l_2-Sensitivity Analysis and Minimization for Generalized Direct-Form II Realization of 2-D Separable-Denominator Digital Filters

Takao Hinamoto
Hiroshima University
Higashi-Hiroshima 739-8527, Japan
Email: hinamoto@ieee.org

Akimitsu Doi
Hiroshima Institute of Technology
Hiroshima 731-5193, Japan
Email: doi@cc.it-hiroshima.ac.jp

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

***Abstract*—In this paper, l_2-sensitivity for the generalized direct-form II state-space realization of two-dimensional (2-D) separable-denominator digital filters is analyzed and minimized with respect to free parameters subject to l_2-scaling constraints. Improved l_2-sensitivity for 2-D separable-denominator digital filters is investigated by taking 0 and ± 1 elements into account. An l_2-sensitivity measure for the generalized direct-form II state-space realization of 2-D separable-denominator digital filters is deduced by applying improved l_2-sensitivity. The l_2-sensitivity measure is then minimized with respect to free parameters subject to l_2-scaling constraints by employing an exhaustive search over a set of discrete values of *finite* cardinality. A numerical example is included to illustrate the validity and effectiveness of the proposed technique.**

I. Introduction

In the finite-word-length (FWL) fixed-point implementation of IIR digital filters, the characteristics of an actual transfer function deviate from the original due to the truncation or rounding of filter coefficients. Several techniques have been presented for synthesizing 2-D separable-denominator digital filter structures with low coefficient sensitivity [1]-[3]. Some of these techniques evaluate the sensitivity by using a mixture of l_1/l_2 norms [1], while the others rely on the use of a pure l_2 norm [2],[3]. The l_2-sensitivity minimization is more natural and reasonable than the conventional l_1/l_2 mixed sensitivity minimization, but it is technically more challenging. Improved l_2-sensitivity considering 0 and ± 1 elements has also been examined for 1-D and 2-D state-space digital filters [4],[5]. Alternatively, delta operator has widely been used in the realization of digital filters to improve FWL performance in the filters with high sampling rate [6]-[10]. Li and Gevers [6] have studied a roundoff noise gain for the delta-operator state-space realization of a transfer function and conditions under which the roundoff noise gain for the optimal delta-operator realization is smaller than that for the optimal shift-operator realization. In [7], the filter is expressed by second-order sections connected in cascade, each section is implemented with a direct form in delta operator, and different direct forms in the delta operator are extensively compared. In [8], the concept of separately scaling the Δs and filter coefficients in the delta transposed direct-form II section is proposed to globally further minimize output roundoff noise gain, as compared to selecting a single optimal Δ only. In [9], a method for obtaining an optimal arbitrary-order delta-operator direct-form II transposed filter has been presented in terms of the roundoff noise gain and coefficient sensitivity. Based on the concept of polynomial operators, a new structure being considered a generalization of the shift operator z-based direct-form II transposed structure for digital filter implementation has been explored [10].

This paper is concerned with analysis and minimization of l_2-sensitivity for the generalized direct-form II state-space realization of 2-D separable-denominator digital filters in [11]. Improved l_2-sensitivity for 2-D separable-denominator state-space digital filters is investigated by taking into account 0 and ± 1 elements. An improved l_2-sensitivity measure is then applied to the generalized direct-form II state-space realization of 2-D separable-denominator digital filters. The l_2-sensitivity measure is then minimized with respect to free parameters subject to l_2-scaling constraints by employing an exhaustive search over a set of discrete values of *finite* cardinality [11]. Unlike [11] where the roundoff noise has been treated, here we deal with l_2-sensitivity for 2-D state-space realization instead.

II. Analysis of Improved l_2-Sensitivity

2-D separable-denominator digital filters can be described by the Roesser model $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_4, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}_1, \boldsymbol{c}_2, d)_{m+n}$ as

$$\begin{bmatrix} \boldsymbol{x}^h(i+1,j) \\ \boldsymbol{x}^v(i,j+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{0} & \boldsymbol{A}_4 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^h(i,j) \\ \boldsymbol{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} u(i,j)$$
$$y(i,j) = \begin{bmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^h(i,j) \\ \boldsymbol{x}^v(i,j) \end{bmatrix} + d\, u(i,j) \tag{1}$$

where $\boldsymbol{x}^h(i,j)$ is an $m \times 1$ horizontal state vector, $\boldsymbol{x}^v(i,j)$ is an $n \times 1$ vertical state vector, $u(i,j)$ is a scalar input, $y(i,j)$ is a scalar output, and $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_4$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ and d are real constant matrices of appropriate dimensions. The filter in (1) is assumed asymptotically stable, separately locally controllable and separately locally observable. The transfer function of the filter in (1) is given by

$$\begin{aligned} H(z_1,z_2) &= [\,\boldsymbol{c}_1 \;\; \boldsymbol{c}_2\,] \begin{bmatrix} z_1\boldsymbol{I}_m - \boldsymbol{A}_1 & -\boldsymbol{A}_2 \\ \boldsymbol{0} & z_2\boldsymbol{I}_n - \boldsymbol{A}_4 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} + d \\ &= \boldsymbol{c}_1(z_1\boldsymbol{I}_m - \boldsymbol{A}_1)^{-1}\boldsymbol{b}_1 + \boldsymbol{c}_2(z_2\boldsymbol{I}_n - \boldsymbol{A}_4)^{-1}\boldsymbol{b}_2 \\ &\quad + \boldsymbol{c}_1(z_1\boldsymbol{I}_m - \boldsymbol{A}_1)^{-1}\boldsymbol{A}_2(z_2\boldsymbol{I}_n - \boldsymbol{A}_4)^{-1}\boldsymbol{b}_2 + d \end{aligned} \tag{2}$$

Respectively, let a_{1kl}, a_{2kl}, and a_{4kl} be the (k,l)th elements of $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, and $\boldsymbol{A}_4$; b_{1k} and b_{2k} be the kth elements of $\boldsymbol{b}_1$

978-1-7281-0397-6/19/$31.00 ©2019 IEEE

and $\boldsymbol{b}_2$; and c_{1l} and c_{2l} be the lth elements of $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. Then an l_2-sensitivity measure for the filter in (1) is defined as

$$\begin{aligned} M = & \sum_{k=1}^{m}\sum_{l=1}^{m}\left|\left|\frac{\partial H(z_1,z_2)}{\partial a_{1kl}}\right|\right|_2^2+\sum_{k=1}^{n}\sum_{l=1}^{n}\left|\left|\frac{\partial H(z_1,z_2)}{\partial a_{4kl}}\right|\right|_2^2 \\ & +\sum_{k=1}^{m}\sum_{l=1}^{n}\left|\left|\frac{\partial H(z_1,z_2)}{\partial a_{2kl}}\right|\right|_2^2+\sum_{k=1}^{m}\left|\left|\frac{\partial H(z_1,z_2)}{\partial b_{1k}}\right|\right|_2^2 \\ & +\sum_{k=1}^{n}\left|\left|\frac{\partial H(z_1,z_2)}{\partial b_{2k}}\right|\right|_2^2+\sum_{l=1}^{m}\left|\left|\frac{\partial H(z_1,z_2)}{\partial c_{1l}}\right|\right|_2^2 \\ & +\sum_{l=1}^{n}\left|\left|\frac{\partial H(z_1,z_2)}{\partial c_{2l}}\right|\right|_2^2+\left|\left|\frac{\partial H(z_1,z_2)}{\partial d}\right|\right|_2^2 \end{aligned} \tag{3}$$

where $||\cdot||_2^2$ denotes the squared l_2-norm defined by

$$||f(z_1,z_2)||_2^2=\frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}|f(z_1,z_2)|^2\,\frac{dz_1dz_2}{z_1z_2}$$

Since coefficients 0 and ± 1 can precisely be realized in the implementation of FWL digital filters, the l_2-sensitivities are not affected by these coefficients. Therefore, by using (2), the sensitivity on the individual element of coefficient matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_4$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ and d can be written as

$$\begin{aligned} \frac{\partial H(z_1,z_2)}{\partial a_{1kl}} &= \psi(a_{1kl})\,\overline{\boldsymbol{e}}_l^T\boldsymbol{F}(z_1,z_2)\boldsymbol{Q}(z_1)\,\overline{\boldsymbol{e}}_k \\ \frac{\partial H(z_1,z_2)}{\partial a_{2kl}} &= \psi(a_{2kl})\,\hat{\boldsymbol{e}}_l^T\boldsymbol{P}(z_2)\boldsymbol{Q}(z_1)\,\overline{\boldsymbol{e}}_k \\ \frac{\partial H(z_1,z_2)}{\partial a_{4kl}} &= \psi(a_{4kl})\,\hat{\boldsymbol{e}}_l^T\boldsymbol{P}(z_2)\boldsymbol{G}(z_1,z_2)\,\hat{\boldsymbol{e}}_k \\ \frac{\partial H(z_1,z_2)}{\partial b_{1k}} &= \psi(b_{1k})\boldsymbol{Q}(z_1)\,\overline{\boldsymbol{e}}_k \\ \frac{\partial H(z_1,z_2)}{\partial b_{2k}} &= \psi(b_{2k})\boldsymbol{G}(z_1,z_2)\,\hat{\boldsymbol{e}}_k \\ \frac{\partial H(z_1,z_2)}{\partial c_{1l}} &= \psi(c_{1l})\,\overline{\boldsymbol{e}}_l^T\boldsymbol{F}(z_1,z_2) \\ \frac{\partial H(z_1,z_2)}{\partial c_{2l}} &= \psi(c_{2l})\,\hat{\boldsymbol{e}}_l^T\boldsymbol{P}(z_2),\quad \frac{\partial H(z_1,z_2)}{\partial d}=\psi(d) \end{aligned} \tag{4}$$

where

$$\psi(x)=\begin{cases}1 & \text{for}\quad x\neq 0,\,\pm 1\\ 0 & \text{for}\quad x=0,\,\pm 1\end{cases}$$

$$\begin{aligned} \boldsymbol{F}(z_1,z_2) &= (z_1\boldsymbol{I}_m-\boldsymbol{A}_1)^{-1}\big[\boldsymbol{b}_1+\boldsymbol{A}_2(z_2\boldsymbol{I}_n-\boldsymbol{A}_4)^{-1}\boldsymbol{b}_2\big] \\ \boldsymbol{G}(z_1,z_2) &= \big[\boldsymbol{c}_2+\boldsymbol{c}_1(z_1\boldsymbol{I}_m-\boldsymbol{A}_1)^{-1}\boldsymbol{A}_2\big](z_2\boldsymbol{I}_n-\boldsymbol{A}_4)^{-1} \\ \boldsymbol{P}(z_2) &= (z_2\boldsymbol{I}_n-\boldsymbol{A}_4)^{-1}\boldsymbol{b}_2,\quad \boldsymbol{Q}(z_1)=\boldsymbol{c}_1(z_1\boldsymbol{I}_m-\boldsymbol{A}_1)^{-1} \end{aligned}$$

and $\overline{\boldsymbol{e}}_k$ ($\hat{\boldsymbol{e}}_k$) is the kth column of an identity matrix $\boldsymbol{I}_m$ ($\boldsymbol{I}_n$). From [2], it follows that

$$\begin{aligned} & \frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{F}(z_1,z_2)\boldsymbol{Q}(z_1)\overline{\boldsymbol{e}}_k\overline{\boldsymbol{e}}_k^T\boldsymbol{Q}^H(z_1) \\ & \qquad\qquad\cdot\boldsymbol{F}^H(z_1,z_2)z_1^{-1}z_2^{-1}\partial z_1\partial z_2 \\ & =\frac{1}{2\pi j}\oint_{|z_1|=1}\boldsymbol{Q}(z_1)\overline{\boldsymbol{e}}_k\overline{\boldsymbol{e}}_k^T\boldsymbol{Q}^H(z_1) \\ & \qquad\cdot\left[\frac{1}{2\pi j}\oint_{|z_2|=1}\boldsymbol{F}(z_1,z_2)\boldsymbol{F}^H(z_1,z_2)\frac{\partial z_2}{z_2}\right]\frac{\partial z_1}{z_1} \\ & =\frac{1}{2\pi j}\oint_{|z_1|=1}\boldsymbol{Q}(z_1)\overline{\boldsymbol{e}}_k\overline{\boldsymbol{e}}_k^T\boldsymbol{Q}^H(z_1)\big(z_1\boldsymbol{I}_m-\boldsymbol{A}_1\big)^{-1} \\ & \qquad\cdot\Big[\boldsymbol{A}_2\boldsymbol{K}^v\boldsymbol{A}_2^T+\boldsymbol{b}_1\boldsymbol{b}_1^T\Big]\big(z_1^{-1}\boldsymbol{I}_m-\boldsymbol{A}_1^T\big)^{-1}\frac{\partial z_1}{z_1} \\ & =\sum_{i=1}^{m}\sigma_{1i}\frac{1}{2\pi j}\oint_{|z_1|=1}\big(z_1\boldsymbol{I}_m-\boldsymbol{A}_1\big)^{-1}\boldsymbol{u}_i\boldsymbol{Q}(z_1)\overline{\boldsymbol{e}}_k \\ & \qquad\cdot\overline{\boldsymbol{e}}_k^T\boldsymbol{Q}^H(z_1)\boldsymbol{u}_i^T\big(z_1^{-1}\boldsymbol{I}_m-\boldsymbol{A}_1^T\big)^{-1}\frac{\partial z_1}{z_1} \end{aligned} \tag{5a}$$

where $\boldsymbol{K}^v$ can be obtained by solving the Lyapunov equation $\boldsymbol{K}^v=\boldsymbol{A}_4\boldsymbol{K}^v\boldsymbol{A}_4^T+\boldsymbol{b}_2\boldsymbol{b}_2^T$, σ_{1i} and $\boldsymbol{u}_i$ denote the ith eigenvalue and eigenvector of $\boldsymbol{A}_2\boldsymbol{K}^v\boldsymbol{A}_2^T+\boldsymbol{b}_1\boldsymbol{b}_1^T$, respectively, namely

$$\boldsymbol{A}_2\boldsymbol{K}^v\boldsymbol{A}_2^T+\boldsymbol{b}_1\boldsymbol{b}_1^T=\sum_{i=1}^{m}\sigma_{1i}\boldsymbol{u}_i\boldsymbol{u}_i^T \tag{5b}$$

Similarly, we obtain

$$\begin{aligned} & \frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\boldsymbol{G}^H(z_1,z_2)\boldsymbol{P}^H(z_2)\hat{\boldsymbol{e}}_l\hat{\boldsymbol{e}}_l^T\boldsymbol{P}(z_2) \\ & \qquad\qquad\cdot\boldsymbol{G}(z_1,z_2)z_1^{-1}z_2^{-1}\partial z_1\partial z_2 \\ & =\sum_{i=1}^{n}\sigma_{2i}\frac{1}{2\pi j}\oint_{|z_1|=1}\big(z_2^{-1}\boldsymbol{I}_n-\boldsymbol{A}_4^T\big)^{-1}\boldsymbol{v}_i\boldsymbol{P}^H(z_2)\hat{\boldsymbol{e}}_l \\ & \qquad\cdot\hat{\boldsymbol{e}}_l^T\boldsymbol{P}(z_2)\boldsymbol{v}_i^T\big(z_2\boldsymbol{I}_n-\boldsymbol{A}_4\big)^{-1}\frac{\partial z_2}{z_2} \end{aligned} \tag{6a}$$

where

$$\boldsymbol{A}_2^T\boldsymbol{W}^h\boldsymbol{A}_2+\boldsymbol{c}_2^T\boldsymbol{c}_2=\sum_{i=1}^{n}\sigma_{2i}\boldsymbol{v}_i\boldsymbol{v}_i^T \tag{6b}$$

and $\boldsymbol{W}^h$ can be obtained by solving the Lyapunov equation $\boldsymbol{W}^h=\boldsymbol{A}_1^T\boldsymbol{W}^h\boldsymbol{A}_1+\boldsymbol{c}_1^T\boldsymbol{c}_1$. It is also noted that

$$\begin{aligned} & \big(z_1\boldsymbol{I}_m-\boldsymbol{A}_1\big)^{-1}\boldsymbol{u}_i\boldsymbol{Q}(z_1)\,\overline{\boldsymbol{e}}_k \\ & \quad=\begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{0}\end{bmatrix}\left(z_1\boldsymbol{I}_{2m}-\begin{bmatrix}\boldsymbol{A}_1 & \boldsymbol{u}_i\boldsymbol{c}_1\\ \boldsymbol{0} & \boldsymbol{A}_1\end{bmatrix}\right)^{-1}\begin{bmatrix}\boldsymbol{0}\\ \overline{\boldsymbol{e}}_k\end{bmatrix} \\ & \hat{\boldsymbol{e}}_l^T\boldsymbol{P}(z_2)\boldsymbol{v}_i^T\big(z_2\boldsymbol{I}_n-\boldsymbol{A}_4\big)^{-1} \\ & \quad=\begin{bmatrix}\hat{\boldsymbol{e}}_l^T & \boldsymbol{0}\end{bmatrix}\left(z_2\boldsymbol{I}_{2n}-\begin{bmatrix}\boldsymbol{A}_4 & \boldsymbol{b}_2\boldsymbol{v}_i^T\\ \boldsymbol{0} & \boldsymbol{A}_4\end{bmatrix}\right)^{-1}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{I}_n\end{bmatrix} \end{aligned} \tag{7}$$

By virtue of (4)–(7), it is possible to express (3) as

$$\begin{aligned} M= & \sum_{k=1}^{m}\sum_{l=1}^{m}\psi(a_{1kl})\begin{bmatrix}\overline{\boldsymbol{e}}_l^T & \boldsymbol{0}\end{bmatrix}\Big(\sum_{i=1}^{m}\sigma_{1i}\boldsymbol{M}_{ki}\Big)\begin{bmatrix}\overline{\boldsymbol{e}}_l\\ \boldsymbol{0}\end{bmatrix} \\ & +\sum_{k=1}^{n}\sum_{l=1}^{n}\psi(a_{4kl})\begin{bmatrix}\boldsymbol{0} & \hat{\boldsymbol{e}}_k^T\end{bmatrix}\Big(\sum_{i=1}^{n}\sigma_{2i}\boldsymbol{N}_{li}\Big)\begin{bmatrix}\boldsymbol{0}\\ \hat{\boldsymbol{e}}_k\end{bmatrix} \\ & +\sum_{k=1}^{m}\sum_{l=1}^{n}\psi(a_{2kl})\,\overline{\boldsymbol{e}}_k^T\boldsymbol{W}^h\overline{\boldsymbol{e}}_k\cdot\hat{\boldsymbol{e}}_l^T\boldsymbol{K}^v\hat{\boldsymbol{e}}_l \\ & +\sum_{k=1}^{m}\psi(b_{1k})\,\overline{\boldsymbol{e}}_k^T\boldsymbol{W}^h\overline{\boldsymbol{e}}_k+\sum_{k=1}^{n}\psi(b_{2k})\,\hat{\boldsymbol{e}}_k^T\boldsymbol{W}^v\hat{\boldsymbol{e}}_k \\ & +\sum_{l=1}^{m}\psi(c_{1l})\,\overline{\boldsymbol{e}}_l^T\boldsymbol{K}^h\overline{\boldsymbol{e}}_l+\sum_{l=1}^{n}\psi(c_{2l})\,\hat{\boldsymbol{e}}_l^T\boldsymbol{K}^v\hat{\boldsymbol{e}}_l+\psi(d) \end{aligned} \tag{8}$$

where Grammians $\boldsymbol{M}_{ki}$ for $k,i=1,2,\cdots,m$, $\boldsymbol{N}_{li}$ for $l,i=1,2,\cdots,n$, $\boldsymbol{K}^h$ and $\boldsymbol{W}^v$ can be obtained by solving the Lyapunov equations

$$\begin{aligned}
\boldsymbol{M}_{ki} &= \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{u}_i\boldsymbol{c}_1 \\ \boldsymbol{0} & \boldsymbol{A}_1 \end{bmatrix} \boldsymbol{M}_{ki} \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{u}_i\boldsymbol{c}_1 \\ \boldsymbol{0} & \boldsymbol{A}_1 \end{bmatrix}^T + \begin{bmatrix} \boldsymbol{0} \\ \overline{\boldsymbol{e}}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{0} \\ \overline{\boldsymbol{e}}_k \end{bmatrix}^T \\
\boldsymbol{N}_{li} &= \begin{bmatrix} \boldsymbol{A}_4 & \boldsymbol{b}_2\boldsymbol{v}_i^T \\ \boldsymbol{0} & \boldsymbol{A}_4 \end{bmatrix}^T \boldsymbol{N}_{li} \begin{bmatrix} \boldsymbol{A}_4 & \boldsymbol{b}_2\boldsymbol{v}_i^T \\ \boldsymbol{0} & \boldsymbol{A}_4 \end{bmatrix} + \begin{bmatrix} \hat{\boldsymbol{e}}_l \\ \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{e}}_l \\ \boldsymbol{0} \end{bmatrix}^T \\
\boldsymbol{K}^h &= \boldsymbol{A}_1\boldsymbol{K}^h\boldsymbol{A}_1^T + \boldsymbol{A}_2\boldsymbol{K}^v\boldsymbol{A}_2^T + \boldsymbol{b}_1\boldsymbol{b}_1^T \\
\boldsymbol{W}^v &= \boldsymbol{A}_4^T\boldsymbol{W}^v\boldsymbol{A}_4 + \boldsymbol{A}_2^T\boldsymbol{W}^h\boldsymbol{A}_2 + \boldsymbol{c}_2^T\boldsymbol{c}_2
\end{aligned}$$

Notice that unlike [2] where eigenvalue-eigenvector decompositions of $\boldsymbol{K}^v$ and $\boldsymbol{W}^h$ are performed to compute (8) with all $\psi(\cdot) = 1$, here we perform eigenvalue-eigenvector decompositions of $\boldsymbol{A}_2\boldsymbol{K}^v\boldsymbol{A}_2^T + \boldsymbol{b}_1\boldsymbol{b}_1^T$ and $\boldsymbol{A}_2^T\boldsymbol{W}^h\boldsymbol{A}_2 + \boldsymbol{c}_2^T\boldsymbol{c}_2$ instead.

III. ANALYSIS AND MINIMIZATION OF l_2-SENSITIVITY FOR GENERALIZED DIRECT-FORM II REALIZATION OF 2-D FILTERS

A. Analysis of l_2-Sensitivity

For the generalized direct-form II structure of a 2-D separable-denominator digital filter, the transfer function in (2) can be converted into [11]

$$H(z_1, z_2) = \frac{\displaystyle\sum_{k=0}^{m}\sum_{l=0}^{n} r_{kl} \prod_{p=0}^{k} \rho_p^h(z_1)^{-1} \prod_{q=0}^{l} \rho_q^v(z_2)^{-1}}{\displaystyle\left(\sum_{k=0}^{m} \alpha_k \prod_{p=0}^{k} \rho_p^h(z_1)^{-1}\right)\left(\sum_{l=0}^{n} \beta_l \prod_{q=0}^{l} \rho_q^v(z_2)^{-1}\right)} \tag{9}$$

where $\alpha_0 = \beta_0 = 1$ and $\rho_0^h(z_1)^{-1} = \rho_0^v(z_2)^{-1} = 1$. The implementations of generalized delta operators $\rho_k^h(z_1)^{-1}$ and $\rho_l^v(z_2)^{-1}$ are depicted in Fig. 1. As an illustrative example, the structure of (9) for a 2-D digital filter with $(m, n) = (3, 3)$ is depicted in Fig. 2 where $u(i, j)$ is a scalar input and $y(i, j)$ is a scalar output.

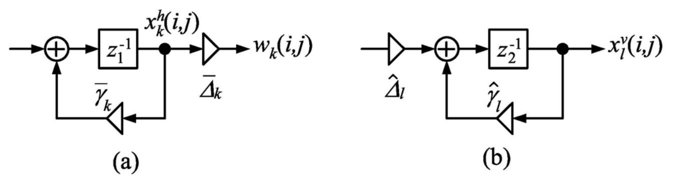

Fig. 1. (a) Implementation of $\rho_k^h(z_1)^{-1}$ and (b) that of $\rho_l^v(z_2)^{-1}$.

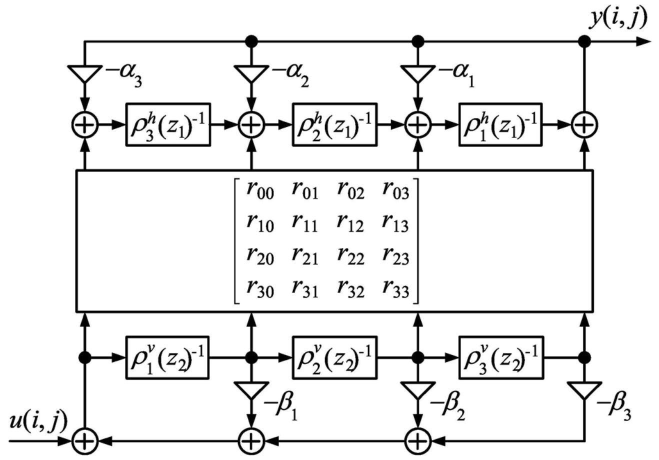

Fig. 2. The 2-D filter structure of (9) with $(m, n) = (3, 3)$.

The 2-D filter in (9) can be realized by the Roesser model $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_4, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{c}_1, \boldsymbol{c}_2, d)_{m,n}$ in (1) with [11]

$$\boldsymbol{A}_1 = \begin{bmatrix} -\alpha_1\overline{\Delta}_1 & \overline{\Delta}_2 & \cdots & 0 \\ -\alpha_2\overline{\Delta}_1 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \overline{\Delta}_m \\ -\alpha_m\overline{\Delta}_1 & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} \overline{\gamma}_1 & 0 & \cdots & 0 \\ 0 & \overline{\gamma}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \overline{\gamma}_m \end{bmatrix}$$

$$\boldsymbol{A}_4 = \begin{bmatrix} -\beta_1\hat{\Delta}_1 & -\beta_2\hat{\Delta}_1 & \cdots & -\beta_n\hat{\Delta}_1 \\ \hat{\Delta}_2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \hat{\Delta}_n & 0 \end{bmatrix} + \begin{bmatrix} \hat{\gamma}_1 & 0 & \cdots & 0 \\ 0 & \hat{\gamma}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \hat{\gamma}_n \end{bmatrix}$$

$$\begin{aligned}
\boldsymbol{A}_2 &= \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix} - \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} [r_{01} \;\; r_{02} \;\; \cdots \;\; r_{0n}] \\
&\quad - \begin{bmatrix} r_{10} \\ r_{20} \\ \vdots \\ r_{m0} \end{bmatrix} [\beta_1 \; \beta_2 \cdots \beta_n] + r_{00} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} [\beta_1 \; \beta_2 \cdots \beta_n] \\
\boldsymbol{b}_1 &= \begin{bmatrix} r_{10} \\ r_{20} \\ \vdots \\ r_{m0} \end{bmatrix} - r_{00} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix}, \qquad \boldsymbol{b}_2 = \begin{bmatrix} \hat{\Delta}_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
\boldsymbol{c}_1 &= [\, \overline{\Delta}_1 \quad 0 \quad \cdots \quad 0 \,] \\
\boldsymbol{c}_2 &= [\, r_{01} \;\; r_{02} \;\; \cdots \;\; r_{0n} \,] - r_{00}\, [\, \beta_1 \; \beta_2 \cdots \beta_n \,], \quad d = r_{00}
\end{aligned} \tag{10}$$

where l_2-scaling constraints can be satisfied by appropriately choosing scalars $\overline{\Delta}_1, \overline{\Delta}_2, \cdots, \overline{\Delta}_m$ and $\hat{\Delta}_1, \hat{\Delta}_2, \cdots, \hat{\Delta}_n$ [11].

For the local state-space realization specified by (10), the l_2-sensitivity measure, say $M_{S\rho}$, can be evaluated by (8) with

$$\begin{aligned}
\psi(a_{1kl}) &= \begin{cases} 1 & \text{for } k = 1, 2, \cdots, m \text{ and } l = 1 \\ 1 & \text{for } k = 1, 2, \cdots, m-1 \text{ and } l = k+1 \\ \psi(\overline{\gamma}_k) & \text{for } k = 2, 3, \cdots, m \text{ and } l = k \\ 0 & \text{otherwise} \end{cases} \\
\psi(a_{4kl}) &= \begin{cases} 1 & \text{for } k = 1 \text{ and } l = 1, 2, \cdots, n \\ 1 & \text{for } k = 2, 3, \cdots, n \text{ and } l = k-1 \\ \psi(\hat{\gamma}_k) & \text{for } k = 2, 3, \cdots, n \text{ and } l = k \\ 0 & \text{otherwise} \end{cases} \\
\psi(a_{2kl}) &= 1 \quad \text{for } k = 1, 2, \cdots, m \text{ and } l = 1, 2, \cdots, n \\
\psi(b_{1k}) &= 1 \quad \text{for } k = 1, 2, \cdots, m \\
\psi(b_{2k}) &= \begin{cases} 1 & \text{for } k = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \psi(c_{1l}) = \begin{cases} 1 & \text{for } l = 1 \\ 0 & \text{otherwise} \end{cases} \\
\psi(c_{2l}) &= 1 \quad \text{for } l = 1, 2, \cdots, n
\end{aligned} \tag{11}$$

which results in

$$
\begin{aligned}
M_{S\rho} = &\sum_{k=1}^{m} \left[\begin{array}{cc} \overline{\boldsymbol{e}}_1^T & \boldsymbol{0} \end{array}\right] \Big(\sum_{i=1}^{m} \sigma_{1i} \boldsymbol{M}_{ki}\Big) \left[\begin{array}{c} \overline{\boldsymbol{e}}_1 \\ \boldsymbol{0} \end{array}\right] \\
&+ \sum_{k=1}^{m-1} \left[\begin{array}{cc} \overline{\boldsymbol{e}}_{k+1}^T & \boldsymbol{0} \end{array}\right] \Big(\sum_{i=1}^{m} \sigma_{1i} \boldsymbol{M}_{ki}\Big) \left[\begin{array}{c} \overline{\boldsymbol{e}}_{k+1} \\ \boldsymbol{0} \end{array}\right] \\
&+ \sum_{k=2}^{m} \psi(\overline{\gamma}_k) \left[\begin{array}{cc} \overline{\boldsymbol{e}}_k^T & \boldsymbol{0} \end{array}\right] \Big(\sum_{i=1}^{m} \sigma_{1i} \boldsymbol{M}_{ki}\Big) \left[\begin{array}{c} \overline{\boldsymbol{e}}_k \\ \boldsymbol{0} \end{array}\right] \\
&+ \sum_{l=1}^{n} \left[\begin{array}{cc} \boldsymbol{0} & \hat{\boldsymbol{e}}_1^T \end{array}\right] \Big(\sum_{i=1}^{n} \sigma_{2i} \boldsymbol{N}_{li}\Big) \left[\begin{array}{c} \boldsymbol{0} \\ \hat{\boldsymbol{e}}_1 \end{array}\right] \\
&+ \sum_{k=2}^{n} \left[\begin{array}{cc} \boldsymbol{0} & \hat{\boldsymbol{e}}_k^T \end{array}\right] \Big(\sum_{i=1}^{n} \sigma_{2i} \boldsymbol{N}_{k-1,i}\Big) \left[\begin{array}{c} \boldsymbol{0} \\ \hat{\boldsymbol{e}}_k \end{array}\right] \\
&+ \sum_{k=2}^{n} \psi(\hat{\gamma}_k) \left[\begin{array}{cc} \boldsymbol{0} & \hat{\boldsymbol{e}}_k^T \end{array}\right] \Big(\sum_{i=1}^{n} \sigma_{2i} \boldsymbol{N}_{ki}\Big) \left[\begin{array}{c} \boldsymbol{0} \\ \hat{\boldsymbol{e}}_k \end{array}\right] \\
&+ \mathrm{tr}[\boldsymbol{W}^h]\mathrm{tr}[\boldsymbol{K}^v] + \mathrm{tr}[\boldsymbol{W}^h] + \hat{\boldsymbol{e}}_1^T \boldsymbol{W}^v \hat{\boldsymbol{e}}_1 \\
&+ \overline{\boldsymbol{e}}_1^T \boldsymbol{K}^h \overline{\boldsymbol{e}}_1 + \mathrm{tr}[\boldsymbol{K}^v] + \psi(d)
\end{aligned} \tag{12}
$$

B. Minimization of l_2-Sensitivity $M_{S\rho}$ w.r.t. $\{\overline{\gamma}_i\}$ and $\{\hat{\gamma}_l\}$

The l_2-sensitivity $M_{S\rho}$ is a function of parameters $\overline{\gamma}_k$ for $k = 1, 2, \cdots, m$ and $\hat{\gamma}_l$ for $l = 1, 2, \cdots, n$. Theoretically, these parameters can be chosen from a set of real numbers. However, our attention is confined to fix-point implementation where the FWL effects are more serious, and therefore we assume in the sequel that $|\overline{\gamma}_k| \le 1$ for $k = 1, 2, \cdots, m$ and $|\hat{\gamma}_l| \le 1$ for $l = 1, 2, \cdots, n$. For a fixed-point implementation of B_c bits, every $\overline{\gamma}_i$ ($\hat{\gamma}_l$) must be truncated or rounded into a B_γ-bit number ($B_\gamma \le B_c$) of the form

$$
v = \pm \sum_{p=1}^{B_\gamma} b_p 2^{-p}, \quad b_p = 0, 1 \ \ \forall\, p \tag{13}
$$

unless $v = \pm 1$. Hence $\{\overline{\gamma}_i\}$ and $\{\hat{\gamma}_l\}$ should take values from a set of discrete values of *finite* cardinality defined by

$$
S_\gamma = \{-1, 1\} \cup \left\{ \pm \sum_{p=1}^{B_\gamma} b_p 2^{-p}, \quad b_p = 0, 1 \ \ \forall\, p \right\} \tag{14}
$$

which contains $(2^{B_\gamma + 1} + 1)$ elements.

Let $\tilde{S}_{\boldsymbol{\gamma}} \in \boldsymbol{R}^{(m+n)\times 1}$ be the discrete set in which $\boldsymbol{\gamma}$ resides, i.e.,

$$
\tilde{S}_{\boldsymbol{\gamma}} = \{\boldsymbol{\gamma} \,|\, \gamma_k \in S_\gamma \ \ \forall\, k\} \tag{15}
$$

where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \cdots, \gamma_{m+n}]^T = [\overline{\gamma}_1, \overline{\gamma}_2, \cdots, \overline{\gamma}_m, \hat{\gamma}_1, \hat{\gamma}_2, \cdots, \hat{\gamma}_n]^T$. Obviously, $M_{S\rho}$ in (12) is a function of vector $\boldsymbol{\gamma}$ because all the Grammians depend on $\boldsymbol{\gamma}$. Consequently, the problem of minimizing $M_{S\rho}$ with respect to vector $\boldsymbol{\gamma}$ can be formulated as

$$
\boldsymbol{\gamma}\big(M_{S\rho}^{opt}\big) = \arg \min_{\boldsymbol{\gamma} \in \tilde{S}_{\boldsymbol{\gamma}}} M_{S\rho} \tag{16}
$$

The l_2-sensitivity measure $M_{S\rho}$ in (12) can be minimized with respect to $m+n$ free parameters $\{\overline{\gamma}_k\}$ and $\{\hat{\gamma}_l\}$ subject to l_2-scaling constraints through an exhaustive search within a finite set of discrete values in a manner similar to that in [11].

IV. A Numerical Example

Consider a 2-D stable separable-denominator digital filter of order $(m, n) = (3, 3)$ in (1) with [11]

$$
\begin{aligned}
\left[\begin{array}{ccc} a_1 & a_2 & a_3 \end{array}\right] &= \left[\begin{array}{ccc} -2.173645 & 1.836929 & -0.599655 \end{array}\right] \\
\left[\begin{array}{ccc} b_1 & b_2 & b_3 \end{array}\right] &= \left[\begin{array}{ccc} -2.280029 & 1.887939 & -0.564961 \end{array}\right]
\end{aligned}
$$

$$
[c_{kl}] = \left[\begin{array}{rrrr} 0.019421 & -0.027724 & 0.011468 & -0.000087 \\ 0.004839 & 0.017545 & -0.050267 & 0.033061 \\ -0.004328 & -0.008847 & 0.096260 & -0.083801 \\ -0.000138 & 0.007979 & -0.052927 & 0.062607 \end{array}\right]
$$

When the l_2-scaling constraints were satisfied by Choosing $\overline{\Delta}_k$ and $\hat{\Delta}_l$ appropriately [11], the filter coefficients in (9) were derived in the case where $\overline{\gamma}_k = \hat{\gamma}_l = 1$ for $k, l = 1, 2, 3$ as

$$
\begin{aligned}
\left[\begin{array}{ccc} \alpha_1 & \alpha_2 & \alpha_3 \end{array}\right] &= \left[\begin{array}{ccc} 0.832776 & 0.586915 & 0.166560 \end{array}\right] \\
\left[\begin{array}{ccc} \beta_1 & \beta_2 & \beta_3 \end{array}\right] &= \left[\begin{array}{ccc} 1.084549 & 0.851201 & 0.344142 \end{array}\right] \\
\left[\begin{array}{ccc} \overline{\Delta}_1 & \overline{\Delta}_2 & \overline{\Delta}_3 \end{array}\right] &= \left[\begin{array}{ccc} 0.992289 & 0.840742 & 0.457913 \end{array}\right] \\
\left[\begin{array}{ccc} \hat{\Delta}_1 & \hat{\Delta}_2 & \hat{\Delta}_3 \end{array}\right] &= \left[\begin{array}{ccc} 0.663843 & 0.580255 & 0.323990 \end{array}\right]
\end{aligned}
$$

$$
[r_{kl}] = 10^{-1} \left[\begin{array}{cccc} 0.1942100 & 0.460026 & 0.370760 & 0.246578 \\ 0.6359199 & 1.877514 & 1.103626 & 1.163573 \\ 0.7625064 & 2.417902 & 3.332929 & 1.812459 \\ 0.5181348 & 1.905871 & 2.841787 & 5.255824 \end{array}\right]
$$

The numerical results obtained by applying the technique similar to that in [11] for minimizing $M_{S\rho}$ in (12) with respect to vector $\boldsymbol{\gamma}$ subject to the l_2-scaling constraints were summarized in Table I in comparison with $\boldsymbol{\gamma}_z = [0,\ 0,\ \cdots,\ 0]^T$ and $\boldsymbol{\gamma}_\delta = [1,\ 1,\ \cdots,\ 1]^T$ where $\boldsymbol{\gamma}(M_{S\rho}^{opt})$ was found to be

$$
\boldsymbol{\gamma}(M_{S\rho}^{opt}) = [\ -0.875 \ \ 0.750 \ \ 0.750 \ \ 0.500 \ \ 1.000 \ \ 1.000]
$$

TABLE I. Performance Comparison Among Various $\boldsymbol{\gamma}$

	$\boldsymbol{\gamma}_z$	$\boldsymbol{\gamma}_\delta$	$\boldsymbol{\gamma}(M_{S\rho}^{opt})$
$M_{S\rho}$	2938.2976	128.8331	104.6422

From Table I, it is observed that the optimal realization yields l_2-sensitivity $\boldsymbol{\gamma}(M_{S\rho}^{opt}) = 104.6422$ that is very close to the minimum l_2-sensitivity $M_{min} = 102.0046$ in [3] where all the coefficients are non-trivial. Unlike [3], here the optimal realization includes many trivial coefficients so that the filter implementation cost can considerably be reduced.

V. Conclusion

In this paper, we have investigated improved l_2-sensitivity for 2-D separable-denominator digital filters by taking 0 and ± 1 elements into account. Unlike the improved l_2-sensitivity for more general 2-D digital filters in [5] where infinite sums are approximated by finite sums, the improved l_2-sensitivity for 2-D separable-denominator digital filters can be computed precisely by solving the Lyapunov equations. An l_2-sensitivity measure for the generalized direct-form II realization of 2-D separable-denominator digital filters has been deduced by applying the improved l_2-sensitivity. The resulting l_2-sensitivity measure has then been minimized with respect to $m + n$ free parameters subject to l_2-scaling constraints by employing an exhaustive search in a finite element space. In the last paragraph of Section IV, our design has been compared with that obtained by simply adapting the general l_2-sensitivity.

REFERENCES

[1] T. Hinamoto, T. Takao and M. Muneyasu, "Synthesis of 2-D separable-denominator digital filters with low sensitivity," *J. Franklin Institute*, vol.329, pp.1063-1080, 1992.

[2] T. Hinamoto and Y. Sugie, "L_2-sensitivity analysis and minimization of 2-D separable-denominator state-space digital filters," *IEEE Trans. Signal Processing*, vol.50, pp.3107-3114, Dec. 2002.

[3] T. Hinamoto, Y. Shibata and W.-S. Lu, "Minimization of L_2-sensitivity for 2-D separable-denominator state-space digital filters subject to L_2-scaling constraints," in *Proc. Asia Pasific Conf. Circuits and Systems (APCCAS 2006)*, pp. 390-3193, Singapore, Dec. 2006.

[4] C. Xiao, " improved L_2-sensitivity for state-space digital system," *IEEE Trans. Signal Process.*, vol. 45, pp. 837-840, Apr. 1997.

[5] Y. Hinamoto and A. Doi, "Analysis of l_2-sensitivity for two-dimensional state-space digital filters by taking into account 0 and ± 1 elements," in *Proc. IEEE Int. Midwest Symp. Circuits Syst.*, pp. 687-690, College Station, TX, Aug. 2014.

[6] G. Li and M. Gevers, "Roundoff Noise minimization using delta-operator realizations," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 629-637, Feb. 1993.

[7] J. Kauraniemi, T. I. Laakso, I. Hartimo and S. J. Ovaska, "Delta operator realizations of direct-form IIR filters," *IEEE Trans. Circuits Syst.-II*, vol. 45, no. 1, pp. 41-52, Jan. 1998.

[8] N. Wong and T.-S. Ng, "Roundoff noise minimization in a modified direct-form delta operator IIR structure," *IEEE Trans. Circuits Syst.-II*, vol. 47, no. 12, pp. 1533-1536, Dec. 2000.

[9] N. Wong and T.-S. Ng, "A generalized direct-form delta operator-based IIR filter with minimum noise gain and sensitivity," *IEEE Trans. Circuits Syst.-II*, vol. 48, no. 4, pp. 425-431, Apr. 2001.

[10] 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, no. 4, pp. 769-778, Apr. 2004.

[11] T. Hinamoto, A. Doi and W.-S. Lu, "Generalized direct-form II realization of 2-D separable-denominator digital filters", in *Proc. 2018 IEEE Int. Symp. Circuits Syst.*, Florence, Italy, May 2018.