsymbol{W}}_{0k}^T) \boldsymbol{D}_{2k} \Big\} \\
& + \frac{1}{2} \sum_{k=1}^{N} \sum_{l=1}^{N} \Big\{ (\hat{\boldsymbol{W}}_{|k-l|,0} + \hat{\boldsymbol{W}}_{|k-l|,0}^T) \boldsymbol{D}_{1k} \boldsymbol{D}_{1l} \\
& + (\hat{\boldsymbol{W}}_{0,|k-l|} + \hat{\boldsymbol{W}}_{0,|k-l|}^T) \boldsymbol{D}_{2k} \boldsymbol{D}_{2l} \\
& + 2(\hat{\boldsymbol{V}}_{kl} + \hat{\boldsymbol{V}}_{kl}^T) \boldsymbol{D}_{1k} \boldsymbol{D}_{2l} \Big\} \Bigg] \hat{\boldsymbol{T}} \boldsymbol{g}_{ij}
\end{aligned} \tag{20}$$

for $i, j = 1, 2, \cdots, n$ where

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

and the gradients of $J(\boldsymbol{x})$ with respect to the high-order EF matrices $\boldsymbol{D}_{1l}$ and $\boldsymbol{D}_{2l}$ are given by

$$\begin{aligned}
\frac{\partial J(\boldsymbol{x})}{\partial \alpha_{rp}} &= -2\,(\hat{\boldsymbol{W}}_{r0})_{pp} + 2 \sum_{l=1}^{N} \beta_{lp} (\hat{\boldsymbol{V}}_{rl})_{pp} \\
&\quad + 2 \sum_{l=1}^{N} \alpha_{lp} (\hat{\boldsymbol{W}}_{|r-l|,0})_{pp} \\
\frac{\partial J(\boldsymbol{x})}{\partial \beta_{rp}} &= -2\,(\hat{\boldsymbol{W}}_{0r})_{pp} + 2 \sum_{l=1}^{N} \alpha_{lp} (\hat{\boldsymbol{V}}_{lr})_{pp} \\
&\quad + 2 \sum_{l=1}^{N} \beta_{lp} (\hat{\boldsymbol{W}}_{0,|r-l|})_{pp}
\end{aligned} \tag{21}$$

for $r = 1, 2, \cdots, N$ and $p = 1, 2, \cdots, n$.

We remark that using the closed-form formulas given in (20) and (21) allows us to quickly and accurately evaluate gradient $\nabla J(\boldsymbol{x})$, which is a key quantity in updating the iterate via (17), hence ensures high efficiency of the proposed algorithm.

4. A Numerical Example

As a numerical example, consider a 2-D digital filter $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}, d)_4$ described by

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & 0 & 0 & -0.00411 \\ 1 & 0 & 0 & 0.08007 \\ 0 & 1 & 0 & -0.42458 \\ 0 & 0 & 1 & 1.04460 \end{bmatrix}$$

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

$$\boldsymbol{b}_1 = \begin{bmatrix} -0.01452 & 0.01234 & 0.02054 & 0.04762 \end{bmatrix}^T$$

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

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

Applying a coordinate transformation defined by

$$\boldsymbol{T}_o = \mathrm{diag}\{0.093632, 0.215308, 0.480320, 1.026019\}$$

to the above state-space parameters, the local controllability Gramian $\boldsymbol{K}_c$ and the local observability Gramian $\boldsymbol{W}_o$ were computed from (9) and (8) with truncation $(0,0) \le (i,j) \le (100,100)$ as

$$\boldsymbol{K}_c = \begin{bmatrix} 1.000000 & -0.881334 & 0.112452 & -0.294461 \\ -0.881334 & 1.000000 & -0.230322 & 0.275434 \\ 0.112452 & -0.230322 & 1.000000 & -0.920316 \\ -0.294461 & 0.275434 & -0.920316 & 1.000000 \end{bmatrix}$$

$$\boldsymbol{W}_o = 10 \begin{bmatrix} 1.337106 & 1.460803 & 1.585050 & 1.595948 \\ 1.460803 & 1.637337 & 1.820902 & 1.858525 \\ 1.585050 & 1.820902 & 2.122581 & 2.269663 \\ 1.595948 & 1.858525 & 2.269663 & 2.672436 \end{bmatrix}$$

respectively. The noise gain of the filter without error feed-forward and EF was then computed from (7) as

$$J_{e1}(\boldsymbol{0},\boldsymbol{0}) = \mathrm{tr}[\boldsymbol{T}_o^T \boldsymbol{W}_o \boldsymbol{T}_o] = 77.694599.$$

Two scenarios were considered, where a static EF (i.e. $N = 1$) and a dynamic EF with $N = 2$ were employed, respectively. In both cases, the EF matrices involved were diagonal.

In the case of $N = 1$, the quasi-Newton algorithm was applied to minimize (16) with tolerance $\varepsilon = 10^{-8}$ in (18). It took the algorithm 64 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 0.168774 & -0.486172 & 0.915628 & -0.736402 \\ -0.728472 & 0.809253 & -0.456256 & -0.571638 \\ -0.242866 & -0.676020 & 1.309537 & 0.357377 \\ 0.746109 & 0.658517 & 0.261039 & 0.302379 \end{bmatrix}$$

$$\boldsymbol{D}_{11} = \mathrm{diag}\{\, 0.513021 \;\; 0.371958 \;\; 0.167229 \;\; 0.452178 \,\}$$

$$\boldsymbol{D}_{21} = \mathrm{diag}\{\, 0.444120 \;\; 0.668166 \;\; 0.349722 \;\; 0.577416 \,\}$$

The minimized noise gain was found to be

$$J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}) = 0.186190.$$

The profile of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D})$ during the first 64 iterations of the algorithm is dipicted in Fig. 2.

In the second case a dynamic EF with $N = 2$ and diagonal $\boldsymbol{D}_{11}$, $\boldsymbol{D}_{12}$, $\boldsymbol{D}_{21}$ and $\boldsymbol{D}_{22}$ was utilized. The quasi-Newton algorithm was applied to minimize (16) with tolerance $\varepsilon = 10^{-8}$ in (18). It took the algorithm 72 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 0.192608 & 0.369715 & -0.061921 & -1.251360 \\ 0.719608 & 0.588335 & 0.378287 & 0.363545 \\ -0.674075 & 0.305026 & 0.708886 & -0.746125 \\ -0.300545 & 0.641049 & -0.877100 & -0.255396 \end{bmatrix}$$

$$\boldsymbol{D}_{11} = \mathrm{diag}\{\, 0.728153 \;\; 0.522761 \;\; 0.556503 \;\; 0.322581 \,\}$$

$$\boldsymbol{D}_{12} = -\mathrm{diag}\{\, 0.231036 \;\; 0.198448 \;\; 0.250199 \;\; -0.144049 \,\}$$

$$\boldsymbol{D}_{21} = \mathrm{diag}\{\, 0.411635 \;\; 1.107838 \;\; 0.406595 \;\; 0.440976 \,\}$$

$$\boldsymbol{D}_{22} = \mathrm{diag}\{\, -0.001412 \;\; -0.492405 \;\; 0.027757 \;\; 0.019747 \,\}$$

whose noise gain was found to be

$$J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D}) = 0.073287.$$

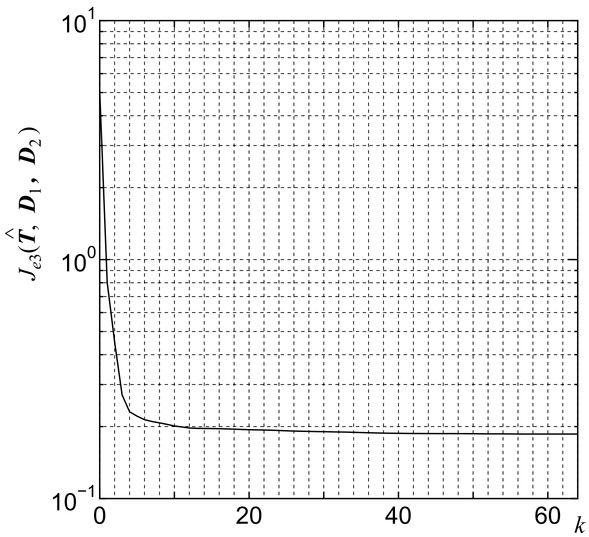

Fig. 2 Profile of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D})$ during the first 64 iterations.

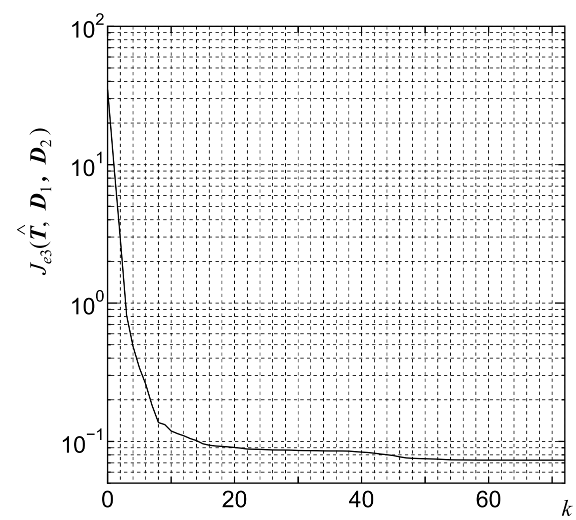

Fig. 3 Profile of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D})$ during the first 72 iterations.

On comparing with the performance achieved by a static EF as reported in the first case, it is observed that the use of a dynamic EF has led to a much reduced noise gain. The profile of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D})$ during the first 72 iterations of the algorithm is dipicted in Fig. 3.

Our experimental study also includes comparisons of the performance in terms of the noise gain $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D})$ in (16) offered by jointly optimized EF (with $N = 1$ and 2) with those obtained by the EF of the same order that were optimized by using the separate optimization method. The main steps involved in the separate optimization method are included in the appendix for the reader's convenience. The numerical results of these eight cases are summarized in Table 1 where the column associated with "Infinite Precision" shows the values of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D})$ derived from the optimal $\hat{\boldsymbol{T}}$ and $\boldsymbol{D}$, and the columns with "3-Bit Quantization" are the values of $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D})$ where each entry of the optimal

Table 1 Performance comparison.

N	Optimization	Infinite Precision	3-Bit Quantization
1	Separate	0.454620	0.461004
	Joint	0.186190	0.221813
2	Separate	0.313562	0.352139
	Joint	0.073287	0.093987

$\boldsymbol{D}$ was rounded to a power-of-two representation with 3 bits after the binary point, and then matrix $\hat{\boldsymbol{T}}$ was updated so as to minimize $J_{e3}(\hat{\boldsymbol{T}}, \boldsymbol{D})$ with $\boldsymbol{D}$ fixed. The results reported above have clearly demonstrated that the use of high-order EF, when jointly optimized with state-space realization, can improve the performance of roundoff noise reduction substantially.

5. Conclusion

The joint optimization problem of *high-order* EF and realization to minimize the effects of roundoff noise subject to l_2-scaling constraints for 2-D digital filters described by the Fornasini-Marchesini second model has been investigated. Linear algebraic techniques have been employed to convert the problem at hand into an unconstrained optimization problem. The resultant unconstrained optimization problem has been solved iteratively by applying an efficient quasi-Newton algorithm. Moreover, closed-form formulas for fast evaluation of the gradient of the objective function have been derived. The proposed technique has been applied to the case where *high-order* EF matrices are diagonal. A numerical example has demonstrated the effectiveness of the proposed technique.

References

[1] T. Thong and B. Liu, "Error spectrum shaping in narrowband recursive digital filters," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-25, no.2, pp.200–203, April 1977.

[2] T.L. Chang and S.A. White, "An error cancellation digital filter structure and its distributed-arithmetic implementation," IEEE Trans. Circuits Syst., vol.CAS-28, no.4, pp.339–342, April 1981.

[3] D.C. Munson and D. Liu, "Narrowband recursive filters with error spectrum shaping," IEEE Trans. Circuits Syst., vol.CAS-28, no.2, pp.160–163, Feb. 1981.

[4] 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 Process., vol.ASSP-30, no.6, pp.963–973, Dec. 1982.

[5] M. Renfors, "Roundoff noise in error-feedback state-space filters," Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'83), pp.619–622, April 1983.

[6] W.E. Higgins and D.C. Munson, "Optimal and suboptimal error-spectrum shaping for cascade-form digital filters," IEEE Trans. Circuits Syst., vol.CAS-31, no.5, pp.429–437, May 1984.

[7] T.I. Laakso and I.O. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," IEEE Trans. Signal Process., vol.40, no.5, pp.1096–1107, May 1992.

[8] P.P. Vaidyanathan, "On error-spectrum shaping in state-space digital filters," IEEE Trans. Circuits Syst., vol.CAS-32, no.1, pp.88–92, Jan. 1985.

[9] D. Williamson, "Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using residue feedback," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-34, no.5, pp.1210–1220, Oct. 1986.

[10] C.T. Mullis and R.A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," IEEE Trans. Circuits Syst., vol.CAS-23, no.9, pp.551–562, Sept. 1976.

[11] S.Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-25, no.4, pp.273–281, Aug. 1977.

[12] 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.CAS-26, no.3, pp.149–153, March 1979.

[13] M. Kawamata and T. Higuchi, "Synthesis of 2-D separable-denominator digital filters with minimum roundoff noise and no overflow oscillations," IEEE Trans. Circuits Syst., vol.CAS-33, no.4, pp.365–372, April 1986.

[14] 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.CAS-33, no.10, pp.965–973, Oct. 1986.

[15] 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, Fundam. Theory Appl., vol.50, no.1, pp.23–33, Jan. 2003.

[16] W.-S. Lu and T. Hinamoto, "Jointly optimized error-feedback and realization for roundoff noise minimization in state-space digital filters," IEEE Trans. Signal Process., vol.53, no.6, pp.2135–2145, June 2005.

[17] T. Hinamoto, H. Ohnishi, and W.-S. Lu, "Roundoff noise minimization for 2-D state-space digital filters using joint optimization of error-feedback and realization," IEEE Trans. Signal Process., vol.54, no.11, pp.4302–4310, Nov. 2006.

[18] T. Hinamoto, A. Doi, and W.-S. Lu, "Jointly optimal high-order error feedback and realization for roundoff noise minimization in 1-D and 2-D state-space digital filters," IEEE Trans. Signal Process., vol.61, no.23, pp.5893–5904, Dec. 2013.

[19] R.P. Roesser, "A discrete state-space model for linear image processing," IEEE Trans. Autom. Control, vol.AC-20, no.1, pp.1–10, Feb. 1975.

[20] E. Fornasini and G. Marchesini, "Doubly-indexed dynamical systems: State-space models and structural properties," Mathematical Systems Theory, vol.12, pp.59–72, 1978.

[21] E. Fornasini and G. Marchesini, "Stability analysis of 2-D systems," IEEE Trans. Circuits Syst., vol.CAS-27, no.12, pp.1210–1217, Dec. 1980.

[22] M. Bisiacco, "State and output feedback stability of 2-D systems," IEEE Trans. Circuits Syst., vol.CAS-32, no.12, pp.1246–1254, Dec. 1985.

[23] T. Hinamoto, T. Hamanaka, and S. Maekawa, "Synthesis of 2-D state-space digital filters with low sensitivity based on the Fornasini-Marchesini model," IEEE Trans. Acoust. Speech Signal Process., vol.28, no.9, pp.1587–1594, Sept. 1990.

[24] T. Hinamoto, "2-D Lyapunov equation and filter design based on the Fornasini-Marchesini second model," IEEE Trans. Circuits Syst. I, Fundam. Theory Appl., vol.40, no.2, pp.102–110, Feb. 1993.

[25] M.E. Valcher, "On the internal stability and asymptotic behavior of 2-D positive systems," IEEE Trans. Circuits Syst. I, Fundam. Theory Appl., vol.44, no.7, pp.602–613, July 1997.

[26] G. Li, "On frequency weighted minimal L_2 sensitivity of 2-D systems using Fornasini-Marchesini LSS model," IEEE Trans. Circuits Syst. I, Fundam. Theory Appl., vol.44, no.7, pp.642–646, July 1997.

[27] T. Hinamoto, T. Oumi, and W.-S. Lu, "Separate/joint optimization

of error feedback and realization for roundoff noise minimization in a class of 2-D state-space digital filters," Multidimens. Syst., Signal Process., vol.18, no.4, pp.327–339, Dec. 2007.
[28] R. Fletcher, Practical Methods of Optimization, 2nd ed., Wiley, New York, 1987.
[29] J.E. Dennis and J.J. More, "Quasi-Newton methods, motivation and theory," SIAM Review, vol.19, pp.46–89, 1977.

Appendix: Analytic Method for Separate Optimization

The separate minimization is carried out in two major steps. First, we fix the high-order EF matrices to $\boldsymbol{D}_{1k} = \boldsymbol{D}_{2k} = \boldsymbol{0}$ for $k = 1, 2, \cdots, N$ in (12) so that the objective function is reduced to

$$J_{e2}(\boldsymbol{T}, \boldsymbol{0}) = \mathrm{tr}[\overline{\boldsymbol{W}}_o - \overline{\boldsymbol{c}}^T \overline{\boldsymbol{c}}] \tag{A·1}$$

which is minimized with respect to matrix $\boldsymbol{T}$ subject to the l_2-scaling constraints in (11). Second, with $\boldsymbol{T}$ optimized in the first step, (12) is minimized under the fixed $\boldsymbol{T}$ with respect to matrices $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{2k}$ for $k = 1, 2, \cdots, N$. To perform the first step, we define the Lagrange function

$$J_o(\boldsymbol{P}, \lambda) = \mathrm{tr}[\boldsymbol{V}\boldsymbol{P}] - \lambda\,(\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n) \tag{A·2}$$

where $\boldsymbol{V} = \boldsymbol{W}_o - \boldsymbol{c}^T\boldsymbol{c}$ and $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$. The optimal coordinate transformation matrix $\boldsymbol{T}$ can be determined by solving the equations

$$\begin{aligned} \frac{\partial J_o(\boldsymbol{P}, \lambda)}{\partial \boldsymbol{P}} &= \boldsymbol{V} - \lambda \boldsymbol{P}^{-1}\boldsymbol{K}_c\boldsymbol{P}^{-1} = \boldsymbol{0} \\ \frac{\partial J_o(\boldsymbol{P}, \lambda)}{\partial \lambda} &= \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n = 0 \end{aligned} \tag{A·3}$$

which are led to [15]

$$\boldsymbol{T} = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\theta_i\right)^{\frac{1}{2}} \mathbf{V}^{-\frac{1}{2}}\left[\mathbf{V}^{\frac{1}{2}}\boldsymbol{K}_\mathbf{c}\boldsymbol{V}^{\frac{1}{2}}\right]^{\frac{1}{4}}\boldsymbol{U} \tag{A·4}$$

where $\theta_1^2, \theta_2^2, \cdots, \theta_n^2$ are the eigenvalues of $\boldsymbol{K}_c\boldsymbol{V}$ and $\boldsymbol{U}$ is an arbitrary $n \times n$ orthogonal matrix.

Suppose that the eigenvalue-eigenvector decomposition of $\left[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}\right]^{\frac{1}{2}}$ can be written as

$$\left[\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{V}^{\frac{1}{2}}\right]^{\frac{1}{2}} = \boldsymbol{R}\,\mathrm{diag}\{\theta_1, \theta_2, \cdots, \theta_n\}\boldsymbol{R}^T \tag{A·5}$$

where $\boldsymbol{R}$ is an $n \times n$ orthogonal matrix. By numerical manupilations [11, p.278], it is possible to obtain an $n \times n$ orthogonal matrix $\boldsymbol{S}$ such that

$$\boldsymbol{S}\boldsymbol{\Lambda}^{-2}\boldsymbol{S}^T = \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \tag{A·6}$$

where

$$\begin{aligned} \boldsymbol{\Lambda} &= \mathrm{diag}\{\lambda_\mathbf{1}, \lambda_\mathbf{2}, \cdots, \lambda_\mathbf{n}\} \\ \lambda_i &= \left(\frac{\theta_1 + \theta_2 + \cdots + \theta_n}{n\theta_i}\right)^{\frac{1}{2}}, \quad i = 1, 2, \cdots, n. \end{aligned}$$

Then, by choosing $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{S}^T$ in (A·4), the l_2-scaling constraints in (11) are satisfied [15].

With definition of (19), it follows from (12) that

$$\begin{aligned} \frac{\partial J_{e2}(\boldsymbol{T}, \boldsymbol{D})}{\partial \alpha_{rp}} &= -2\left(\overline{\boldsymbol{W}}_{r0}\right)_{pp} + 2\sum_{l=1}^{N}\left(\alpha_{lp}w_{rl} + \beta_{lp}v_{rl}\right) = 0 \\ \frac{\partial J_{e2}(\boldsymbol{T}, \boldsymbol{D})}{\partial \beta_{rp}} &= -2\left(\overline{\boldsymbol{W}}_{0r}\right)_{pp} + 2\sum_{l=1}^{N}\left(\beta_{lp}u_{rl} + \alpha_{lp}v_{lr}\right) = 0 \end{aligned} \tag{A·7}$$

for $r = 1, 2, \cdots, N$ and $p = 1, 2, \cdots, n$ where

$$w_{rlp} = \left(\overline{\boldsymbol{W}}_{|r-l|,0}\right)_{pp}, \qquad v_{rlp} = \left(\overline{\boldsymbol{V}}_{rl}\right)_{pp}$$
$$u_{rlp} = \left(\overline{\boldsymbol{W}}_{0,|r-l|}\right)_{pp}.$$

By virtue of (A·7), we obtain

$$\begin{aligned} \boldsymbol{X}_{1p}\boldsymbol{\alpha}_p + \boldsymbol{Y}_p\boldsymbol{\beta}_p &= \boldsymbol{z}_{1p} \\ \boldsymbol{X}_{2p}\boldsymbol{\beta}_p + \boldsymbol{Y}_p^T\boldsymbol{\alpha}_p &= \boldsymbol{z}_{2p} \end{aligned} \tag{A·8}$$

for $p = 1, 2, \cdots, n$ where

$$\begin{aligned} \boldsymbol{\alpha}_p &= \begin{bmatrix} \alpha_{1p} & \alpha_{2p} & \cdots & \alpha_{Np} \end{bmatrix}^T \\ \boldsymbol{\beta}_p &= \begin{bmatrix} \beta_{1p} & \beta_{2p} & \cdots & \beta_{Np} \end{bmatrix}^T \\ \boldsymbol{X}_{1p} &= \begin{bmatrix} w_{11p} & w_{12p} & \cdots & w_{1Np} \\ w_{21p} & w_{22p} & \cdots & w_{2Np} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1p} & w_{N2p} & \cdots & w_{NNp} \end{bmatrix} \\ \boldsymbol{X}_{2p} &= \begin{bmatrix} u_{11p} & u_{12p} & \cdots & u_{1Np} \\ u_{21p} & u_{22p} & \cdots & u_{2Np} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1p} & u_{N2p} & \cdots & u_{NNp} \end{bmatrix} \\ \boldsymbol{Y}_p &= \begin{bmatrix} v_{11p} & v_{12p} & \cdots & v_{1Np} \\ v_{21p} & v_{22p} & \cdots & v_{2Np} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N1p} & v_{N2p} & \cdots & v_{NNp} \end{bmatrix} \\ \boldsymbol{z}_{1p} &= \begin{bmatrix} (\overline{\boldsymbol{W}}_{10})_{pp} & (\overline{\boldsymbol{W}}_{20})_{pp} & \cdots & (\overline{\boldsymbol{W}}_{N0})_{pp} \end{bmatrix}^T \\ \boldsymbol{z}_{2p} &= \begin{bmatrix} (\overline{\boldsymbol{W}}_{01})_{pp} & (\overline{\boldsymbol{W}}_{02})_{pp} & \cdots & (\overline{\boldsymbol{W}}_{0N})_{pp} \end{bmatrix}^T. \end{aligned}$$

The solution of linear simultaneous equations in (A·8) is given by

$$\begin{aligned} \boldsymbol{\alpha}_p &= (\boldsymbol{X}_{1p} - \boldsymbol{Y}_p\boldsymbol{X}_{2p}^{-1}\boldsymbol{Y}_p^T)^{-1}(\boldsymbol{z}_{1p} - \boldsymbol{Y}_p\boldsymbol{X}_{2p}^{-1}\boldsymbol{z}_{2p}) \\ \boldsymbol{\beta}_p &= (\boldsymbol{X}_{2p} - \boldsymbol{Y}_p^T\boldsymbol{X}_{1p}^{-1}\boldsymbol{Y}_p)^{-1}(\boldsymbol{z}_{2p} - \boldsymbol{Y}_p^T\boldsymbol{X}_{1p}^{-1}\boldsymbol{z}_{1p}) \end{aligned} \tag{A·9}$$

for $p = 1, 2, \cdots, n$.

Akimitsu Doi received the B.E. degree in Electronic Engineering from Tottori University in 1992, and M.E. and Ph.D. degrees in System Engineering from Hiroshima University in 1994 and 1997, respectively. From 1997 to 2002, he was a Research Associate with the Faculty of Engineering, Hiroshima University, Hiroshima, Japan. From 2002 to 2005, he was a Lecturer in the Faculty of Engineering, Hiroshima Institute of Technology, Hiroshima, Japan. Since 2005, he has been an Associate Professor at the same Institute. His research interests are in the area of digital signal and image processing.

Takao Hinamoto received the B.E. degree from Okayama University, Okayama, Japan, in 1969, the M.E. degree from Kobe University, Kobe, Japan, in 1971, and the Dr. Eng. degree from Osaka University, Osaka, Japan, in 1977, all in electrical engineering. From 1972 to 1988, he was with the Faculty of Engineering, Kobe University, Kobe, Japan. From 1979 to 1981, he was a Visiting Member of Staff in the Department of Electrical Engineering, Queen's University, Kingston, ON, Canada, on leave from Kobe University. During 1988–1991, he was Professor of electronic circuits in the Faculty of Engineering, Tottori University, Tottori, Japan. During 1992–2009, he was Professor of Electronic Control in the Department of Electrical Engineering, Hiroshima University, Hiroshima, Japan. Since 2009, he has been an Emeritus Professor of Hiroshima University, and a Teaching Member of Staff at Hiroshima Institute of Technology, Hiroshima, Japan. His research interests include digital signal processing, system theory, and control engineering. He has published over 400 papers in these areas. He is co-editor and co-author of Two-Dimensional Signal and Image Processing (Tokyo, Japan: SICE, 1996). Dr. Hinamoto was Guest Editor of the special sections on Digital Signal Processing, and Adaptive Signal Processing and Its Applications in the IEICE Trans. on Fundamentals in August 1998 and March 2005, respectively. He was Co-Guest Editor of the special section on Recent Advances in Circuits and Systems in the July and August 2005 issues of IEICE Trans. on Information and Systems. He was Chair of the 12th DSP Symposium held in Hiroshima, Japan, in November 1997, sponsored by the DSP Technical Committee of IEICE. From 1993 to 2000, he was a Senator or Member of the Board of Directors in the Society of Instrument and Control Engineers (SICE), and from 1999 to 2001 he was Chair of the Chugoku Chapter of SICE. From 2003 to 2004, he served as Chair of the DSP Technical Committee of IEICE and Chair of the Chugoku Chapter of IEICE. Dr. Hinamoto received the IEEE Third Millennium Medal in January 2000. From 1993 to 1995 he was an Associate Editor of the IEEE Trans. on Circuits and Systems II. During 2002–2003 and 2006–2007 he was an Associate Editor of the IEEE Trans. on Circuits and Systems I. He also served as General Chair of the 47th IEEE Int. Midwest Symposium on Circuits and Systems held in Hiroshima, Japan, in July 2004. Since 1995, he has been a Steering Committee Member of the IEEE Int. Midwest Symposium on Circuits and Systems. Since 1998, he has been a Digital Signal Processing Technical Committee Member in the IEEE Circuits and Systems Society. He played a leading role in establishing the Hiroshima Section of IEEE and served as Interim Chair of the Section. Dr. Hinamoto is a Life Fellow of IEEE as well as a Fellow of SICE.

Wu-Sheng Lu received the B.Sc. degree in Mathematics from Fudan University, Shanghai, China, in 1964, the M.S. degree in electrical engineering, and the Ph.D. degree in control science from the University of Minnesota, Minneapolis, USA, in 1983 and 1984, respectively. He was a post-doctoral fellow at the University of Victoria, Victoria, B.C., Canada, in 1985 and held a visiting assistant professorship with the University of Minnesota in 1986. Since 1987, he has been with the University of Victoria where he is currently a Professor. His current teaching and research interests are in the general areas of digital signal processing and application of optimization methods. He is the co-author with A. Antoniou of Two-Dimensional Digital Filters (Marcel Dekker, 1992). and Practical Optimization: Algorithms and Engineering Applications (Springer, 2007). Dr. Lu served as an Associate Editor of the Canadian Journal of Electrical and Computer Engineering in 1989, and was Editor of the same journal from 1990 to 1992. He was an Associate Editor of the IEEE Trans. on Circuits and Systems II from 1993 to 1995, for IEEE Trans. on Circuits and Systems I from 1999 to 2001 and from 2004 to 2005, for International Journal of Multidimensional Systems and Signal Processing from 1997 to 2006, and for Circuits, Systems, and Signal Processing from 2009 to 2011. Dr. Lu is a Life Fellow of IEEE.