Synthesis of 2-D State-Space Fixed-Point Digital-Filter Structures with Minimum Roundoff Noise

WU-SHENG LU, MEMBER, IEEE, AND ANDREAS ANTONIOU, FELLOW, IEEE

Abstract—**Based on a roundoff-noise analysis, a general synthesis procedure is developed which leads to an optimal local state-space 2-D digital-filter realization that minimizes the output-noise power due to roundoff subject to a scaling condition on the state variables. The output-noise power and the signal scaling condition are closely related to two positive-definite matrices W and K. These matrices provide two sets of invariants, called the 2-D second-order modes of the filter, which play a crucial role in the minimization of the output-noise power. With the availability of matrices W and K, the 2-D similarity transformation that yields an optimal state-space realization can be obtained by solving separately two 1-D optimization problems so that the well-developed techniques for minimizing roundoff noise in 1-D state-space digital filters can also be used for minimizing roundoff noise in 2-D state-space digital filters.**

I. Introduction

THE MINIMIZATION of roundoff noise in digital filters is of considerable practical significance since it leads to implementations with optimal signal-to-noise ratio. Minimum roundoff noise can be achieved in 1-D infinite-impulse-response (IIR) digital filters by using the method of Mullis and Roberts [1], [2], or that of Hwang [3], or by using the state-space structures of Bomar [4]. In the two-dimensional (2-D) case, the effects of finite precision in the implementation of recursive digital filters were considered in [5]–[7]. Recently, the synthesis of 2-D separable denominator digital filters with minimum roundoff noise has been considered in [8]. The 2-D counterpart of the fundamental work in [1]–[3], however, is not available to date and continues to be a significant open problem [9, p. 128], [10, p. 280].

The objective of this paper is to provide a solution to the following synthesis problem: given the transfer function of a 2-D digital filter, find the state-space realization that minimizes the output-noise power due to the roundoff of products, subject to l_2-norm dynamic range constraints. Based on Roesser's local state-space model (LSS) [11], a 2-D roundoff-noise analysis is carried out from which an explicit expression of the output-noise variance is derived. It is then shown that the output-noise power and the dynamic range constraints on state variables are naturally related to two positive-definite matrices W and K first reported in the literature by Mertzios [12]. These matrices provide two sets of 2-D second-order modes which play a crucial role in the minimization of roundoff noise. The main results of this paper are illustrated by two examples.

Manuscript received August 27, 1985; revised March 5, 1986. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

W.-S. Lu is with the Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

A. Antoniou is with the Department of Electrical Engineering, University of Victoria, Victoria, B.C., Canada, V8W 2Y2.

IEEE Log Number 8609853.

II. Local State-Space Model

A single-input single-output 2-D digital filter can be represented by the local state-space model (LSS) due to Roesser [11] given by

$$\begin{bmatrix} v(i+1,j) \\ h(i,j+1) \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \begin{bmatrix} v(i,j) \\ h(i,j) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} u(i,j)$$

$$\equiv A \begin{bmatrix} v(i,j) \\ h(i,j) \end{bmatrix} + bu(i,j) \tag{1a}$$

$$y(i,j) = [c_1 \quad c_2] \begin{bmatrix} v(i,j) \\ h(i,j) \end{bmatrix} + du(i,j)$$

$$\equiv c \begin{bmatrix} v(i,j) \\ h(i,j) \end{bmatrix} + du(i,j) \tag{1b}$$

where $v(i,j) \in R^m$, $h(i,j) \in R^n$, $A_1 \in R^{m \times m}$, and $A_4 \in R^{n \times n}$. Throughout the paper it is assumed that

$$\det \begin{bmatrix} I_m - z_1 A_1 & -z_1 A_2 \\ -z_2 A_3 & I_n - z_2 A_4 \end{bmatrix} \neq 0$$

$$\text{for } (z_1, z_2) \in \{(z_1, z_2) : |z_1| \leqslant 1, |z_2| \leqslant 1\} \tag{2}$$

which implies the asymptotical stability of the filter.

As in [11], let

$$A_{00} = I_{m+n}, A_{10} = \begin{bmatrix} A_1 & A_2 \\ 0 & 0 \end{bmatrix}, A_{01} = \begin{bmatrix} 0 & 0 \\ A_3 & A_4 \end{bmatrix}$$

$$A_{ij} = A_{10} A_{i-1,j} + A_{01} A_{i,j-1} \qquad \text{for } (i,j) > (0,0)$$

$$A_{-i,j} = A_{i,-j} = 0 \qquad \text{for } i \geqslant 1, j \geqslant 1. \tag{3}$$

The transfer function of the filter can be expressed in terms of A_{ij}, b, c, and d as

$$g(z_1, z_2) = c(I - z_1 A_{10} - z_2 A_{01})^{-1} \begin{bmatrix} z_1 b_1 \\ z_2 b_2 \end{bmatrix} + d$$

$$= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} c \left(A_{i-1,j} \begin{bmatrix} b_1 \\ 0 \end{bmatrix} + A_{i,j-1} \begin{bmatrix} 0 \\ b_2 \end{bmatrix} \right) z_1^i z_2^j + d. \tag{4}$$

0098-4094/86/1000-0965$01.00

Since the filter is stable, we have

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\left|c\left(A_{i-1,j}\begin{bmatrix}b_1\\0\end{bmatrix}+A_{i,j-1}\begin{bmatrix}0\\b_2\end{bmatrix}\right)\right|<\infty \tag{5}$$

and since (5) holds for all b and c, we deduce

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}|A_{ij}|<\infty \tag{6}$$

where $|A_{ij}|$ represents matrix A_{ij} with its entries replaced by their absolute values.

It is known that a 2-D similarity transformation matrix for the LSS model should be of form

$$T=\begin{bmatrix}T_1 & 0\\0 & T_2\end{bmatrix}\equiv T_1\oplus T_2 \tag{7}$$

where $T_1\in R^{m\times m}$, $T_2\in R^{n\times n}$, and $\oplus$ denotes the direct sum. This transformation leads to an equivalent realization characterized by $(\tilde{A},\tilde{b},\tilde{c})$, where

$$\tilde{A}=T^{-1}AT,\quad \tilde{b}=T^{-1}b,\quad \text{and } \tilde{c}=cT. \tag{8}$$

Once matrices $\tilde{A}_{ij}$ are defined by analogy with (3), it is easy to show that

$$\tilde{A}_{ij}=T^{-1}A_{ij}T \qquad \text{for } i\geq 0,\ j\geq 0. \tag{9}$$

III. A Roundoff Noise Analysis

A general analysis of roundoff noise applicable to 1-D state-space digital filters has been given by Hwang [13]. With a few minor modifications, a noise model can similarly be established for 2-D digital filters using Roesser's local state-space description. This model is then used to derive an explicit expression for the output-noise power.

A. Derivation of Noise Model

If finite wordlength effects due to input, coefficient, and product quantization are taken into consideration, the state-space model of an actual filter becomes

$$\begin{bmatrix}\bar{v}(i+1,j)\\\bar{h}(i,j+1)\end{bmatrix}=\bar{A}\begin{bmatrix}\bar{v}(i,j)\\\bar{h}(i,j)\end{bmatrix}+\bar{b}\bar{u}(i,j)+\alpha(i,j)+\beta(i,j) \tag{10a}$$

$$\bar{y}(i,j)=\bar{c}\begin{bmatrix}\bar{v}(i,j)\\\bar{h}(i,j)\end{bmatrix}+\bar{d}\bar{u}(i,j)+\gamma(i,j)+\delta(i,j) \tag{10b}$$

where $\bar{A}=A+\Delta A$, $\bar{b}=b+\Delta b$, $\bar{c}=c+\Delta c$, and $\bar{d}=d+\Delta d$ are the finite wordlength implementations of matrices A, b, c, and d, respectively; $\bar{u}=u+\Delta u$; $\alpha(i,j)\in R^{m+n}$, $\beta(i,j)\in R^{m+n}$, $\gamma(i,j)\in R$, and $\delta(i,j)\in R$ are random errors generated by product quantization in (10).

Define the state-error vector as

$$\begin{bmatrix}\Delta v(i,j)\\\Delta h(i,j)\end{bmatrix}=\begin{bmatrix}\bar{v}(i,j)\\\bar{h}(i,j)\end{bmatrix}-\begin{bmatrix}v(i,j)\\h(i,j)\end{bmatrix}$$

and let

$$\bar{A}=\begin{bmatrix}\bar{A}_1 & \bar{A}_2\\\bar{A}_3 & \bar{A}_4\end{bmatrix},\ \bar{A}_{10}=\begin{bmatrix}\bar{A}_1 & \bar{A}_2\\0 & 0\end{bmatrix},$$

$$\bar{A}_{01}=\begin{bmatrix}0 & 0\\\bar{A}_3 & \bar{A}_4\end{bmatrix}$$

$$\Delta A=\begin{bmatrix}\Delta A_1 & \Delta A_2\\\Delta A_3 & \Delta A_4\end{bmatrix},$$

$$\Delta A_{10}=\begin{bmatrix}\Delta A_1 & \Delta A_2\\0 & 0\end{bmatrix},\ \Delta A_{01}=\begin{bmatrix}0 & 0\\\Delta A_3 & \Delta A_4\end{bmatrix}$$

$$\bar{\tau}(i,j)=\bar{b}\Delta u(i,j)+\Delta b u(i,j)+\alpha(i,j)+\beta(i,j)\equiv\begin{bmatrix}\bar{\tau}_1(i,j)\\\bar{\tau}_2(i,j)\end{bmatrix}$$

and

$$\Delta y(i,j)=\bar{y}(i,j)-y(i,j).$$

Straightforward manipulation yields

$$\begin{bmatrix}\Delta v(i,j)\\\Delta h(i,j)\end{bmatrix}=\bar{A}_{10}\begin{bmatrix}\Delta v(i-1,j)\\\Delta h(i-1,j)\end{bmatrix}+\bar{A}_{01}\begin{bmatrix}\Delta v(i,j-1)\\\Delta h(i,j-1)\end{bmatrix}+\Delta A_{10}\begin{bmatrix}v(i-1,j)\\h(i-1,j)\end{bmatrix}+\Delta A_{01}\begin{bmatrix}v(i,j-1)\\h(i,j-1)\end{bmatrix}+\begin{bmatrix}\bar{\tau}_1(i-1,j)\\\bar{\tau}_2(i,j-1)\end{bmatrix} \tag{11a}$$

$$\Delta y(i,j)=\bar{c}\begin{bmatrix}\Delta v(i,j)\\\Delta h(i,j)\end{bmatrix}+\Delta c\begin{bmatrix}v(i,j)\\h(i,j)\end{bmatrix}+\gamma(i,j)+\delta(i,j). \tag{11b}$$

If we assume that $\{\bar{A},\bar{b},\bar{c},\bar{d}\}=\{A,b,c,d,\}$ and $\bar{u}(i,j)=u(i,j)$, that is, if the errors in $y(i,j)$ are due to product quantization only, then model (11) becomes

$$\begin{bmatrix}\Delta v(i,j)\\\Delta h(i,j)\end{bmatrix}=A_{10}\begin{bmatrix}\Delta v(i-1,j)\\\Delta h(i-1,j)\end{bmatrix}+A_{01}\begin{bmatrix}\Delta v(i,j-1)\\\Delta h(i,j-1)\end{bmatrix}+\begin{bmatrix}\tau_1(i-1,j)\\\tau_2(i,j-1)\end{bmatrix} \tag{12a}$$

$$\Delta y(i,j)=c\begin{bmatrix}\Delta v(i,j)\\\Delta h(i,j)\end{bmatrix}+\gamma(i,j)+\delta(i,j) \tag{12b}$$

where

$$\begin{bmatrix}\tau_1(i,j)\\\tau_2(i,j)\end{bmatrix}=\alpha(i,j)+\beta(i,j).$$

B. Output-Noise Power

For any fixed $(i,j)\geq(0,0)$, noise model (12) gives

$$\Delta y(i,j)=c\sum\sum_{(0,0)<(l,k)\leq(i,j)}\left(A_{l-1,k}\begin{bmatrix}\tau_1(i-l-1,j-k)\\0\end{bmatrix}+A_{l,k-1}\begin{bmatrix}0\\\tau_2(i-l,j-k-1)\end{bmatrix}\right)+\gamma(i,j)+\delta(i,j). \tag{13}$$

Let us assume statistical independence among noise sources and among distinct samples for each noise source [14, p. 288]. If quantization of products is carried out by rounding, then the variance of each noise source is $E_0^2/12$, where E_0 is the quantization step size. By (13), the expected square error is

$$E(\Delta y^2(i,j)) = \frac{E_0^2}{12} c \left\{ \sum\sum_{(0,0)<(l,k)\leq(i,j)} A_{l-1,k} \begin{bmatrix} Q_1 & 0 \\ 0 & 0 \end{bmatrix} A_{l-1,k}^t + A_{l,k-1} \begin{bmatrix} 0 & 0 \\ 0 & Q_2 \end{bmatrix} A_{l,k-1}^t \right\} c^t + \frac{E_0^2}{12}(\mu+\nu) \quad (14)$$

where

$$\frac{E_0^2}{12} Q_l = E[\tau_l \tau_l^t] = E[\alpha_l \alpha_l^t] + E[\beta_l \beta_l^t], \qquad l = 1,2 \quad (15)$$

are positive definite diagonal matrices of dimensions m and n, respectively, and μ and ν are the numbers of constants in c and d, respectively, which are neither zero nor one.

Concerning the convergence of series (14) as $i \to \infty$ and $j \to \infty$, (6) gives

$$\left| c \left(\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A_{ij} A_{ij}^t \right) c^t \right| = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \|cA_{ij}\|^2 \leq \|c\|^2 \cdot \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \|A_{ij}\|^2 < \infty$$

where $\|\cdot\|$ denotes the Euclidean norm, and thus

$$\left| c \left(\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A_{ij} \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix} A_{ij}^t \right) c^t \right| \leq q^* \left| c \left(\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A_{ij} A_{ij}^t \right) c^t \right| < \infty$$

where q^* represents the largest element in the diagonal of $Q_1 \oplus Q_2$. Therefore, viewing the output noise $\Delta y(i,j)$ as a stochastic process, its variance can be calculated as

$$\begin{aligned} E(\Delta y^2) &= \lim_{\substack{i\to\infty \\ j\to\infty}} E\left[\frac{1}{(i+1)(j+1)-1} \sum\sum_{(0,0)\leq(l,k)<(i,j)} \Delta y^2(l,k) \right] \\ &= \frac{E_0^2}{12} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} cA_{ij} Q A_{ij}^t c^t + \frac{E_0^2}{12}(\mu+\nu) \\ &= \frac{E_0^2}{12} \operatorname{tr}[QW] + \frac{E_0^2}{12}(\mu+\nu) \end{aligned} \quad (16)$$

where

$$Q = Q_1 \oplus Q_2 \quad (17)$$

$$W = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A_{ij}^t c^t c A_{ij} \quad (18)$$

and $\operatorname{tr}[\cdot]$ denotes the trace of a matrix. The computation of matrix W can be carried out as shown in Appendix A. It is important to note that W in (18) is positive-definite if (A, c) is a 2-D locally observable pair, namely

$$\operatorname{rank}\left[c^t \quad (cA_{01})^t \cdots (cA_{0n})^t \quad (cA_{10})^t \cdots (cA_{1n})^t \cdots (cA_{m,n-1})^t \right]^t = m+n. \quad (19)$$

The proof can be found in Appendix B.

An important feature of expression (16) is its dependence on the coordinate system. Consequently, if transformation (7) is applied, the output-noise power of realization (8) is given by

$$E(\Delta \tilde{y}^2) = \frac{E_0^2}{12} \operatorname{tr}[\tilde{Q}\tilde{W}] + \frac{E_0^2}{12}(\mu+\nu) \quad (20)$$

where

$$\tilde{W} = T^t W T = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \tilde{A}_{ij}^t \tilde{c}^t \tilde{c} \tilde{A}_{ij} \quad (21)$$

and $\tilde{Q} = \tilde{Q}_1 \oplus \tilde{Q}_2$. In the next section, an optimal similarity transformation T will be found, which minimizes $E(\Delta \tilde{y}^2)$ under a set of dynamic range constraints.

IV. Roundoff Noise Minimization

A. Scaling Condition

To prevent overflow in the digital filter, signal scaling will be applied based on the l_2 norm, which accordingly gives a set of dynamic range constraints on the local state variables.

Let $f(i,j)$ be the state for an impulse input. By (4)

$$f(i,j) = A_{i-1,j} \begin{bmatrix} b_1 \\ 0 \end{bmatrix} + A_{i,j-1} \begin{bmatrix} 0 \\ b_2 \end{bmatrix}$$

and, consequently, the local state at (i,j) due to an input sequence $\{u(l,k), (0,0) \leq (l,k) < (i,j)\}$ is obtained as

$$\begin{bmatrix} v(i,j) \\ h(i,j) \end{bmatrix} = \sum\sum_{(0,0)<(l,k)\leq(i,j)} f(l,k) u(i-l, j-k). \quad (22)$$

If e_p is the pth column of the identity matrix of dimension $(m+n)$, then the pth component of the local state in (22) can be estimated as

$$\begin{aligned} \left| e_p^t \begin{bmatrix} v(i,j) \\ h(i,j) \end{bmatrix} \right|^2 &= \left[\sum\sum_{(l,k)} e_p^t f(l,k) u(i-l, j-k) \right]^2 \\ &\leq e_p^t \left(\sum\sum_{(l,k)} f(l,k) f^t(l,k) \right) \cdot e_p \sum\sum_{(l,k)} u^2(i-l, j-k) \\ &\leq e_p^t K e_p \|u\|^2 \end{aligned} \quad (23)$$

where

$$K = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f(i,j) f^t(i,j). \quad (24)$$

Notice that $e_p^t K e_p$ is the pth diagonal element of K and, therefore, if all the diagonal elements of K are equal to one, (24) implies that the amplitude of each state component is no more than $\|u\|$. The dynamic range constraints on the state variables are, therefore, given by

$$K = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & * & \\ & & 1 & & & \\ & & & 1 & & \\ & * & & & \ddots & \\ & & & & & 1 \end{bmatrix}. \tag{25}$$

Further, once a similarity transformation T is used to reduce the output-noise power, the above dynamic range constraints should also be satisfied by the transformed realization. Since

$$\tilde{K} \equiv \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \tilde{f}(i,j)\tilde{f}^t(i,j) = T^{-1}KT^{-t} \tag{26}$$

the dynamic range constraints on the new state variables $\tilde{v}(i,j)$ and $\tilde{h}(i,j)$ are given by

$$\tilde{K} = T^{-1}KT^{-t} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & * & \\ & & 1 & & & \\ & & & 1 & & \\ & * & & & \ddots & \\ & & & & & 1 \end{bmatrix}. \tag{27}$$

The computation issues of matrix K are detailed in Appendix A. Note that K is positive-definite if (A, b) is a locally reachable pair, i.e.,

$$\operatorname{rank}[f(1,0) \quad f(0,1) \cdots f(m,n)] = m+n. \tag{28}$$

The proof is given in Appendix B.

B. The 2-D Second-Order Modes

By analogy with the 1-D case, (26) and (21) given $\tilde{K}\tilde{W} = T^{-1}KWT$ and, therefore, the eigenvalues of KW are invariant under a 2-D similarity transformation [12]. It turns out, however, that these invariants are not significant in the present optimization problem. This is because the related optimization problem using these invariants would lead to a non-block-diagonal similarity transformation which is not acceptable. An alternative approach is, therefore, explored. If we denote

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{12}^t & K_{22} \end{bmatrix}, \quad W = \begin{bmatrix} W_{11} & W_{12} \\ W_{12}^t & W_{22} \end{bmatrix}$$
$$\tilde{K} = \begin{bmatrix} \tilde{K}_{11} & \tilde{K}_{12} \\ \tilde{K}_{12}^t & \tilde{K}_{22} \end{bmatrix}, \quad \tilde{W} = \begin{bmatrix} \tilde{W}_{11} & \tilde{W}_{12} \\ \tilde{W}_{12}^t & \tilde{W}_{22} \end{bmatrix} \tag{29}$$

then (26) and (21) yield

$$\tilde{K}_{11}\tilde{W}_{11} = T_1^{-1}K_{11}W_{11}T_1 \tag{30}$$

and

$$\tilde{K}_{22}\tilde{W}_{22} = T_2^{-1}K_{22}W_{22}T_2. \tag{31}$$

In other words, the eigenvalues of $\Phi \equiv K_{11}W_{11}$ and $\Psi \equiv K_{22}W_{22}$, denoted by $\phi \equiv \{\phi_i,\ 1 \leqslant i \leqslant m\}$ and $\psi \equiv \{\psi_j,\ 1 \leqslant j \leqslant n\}$, are invariant under a 2-D similarity transformation. In the rest of this paper, we call sets $\{\phi_i,\ 1 \leqslant i \leqslant m\}$ and $\{\psi_j,\ 1 \leqslant j \leqslant n\}$ the 2-D second-order modes of the filter. As will be shown below, these invariants are crucial in obtaining an explicit solution to a simplified optimization problem.

C. Simplified Optimization Problem

From now on we assume that matrix $\tilde{Q}$ in (20) is equal to $(m+n+1)I$ and is independent of the similarity transformation used. In other words, it is assumed that the computation of each component of the new state variables $\tilde{v}(i+1, j)$ and $\tilde{h}(i, j+1)$ always involves $(m+n+1)$ multiplications, i.e., both matrix $\tilde{A}$ and vector $\tilde{b}$ have neither zero or one entries. This assumption is obviously pessimistic. It, however, greatly simplifies our problem, and numerical experience for the case of 1-D filters [1] has shown that the application of the noise model with this assumption is valid. Expression (20) now becomes

$$E(\Delta\tilde{y}^2) = \frac{(m+n+1)E_0^2}{12}\tilde{G} + \frac{E_0^2}{12}(\mu+\nu) \tag{32}$$

where

$$\tilde{G} = \operatorname{tr}(T^tWT) \tag{33}$$

referred to as the unit noise of realization $(T^{-1}AT, T^{-1}b, cT, d)$ is the only transformation-dependent quantity in (32). Our objective can be formulated as the following optimization problem: given a reachable, observable, and stable LSS realization $\{A, b, c, d\}$, find a 2-D similarity transformation $T = T_1 \oplus T_2$ such that realization $\{\tilde{A}, \tilde{b}, \tilde{c}, \tilde{d}\} \equiv \{T^{-1}AT, T^{-1}b, cT, d\}$ minimizes unit noise $\tilde{G}$ subject to constraint (27).

Since matrices $K_{11} \in R^{m\times m}$ and $K_{22} \in R^{n\times n}$ in K (see (29)) are both positive-definite, one can find a nonsingular, block-diagonal matrix $P = P_1 \oplus P_2$ such that

$$\bar{K} \equiv P^{-1}KP^{-t} = \begin{bmatrix} I_m & \Gamma \\ \Gamma^t & I_n \end{bmatrix} \tag{34}$$

where Γ is a certain $m \times n$ real matrix. Now if

$$\tilde{T} = P^{-1}T \tag{35}$$

$\tilde{T}$ is also block-diagonal and can be used as a 2-D similarity transformation. Substituting (35) in (33) and (27) gives

$$\tilde{G} = \operatorname{tr}(\tilde{T}^t\bar{W}\tilde{T}) \tag{36}$$

and

$$\tilde{K} = \tilde{T}^{-1}\begin{bmatrix} I_m & \Gamma \\ \Gamma^t & I_n \end{bmatrix}\tilde{T}^{-t} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & * & \\ & & 1 & & & \\ & & & 1 & & \\ & * & & & \ddots & \\ & & & & & 1 \end{bmatrix} \tag{37}$$

respectively, where

$$\overline{W} = P^t W P \equiv \begin{bmatrix} \overline{W}_{11} & \overline{W}_{12} \\ \overline{W}_{12}^t & \overline{W}_{22} \end{bmatrix}. \tag{38}$$

A further simplification of the problem can now be made through the use of the singular value decomposition (SVD) of matrix $\tilde{T}$ [15, ch. 6]. We can express $\tilde{T}$ as

$$\tilde{T} = \begin{bmatrix} \tilde{T}_1 & 0 \\ 0 & \tilde{T}_2 \end{bmatrix} = \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}^t \equiv R \Lambda S^t \tag{39}$$

where $R_1 \in R^{m\times m}$, $R_2 \in R^{n\times n}$, $S_1 \in R^{m\times m}$, and $S_2 \in R^{n\times n}$ are orthogonal matrices; $\Lambda_1 = \text{diag}\{\lambda_{11} \cdots \lambda_{1m}\}$, $\Lambda_2 = \text{diag}\{\lambda_{21} \cdots \lambda_{2n}\}$; $\lambda_{11} \geqslant \cdots \geqslant \lambda_{1m} > 0$, and $\lambda_{21} \geqslant \cdots \geqslant \lambda_{2n} > 0$ are the singular values of matrix $\tilde{T}$. By substituting (39) in (36) and (37), we have

$$\tilde{G} = \text{tr}(S\Lambda R^t \overline{W} R \Lambda S^t) = \text{tr}(\Lambda^2 R^t \overline{W} R) = \sum_{i=1}^{m} \lambda_{1i}^2 u_i^2 + \sum_{i=1}^{n} \lambda_{2i}^2 u_{m+i}^2 \tag{40}$$

where u_i^2 is the ith diagonal element of $R^t \overline{W} R$, and

$$\tilde{K} = S\Lambda^{-1} R^t \begin{bmatrix} I_m & \Gamma \\ \Gamma^t & I_n \end{bmatrix} R \Lambda^{-1} S^t = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & * & \\ & & 1 & & & \\ & & & 1 & & \\ & * & & & \ddots & \\ & & & & & 1 \end{bmatrix}.$$

This yields the simpler constraint

$$\tilde{K} = S \begin{bmatrix} \Lambda_1^{-2} & * \\ * & \Lambda_2^{-2} \end{bmatrix} S^t = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & * & \\ & & 1 & & & \\ & & & 1 & & \\ & * & & & \ddots & \\ & & & & & 1 \end{bmatrix}. \tag{41}$$

Since any 2-D similarity transformation $\tilde{T}$ can be obtained by properly choosing block-orthogonal matrices R, S, and diagonal matrix Λ in (39), the above optimization problem is now reduced to the following one: given as LSS realization $\{A, b, c, d\}$ as is (1), choose two block-orthogonal matrices R, S and a set of positive real numbers $\{\lambda_{1i}, \lambda_{2j}, 1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n\}$ such that $\tilde{G}$ in (40) is minimized subject to constraint (41).

D. A Solution of the Simplified Problem

A remarkable feature of the problem of minimizing $\tilde{G}$ in (40) subject to constraint (41) is that the free parameter R appears only in $\tilde{G}$ and free parameter S appears only in $\tilde{K}$. Therefore, one can choose R independently of S. Making use of this advantage, we conclude that constraint (41) can be characterized by two simple equalities on the λ's. This result is stated as a lemma below.

Lemma 1: There exist a block-orthogonal matrix S and a diagonal matrix Λ such that constraint (41) is satisfied if, and only if

$$\sum_{i=1}^{m} \frac{1}{\lambda_{1i}^2} = m \quad \text{and} \quad \sum_{i=1}^{n} \frac{1}{\lambda_{2i}^2} = n. \tag{42}$$

Proof: Substituting $S = S_1 \oplus S_2$ in (41), the constraint condition becomes

$$\tilde{K} = \begin{bmatrix} S_1 \Lambda_1^{-2} S_1^t & * \\ * & S_2 \Lambda_2^{-2} S_2^t \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & * & \\ & & 1 & & & \\ & & & 1 & & \\ & * & & & \ddots & \\ & & & & & 1 \end{bmatrix}. \tag{43}$$

This implies that (42) holds since S_1 and S_2 are both orthogonal. Conversely, by Lemma 3 of [3], conditions (42) imply that there exist two orthogonal matrices $S_1 \in R^{m\times m}$ and $S_2 \in R^{n\times n}$ such that

$$S_1 \Lambda_1^{-2} S_1^t = \begin{bmatrix} 1 & & * \\ & \ddots & \\ * & & 1 \end{bmatrix}_{m\times m}$$

and

$$S_2 \Lambda_2^{-2} S_2^t = \begin{bmatrix} 1 & & * \\ & \ddots & \\ * & & 1 \end{bmatrix}_{n\times n}.$$

These equalities give (41) where $S = S_1 \oplus S_2$. □

The following lemma gives two sets $\{\lambda_{1i}^*,\ 1 \leqslant i \leqslant m\}, \{\lambda_{2i}^*,\ 1 \leqslant i \leqslant n\}$ in terms of u_i $(1 \leqslant i \leqslant m+n)$, which minimize (40) subject to condition (42).

Lemma 2: For any positive real numbers $\{u_i,\ 1 \leqslant i \leqslant m+n\}$, the minimization problem

$$\min_{\lambda} \left[\sum_{i=1}^{m} \lambda_{1i}^2 u_i^2 + \sum_{i=1}^{n} \lambda_{2i}^2 u_{m+i}^2 \right] \tag{44}$$

subject to (42) can be solved by selecting

$$\lambda_{1j}^* = \left[\frac{\frac{1}{m} \sum_{i=1}^{m} u_i}{u_j} \right]^{1/2}, \quad 1 \leqslant j \leqslant m \tag{45}$$

$$\lambda_{2j}^* = \left[\frac{\frac{1}{n} \sum_{i=1}^{n} u_{m+i}}{u_{m+j}} \right]^{1/2}, \quad 1 \leqslant j \leqslant n \tag{46}$$

and then solving the minimization problem

$$\min_{\lambda} \tilde{G} = \min_{\lambda} \left[\sum_{i=1}^{m} \lambda_{1i}^2 u_i^2 + \sum_{i=1}^{n} \lambda_{2i}^2 u_{m+i}^2 \right] = \frac{\left(\sum_{i=1}^{m} u_i \right)^2}{m} + \frac{\left(\sum_{i=1}^{n} u_{m+i} \right)^2}{n}. \quad (47)$$

Proof: Define

$$F = \sum_{i=1}^{m} \lambda_{1i}^2 u_i^2 + \sum_{i=1}^{n} \lambda_{2i}^2 u_{m+i}^2 + \eta_1 \left(\sum_{i=1}^{m} \frac{1}{\lambda_{1i}^2} - m \right) + \eta_2 \left(\sum_{i=1}^{n} \frac{1}{\lambda_{2i}^2} - n \right)$$

where η_1 and η_2 are the Lagrange multipliers. Routine calculus manipulation leads to solutions (45) and (46). Substituting (45) and (46) in $\tilde{G}$ given by (40) yields (47). □

The above lemma reduces the main optimization problem to

$$\min_{u} \left[\frac{\left(\sum_{i=1}^{m} u_i \right)^2}{m} + \frac{\left(\sum_{i=1}^{n} u_{m+i} \right)^2}{n} \right] \quad (48)$$

where u_i is the square root of the ith diagonal element of $R^t \overline{W} R$. By the invariance of the 2-D second-order modes under a similarity transformation, the optimum values of the u's which minimize the expression in (48) are obtained as follows.

Lemma 3: The minimum value in (48) can be achieved if, and only if, the block-orthogonal matrix R is chosen such that $R_1^t \overline{W}_{11} R_1$ and $R_2^t \overline{W}_{22} R_2$ are both diagonal. Moreover

$$\min_{u} \left[\frac{\left(\sum_{i=1}^{m} u_i \right)^2}{m} + \frac{\left(\sum_{i=1}^{n} u_{m+i} \right)^2}{n} \right] = \frac{\left(\sum_{i=1}^{m} \phi_i \right)^2}{m} + \frac{\left(\sum_{i=1}^{n} \psi_i \right)^2}{n} \quad (49)$$

where the ϕ's and ψ's are the 2-D second-order modes of the filter considered.

Proof: Notice first that $\{u_i^2,\ 1 \leqslant i \leqslant m\}$ and $\{u_{m+i}^2,\ 1 \leqslant i \leqslant n\}$ are also the diagonal elements of matrices $R_1 \overline{W}_{11} R_1^t$ and $R_2 \overline{W}_{22} R_2^t$, respectively. Thus, by Lemma 2 of [3], the minimum values of

$$\sum_{i=1}^{m} u_i \text{ and } \sum_{i=1}^{n} u_{m+i}$$

will be achieved if, and only if, $R_1 \overline{W}_{11} R_1^t$ and $R_2 \overline{W}_{22} R_2^t$ are diagonal. Further, by (34) and (38) the invariance of the 2-D second-order modes implies that the eigenvalues of matrices $R_1 \overline{W}_{11} R_1^t$ and $R_2 \overline{W}_{22} R_2^t$ are $\phi = \{\phi_i,\ 1 \leqslant i \leqslant m\}$ and $\psi = \{\psi_i,\ 1 \leqslant i \leqslant n\}$. Therefore, (49) holds. □

We are now in a position to summarize the main results of this section.

Theorem 1: Given a reachable and observable realization $\{A, b, c, d\}$ satisfying stability condition (2), there exists a 2-D similarity transformation $T = T_1 \oplus T_2$, such that realization $\{T^{-1}AT, T^{-1}b, cT, d\}$ minimizes the output-noise power $E(\Delta \tilde{y}^2)$ in (32) subject to dynamic range constraint (27). The computation of the desired transformation matrix T can be carried out by the following procedure.

1) Compute positive-definite matrices K and W via (24) and (18), respectively.

2) Find matrix $P = P_1 \oplus P_2$ such that (34) holds.

3) Compute $\overline{W}$ via (38).

4) Find block-orthogonal matrix $R = R_1 \oplus R_2$ such that $R_1 \overline{W}_{11} R_1^t$ and $R_2 \overline{W}_{22} R_2^t$ are diagonal, i.e., $R_1 \overline{W}_{11} R_1^t = \operatorname{diag}\{u_1 \cdots u_m\}$, $R_2 \overline{W}_{22} R_2^t = \operatorname{diag}\{u_{m+1} \cdots u_{m+n}\}$.

5) Compute

$$\Lambda^* = \begin{bmatrix} \Lambda_1^* & 0 \\ 0 & \Lambda_2^* \end{bmatrix}$$

where $\Lambda_1^* = \operatorname{diag}\{\lambda_{11}^* \cdots \lambda_{1m}^*\}$ and $\Lambda_2^* = \operatorname{diag}\{\lambda_{21}^* \cdots \lambda_{2n}^*\}$ are given by (45) and (46), respectively.

6) Find block-orthogonal matrix $S = S_1 \oplus S_2$ such that

$$S_1 \Lambda_1^{*-2} S_1^t = \begin{bmatrix} 1 & & * \\ & \ddots & \\ * & & 1 \end{bmatrix}_{m \times m}$$

and

$$S_2 \Lambda_2^{*-2} S_2^t = \begin{bmatrix} 1 & & * \\ & \ddots & \\ * & & 1 \end{bmatrix}_{n \times n}$$

by the algorithm given in the appendix of [3].

7) Form

$$T = PR\Lambda^* S^t.$$

E. Other Issues

We conclude this section with a brief discussion on several issues relevant to the main results presented.

1) As for the 1-D case [1], the signal scaling conditions based on the l_2 norm can be expressed as

$$\delta^2 \|f_k\|^2 = \left(E_0 2^{l-1} \right)^2 \qquad k = 1, \cdots, m+n \quad (50)$$

where l is the wordlength, δ is a parameter which determines the probability of overflow, and $\|f_k\|$ denotes the l_2 norm of doubly-indexed sequence composed of the kth component of $f(i, j)$, i.e.,

$$\|f_k\|^2 = e_k^t \left(\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f(i, j) f^t(i, j) \right) e_k = K_{kk}, \qquad k = 1, \cdots, m+n.$$

Thus, scaling condition (50) becomes

$$K_{kk} = \left(\frac{E_0 2^{l-1}}{\delta} \right)^2, \qquad k = 1, \cdots, m+n. \quad (51)$$

We observe that condition (51) is the same as condition (25) up to a constant factor and, therefore, no essential

differences will occur when scaling condition (25) is replaced by (51).

2) Through the same argument as in [3], we can establish a lower bound for $\tilde{G}$ in (33) as

$$\tilde{G} \geq (m+n)(\det KW)^{(1/m+n)}. \tag{52}$$

Note that $\det(KW)$ is invariant under similarity transformation so that the lower bound given in (52) is coordinate independent.

3) From Theorem 1, it is observed that once matrices K and W are computed, the desirable transformation $T = T_1 \oplus T_2$ can be obtained by solving *separately* two 1-D minimization problems as follows.

(1) Find a nonsingular transformation T_1 of dimension m such that

$$\tilde{G}_1 = \operatorname{tr}(T_1^t W_{11} T_1) \tag{53}$$

is minimized subject to

$$T_1^{-1} K_{11} T_1^{-t} = \begin{bmatrix} 1 & & * \\ & \ddots & \\ * & & 1 \end{bmatrix} \tag{54}$$

where W_{11} and K_{11} are the positive-definite matrices of dimension m given by (29).

(2) Find a nonsingular transformation T_2 of dimension n such that

$$\tilde{G}_2 = \operatorname{tr}(T_2^t W_{22} T_2) \tag{55}$$

is minimized subject to

$$T_2^{-1} K_{22} T_2^{-t} = \begin{bmatrix} 1 & & * \\ & \ddots & \\ * & & 1 \end{bmatrix} \tag{56}$$

where W_{22} and K_{22} are the positive-definite matrices of dimension n given by (29).

In other words, upon the availability of matrices W and K defined in (18) and (24), the well-developed synthesis approaches in [1] and [3] can be used to obtain a state-space realization for a 2-D digital filter with minimum roundoff noise.

V. Examples

Example 1

Let us consider an arbitrary 2-D filter of order (1,1) represented by state-space model (1), where

$$A = \left[\begin{array}{c|c} a_1 & a_2 \\ \hline a_3 & a_4 \end{array}\right], \quad b = \begin{bmatrix} b_1 \\ \hline b_2 \end{bmatrix}, \quad c = [c_1 \mid c_2].$$

We assume that the filter is BIBO stable so that K and W in (24) and (18) are well defined. We also assume that (A, c) is a 2-D locally observable pair and that (A, b) is a 2-D locally reachable pair, i.e.,

$$\operatorname{rank}\begin{bmatrix} c \\ cA_{01} \\ cA_{10} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} c_1 & c_2 \\ c_2 a_3 & c_2 a_4 \\ c_1 a_1 & c_1 a_2 \end{bmatrix} = 2$$

and

$$\operatorname{rank}\begin{bmatrix} b_1 & 0 & a_2 b_2 \\ 0 & b_2 & a_3 b_1 \end{bmatrix} = 2.$$

These assumptions guarantee that the resulting matrices K and W are positive definite. Let us suppose that K and W have been computed through (18) and (24), and are given by

$$K = \begin{bmatrix} k_1 & k_2 \\ k_2 & k_4 \end{bmatrix} \text{ and } W = \begin{bmatrix} w_1 & w_2 \\ w_2 & w_4 \end{bmatrix}.$$

It is easy to verify that

$$P = \begin{bmatrix} \sqrt{k_1} & 0 \\ 0 & \sqrt{k_4} \end{bmatrix} \text{ and } R = \Lambda^* = S = I.$$

Thus, $T = P$ and the optimal realization is given by

$$\tilde{A} = TAT^{-1} = \left[\begin{array}{c|c} a_1 & a_2 \sqrt{\dfrac{k_1}{k_4}} \\ \hline a_3 \sqrt{\dfrac{k_4}{k_1}} & a_4 \end{array}\right], \quad \tilde{b} = Tb = \begin{bmatrix} b_1 \sqrt{k_1} \\ \hline b_2 \sqrt{k_4} \end{bmatrix}$$

$$\tilde{c} = cT^{-1} = \begin{bmatrix} \dfrac{c_1}{\sqrt{k_1}} & \dfrac{c_2}{\sqrt{k_4}} \end{bmatrix}, \text{ and } \tilde{d} = d.$$

Example 2

As a numerical example, we consider a stable state-space digital filter of order (2,2) modelled by (1) where

$$A = \left[\begin{array}{cc|cc} 1.88899 & -0.91219 & -1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ \hline 0.02771 & -0.0258 & 1.88899 & 1.0 \\ -0.0258 & 0.02431 & -0.91219 & 0.0 \end{array}\right]$$

$$b^T = [0.219089 \quad 0.0 \mid -0.028889 \quad 0.091219], \text{ and}$$

$$c = [0.28889 \quad -0.091219 \mid -0.219089 \quad 0.0].$$

Since the system is stable, we may use finite sums

$$\sum_{i=0}^{M} \sum_{j=0}^{N} A_{ij}^t c^t c A_{ij} \quad \text{and} \quad \sum_{i=0}^{M} \sum_{j=0}^{N} f(i,j) f^t(i,j)$$

to approximate W and K, respectively. Taking $M = N = 240$, numerical computation gives

$$W = \left[\begin{array}{cc|cc} 1.133630 & -1.032898 & 0.977893 & 1.774435 \\ -1.032898 & 0.965161 & -0.941089 & -1.672273 \\ \hline 0.977893 & -0.941089 & 87.124460 & 85.257248 \\ 1.774435 & -1.672273 & 85.257248 & 87.172446 \end{array}\right]$$

and

$$K = \left[\begin{array}{cc|cc} 87.124446 & 85.257248 & 1.639820 & -1.539081 \\ 85.257248 & 87.172445 & 1.321218 & -1.233185 \\ \hline 1.639820 & 1.321218 & 1.133630 & -1.032898 \\ -1.539081 & -1.233185 & -1.032898 & 0.965161 \end{array}\right].$$

The unit noise of this filter after scaling is the sum of products of corresponding diagonal entries in W and K and is given by

$$G_0 = \sum_{i=1}^{4} w_{ii} k_{ii} = 365.804889.$$

Following the procedure of Theorem 1, the desirable similarity transformation is calculated as

$$T=\left[\begin{array}{cc|cc} -3.309820 & 10.442754 & \multicolumn{2}{c}{0} \\ -5.328425 & 10.609476 & & \\ \hline \multicolumn{2}{c|}{0} & 0.915337 & 0.261280 \\ & & -0.949348 & -0.065936 \end{array}\right].$$

The characterization of the optimal state-space structure of this filter can now be obtained as

$$\tilde{A}=T^{-1}AT=\left[\begin{array}{cc|cc} 0.964470 & -0.119001 & -0.473073 & -0.135037 \\ 0.172420 & 0.924519 & -0.237593 & -0.067820 \\ \hline 0.045371 & 0.010522 & 0.888409 & 0.181558 \\ 0.016180 & 0.023012 & -0.128139 & 1.000580 \end{array}\right]$$

$$\tilde{b}=T^{-1}b=\left[\begin{array}{c} 0.113231 \\ 0.056868 \\ \hline -0.116834 \\ 0.298736 \end{array}\right]$$

and

$$\tilde{c}=cT=[0.390436 \quad -0.666105 \mid -0.200540 \quad -0.057243].$$

The corresponding matrices $\tilde{W}$ and $\tilde{K}$ are

$$\tilde{W}=\begin{bmatrix} 3.389308 & 0.000009 & -1.256271 & 0.264211 \\ 0.000009 & 3.389308 & -0.539958 & 0.007450 \\ -1.256271 & -0.539958 & 3.389006 & -0.000005 \\ 0.264211 & 0.007450 & -0.000005 & 3.389074 \end{bmatrix}$$

and

$$\tilde{K}=\begin{bmatrix} 0.999957 & 0.475650 & 0.172402 & 0.067288 \\ 0.475650 & 0.999974 & 0.204635 & 0.096822 \\ 0.172402 & 0.204635 & 0.999968 & 0.475657 \\ 0.067288 & 0.096822 & 0.475657 & 0.999982 \end{bmatrix}.$$

These give the unit noise of realization $(\tilde{A}, \tilde{b}, \tilde{c})$ as

$$\tilde{G}=\sum_{i=1}^{4}\tilde{w}_{ii}\tilde{k}_{ii}=13.555811$$

which is quite close to the lower bound of $\tilde{G}$ given in (52):

$$(m+n)(\det KW)^{1/(m+n)}=11.949383.$$

VI. Conclusions

Based on a roundoff-noise analysis, a general synthesis procedure has been presented which leads to an optimal local state-space 2-D filter that minimizes the output-noise power due to roundoff subject to a signal scaling condition.

Two matrices W and K derived from the calculation of the output-noise power and the signal scaling condition, respectively, provide two sets of 2-D second-order modes of the filter. It has been demonstrated that these modes play a crucial role in the minimization of roundoff noise. The general synthesis procedure has been illustrated by two examples.

It should be pointed out that the approach presented in this paper can be extended to the N-dimensional case where $N>2$ in a straightforward manner provided that the multidimensional LSS model proposed in [16] is used.

Appendix A

In this appendix, we outline two approaches for computing matrices W and K defined by (18) and (24), respectively.

A. Truncation Method

A straightforward way of computing W and K is to use the truncated double sums

$$W\approx\sum_{i=0}^{M}\sum_{j=0}^{N}A_{ij}^{t}c^{t}cA_{ij} \quad \text{and} \quad K\approx\sum_{i=0}^{M}\sum_{j=0}^{N}f(i,j)f^{t}(i,j) \tag{A1}$$

where M and N are positive integers. Through the use of recursive formula (3), the above finite sums can readily be programmed. In general, when M and N are taken to be large enough, the sums in (A1) will represent good approximations of W and K. However, experience has shown that even for small dimensions m and n, this approach needs a considerable amount of computation time.

B. Evaluation of K_{ii} and W_{ii} $(i=1,2)$ via a Lyapunov Approach

Define

$$F(z,w)=[I(z,w)-A]^{-1}b$$
$$G(z,w)=c[I(z,w)-A]^{-1}$$

and

$$I(z,w)=I_m z\oplus I_n w$$

where $\oplus$ denotes direct sum, and let

$$\hat{K}=\frac{1}{(2\pi j)^2}\oint_{|z|=1}\oint_{|w|=1}F(z,w)F^{*}(z,w)\frac{dw}{w}\frac{dz}{z} \tag{A2}$$

and

$$\hat{W}=\frac{1}{(2\pi j)^2}\oint_{|z|=1}\oint_{|w|=1}G^{*}(z,w)G(z,w)\frac{dw}{w}\frac{dz}{z} \tag{A3}$$

where * represents conjugate transpose. The use of the residue theorem leads to

$$K_{11}=[I_m \quad 0]\hat{K}\begin{bmatrix} I_m \\ 0 \end{bmatrix},\quad K_{22}=[0 \quad I_n]\hat{K}\begin{bmatrix} 0 \\ I_n \end{bmatrix}$$
$$W_{11}=[I_m \quad 0]\hat{W}\begin{bmatrix} I_m \\ 0 \end{bmatrix},\quad W_{22}=[0 \quad I_n]\hat{W}\begin{bmatrix} 0 \\ I_n \end{bmatrix}.$$

Thus

$$K_{11}=\frac{1}{(2\pi j)^2}\oint_{|z|=1}\oint_{|w|=1}[I_m \quad 0]F(z,w)\cdot([I_m \quad 0]F(z,w))^{*}\frac{dw}{w}\frac{dz}{z} \tag{A4}$$

where $[I_m \quad 0]F(z,w)$ can be written as

$$[I_m \quad 0]F(z,w)=(zI-\tilde{A}(w))^{-1}\tilde{b}(w)$$

with

$$\tilde{A}_1(w)=A_1+A_2(wI-A_4)^{-1}A_3$$

and

$$\tilde{b}_1(w)=b_1+A_2(wI-A_4)^{-1}b_2.$$

Hence, (A4) becomes

$$K_{11} = \frac{1}{2\pi j}\oint_{|w|=1}\frac{dw}{w}\frac{1}{2\pi j}\oint_{|z|=1}(zI-\tilde{A}_1(w))^{-1}\tilde{b}_1(w)\cdot\tilde{b}_1^*(w)(z^*I-\tilde{A}_1^*(w))^{-1}\frac{dz}{z}$$
$$= \frac{1}{2\pi j}\oint_{|w|=1}\tilde{K}_1(w)\frac{dw}{w} \qquad \text{(A5)}$$

where $\tilde{K}_1(w)$ is the positive-definite Hermitian solution of the Lyapunov equation

$$\tilde{A}_1(w)\tilde{K}_1(w)\tilde{A}_1^*(w)-\tilde{K}_1(w) = -\tilde{b}_1(w)\tilde{b}_1^*(w). \qquad \text{(A6)}$$

The integral in (A5) can be computed through the use of the residue theorem.

K_{22}, W_{11}, and W_{22} can be evaluated in a similar manner.

Appendix B

Theorem B.1: If (A, c) is a 2-D locally observable pair, i.e., condition (19) holds, then matrix W defined by (18) is positive-definite.

Proof: If x is an $(n+m)$-dimensional vector such that $x^tWx = 0$, then

$$0 = x^tWx = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}x^tA_{ij}^tc^tcA_{ij}x = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\|cA_{ij}x\|^2.$$

Hence

$$cA_{ij}x = 0 \qquad \text{for all } i \geq 0,\ j \geq 0$$

which implies

$$x^t\left[c^t \quad (cA_{01})^t\cdots(cA_{0n})^t \quad (cA_{10})^t\cdots(cA_{1n})^t \cdots (cA_{m,n-1})^t\right]^t = 0. \qquad \text{(B1)}$$

By condition (19), equation (B1) implies that $x = 0$ and, therefore, W is positive-definite. □

Theorem B.2: If (A, b) is a 1-D reachable pair, i.e., condition (28) holds, then matrix K defined by (24) is positive-definite.

Proof: If x is an $(m+n)$-dimensional vector such that $x^tKx = 0$, then

$$0 = x^tKx = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\|x^tf(i,j)\|^2$$

which yields

$$x^tf(i,j) = 0 \qquad \text{for } i \geq 0,\ j \geq 0. \qquad \text{(B2)}$$

Therefore

$$x^t[f(1,0) \quad f(0,1)\cdots f(m,n)] = 0.$$

Condition (28) now implies that $x = 0$ and, hence, K is positive definite. □

References

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

[2] ____, "Roundoff noise in digital filters: Frequency transformations and invariants," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, 538–550, Dec. 1976.

[3] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 256–262, Aug. 1977.

[4] B. W. Bomar, "State-space structures for the realization of low roundoff noise digital filters," dissertation, Univ. Tennessee, 1983.

[5] M.-D. Ni and J. K. Aggarwal, "Two-dimensional digital filtering and its error analysis," *IEEE Trans. Computers*, vol. C-23, 942–954, Sept. 1974.

[6] S. H. Mneney and A. N. Venetsanopoulos, "Finite register length effects in 2-D digital filters," in *Proc. 22nd Midwest Symp. on Circuits and Syst.*, Apr. 1979.

[7] B. G. Mertzios and A. N. Venetsanopoulos, "Combined error at the output of 2-D recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, 888–891, Oct. 1984.

[8] M. Kawamata and T. Higuchi, "Synthesis of 2-D separable denominator digital filters with minimum roundoff noise and no overflow oscillations," in *Proc. 1985 Int. Symp. on Circuits and Syst.* (Kyoto, Japan), June 1985, pp. 1087–1091.

[9] A. S. Willsky, *Digital Signal Processing and Control and Estimation Theory*. Cambridge, MA: MIT Press, 1979.

[10] N. K. Bose, *Applied Multidimensional Systems Theory*. New York: Van Nostrand Reinhold, 1982.

[11] R. P. Roesser, "A discrete state-space model for linear image processing," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 1–10, Feb. 1975.

[12] B. G. Mertzios, "On the roundoff noise in 2-D state-space digital filtering," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 201–204, Feb. 1985.

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

[14] A. Antoniou, *Digital Filters: Analysis and Design*. New York: McGraw-Hill, 1979.

[15] G. W. Stewart, *Introduction to Matrix Computations*. New York: Academic Press, 1973.

[16] D. S. K. Chan, "The structure of recursible multidimensional discrete systems," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 663–673, Aug. 1980.

Wu-Sheng Lu (S'81–M'86) received the B.S. and M.S. degrees in mathematics from Fundan University and East China Normal University, China, in 1964 and 1980, respectively. He received the M.S. degree in electrical engineering and the Ph.D. degree in control science from the University of Minnesota, in 1983 and 1984, respectively.

From October 1984 to December 1985, he was with the Department of Electrical Engineering, University of Victoria, Canada, as a Postdoctoral Fellow. Currently, he is a Visiting Assistant Professor of Electrical Engineering at the University of Minnesota. His research interests include systems theory, and analysis and synthesis of multidimensional digital filters.

Andreas Antoniou (M'69–SM'79–F'82) received the B.Sc.(Eng.) and Ph.D. degrees in electrical engineering from London University, London, UK, in 1963 and 1966, respectively.

From 1966 to 1969, he was Senior Scientific Officer at the Post Office Research Department, London, England, and from 1969 to 1970, he was a member of the Scientific Staff at the R&D Laboratories of Northern Electric Company Ltd., Ottawa, Ontario, Canada. From 1970 to 1983, he served in the Department of Electrical Engineering, Concordia University, Montreal, Quebec, Canada, as Professor from June 1973 and as Chairman from December 1977. On July 1, 1983, he was appointed Founding Chairman of the Department of Electrical Engineering, University of Victoria, Victoria, B.C., Canada.

His teaching and research interests are in the areas of electronics, network synthesis, digital system design, active and digital filters, and digital signal processing. He has published a number of papers on electronic circuits, active filters, and digital filters. He has authored *Digital Filters: Analysis and Design* (New York: McGraw-Hill). One of his papers on gyrator circuits was awarded the Ambrose Fleming Premium by the Institution of Electrical Engineers, UK.

Dr. Antoniou is a Member of the Association of Professional Engineers of British Columbia and a Fellow of the Institution of Electrical Engineers. He was Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS during the period June 1983 to May 1985. He is now serving as Editor of the same TRANSACTIONS.