P42

Roundoff Noise Minimization Using Delta-Operator Realizations

Gang Li and Michel Gevers, *Fellow, IEEE*

Abstract—**We examine the possible advantages of using delta operator state space realizations rather than shift operator realizations of transfer functions in terms of minimizing the roundoff noise gain of the realization. We first give several conditions under which the optimal roundoff noise gain for delta operator realizations is smaller than the optimal gain for shift operator realizations. We then illustrate that even sparse (and hence nonoptimal) delta operator realizations can have smaller roundoff noise gain than the optimal shift operator realizations.**

I. Introduction

One of the interesting problems in the state-variable implementations of transfer functions using finite arithmetic computations is the search for implementations that minimize the roundoff noise gain of the realization. Within the class of usual shift operator state realizations, the roundoff noise gain G_z is equal to the trace of the observability Gramian. Subject to a commonly used dynamic range constraint on the states of the realization, the set of realizations minimizing this roundoff noise gain has been completely characterized by Hwang [1] and Mullis and Roberts [2].

In [3], Williamson proposed the use of residue feedback to reduce the roundoff noise gain for shift-operator realizations and compared it with the optimal gain for realizations without residue feedback. He introduced the concept of residue modes and showed that the superiority of the optimal realizations with residue feedback over the optimal realizations without residue feedback hinged on whether the sum of the residue modes was smaller than the sum of the Hankel singular values.

In this paper, we study the roundoff noise gain G_δ for state variable realizations implemented in the delta operator popularized by Middleton and Goodwin [4], with the aim of examining under what conditions the optimal δ-operator realization roundoff noise gain $G_\delta^{\min}$ is smaller than the optimal shift-operator realization roundoff noise gain $G_z^{\min}$. We first show that the δ-operator implementation is in fact a special case of residue feedback. Therefore, following [3], $G_\delta^{\min}$ will be smaller than $G_z^{\min}$ if and only if the sum of the residue modes is smaller than the sum of the Hankel singular values. We give a few new conditions (i.e., sharper than those in [3]) under which this holds.

A drawback of optimal realizations (i.e., realizations minimizing roundoff noise gains) is that they are typically fully parametrized. This is of course a disadvantage because it maximizes the number of computations. In the last part of this paper, we show that in situations where $G_\delta^{\min}$ is smaller than $G_z^{\min}$, one can obtain nonoptimal sparse δ-form state space realizations whose roundoff noise gain could still be smaller than $G_z^{\min}$.

Manuscript received October 21, 1990; revised December 13, 1991.

G. Li was with the Laboratoire d'Automatique Dynamique et Analyse des Systèmes, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. He is now with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

M. Gevers is with the Laboratoire d'Automatique Dynamique et Analyse des Systèmes, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium.

IEEE Log Number 9205094.

II. The Roundoff Noise Gain of Shift- and Delta-Operator Realizations

A stable strictly causal linear time-invariant system is parametrized as follows in the usual shift operator z:

$$H_z(z) = \frac{\sum_{i=1}^{n} b_i z^{n-i}}{z^n + \sum_{i=1}^{n} a_i z^{n-i}}. \tag{2.1}$$

Defining $\delta = (z - 1)/\Delta$, with $\Delta > 0$, we can alternatively represent the transfer function (2.1) as

$$H_\delta(\delta) = \frac{\sum_{i=1}^{n} \beta_i \delta^{n-i}}{\delta^n + \sum_{i=1}^{n} \alpha_i \delta^{n-i}}. \tag{2.2}$$

The introduction of δ-operator realizations in digital filtering with a view to reducing coefficient sensitivity and signal roundoff noise can probably be traced back to the work of Agarwal and Burrus [5]. This technique was later called "delay replacement" in [6] and [7]. The reasons for using δ-operator models rather than z-operator models have been abundantly developed by Middleton and Goodwin [4]. For reasons of brevity, we shall in future often use the operator ϱ to denote either z or δ. Each of the transfer functions (2.1) and (2.2) can be realized in state-variable form. Using the ϱ-operator, we obtain

$$\begin{cases} \varrho x_t = A_\varrho x_t + B_\varrho u_t \\ y_t = C_\varrho x_t \end{cases} \tag{2.3a}$$

with

$$H_\varrho(\varrho) = C_\varrho(\varrho I - A_\varrho)^{-1} B_\varrho. \tag{2.3b}$$

1053-587X/93$03.00

For $\varrho = z$, (2.3a) is implemented in the usual way using shift registers

$$\begin{cases} x_{t+1} = A_z x_t + B_z u_t \\ y_t \quad = C_z x_t. \end{cases} \tag{2.4a}$$

For $\varrho = \delta$, the δ-operator state-model can be directly evaluated using

$$\begin{cases} x_t = \delta^{-1}(A_\delta x_t + B_\delta u_t) \\ y_t = C_\delta x_t. \end{cases} \tag{2.4b}$$

It uses the basic building block δ^{-1} instead of the classical shift operator z^{-1}. Denoting $A_\delta x_t + B_\delta u_t = w_t$, then

$$x_t \triangleq \delta^{-1} w_t \qquad \text{means that } x_{t+1} = x_t + \Delta w_t. \tag{2.5}$$

We shall denote by S_ϱ ($\varrho = z$ or δ) the following sets of equivalent state-space realizations:

$$S_\varrho \triangleq \{(A_\varrho, B_\varrho, C_\varrho): H_\varrho(\varrho) = C_\varrho(\varrho I - A_\varrho)^{-1}B_\varrho\}. \tag{2.6}$$

It is easy to verify that to each realization $(A_\delta, B_\delta, C_\delta) \in S_\delta$ as in (2.4b) there corresponds a realization $(A_z, B_z, C_z) \in S_z$ as in (2.4a) through the following relationship:

$$\begin{cases} A_z = \Delta A_\delta + I \\ B_z = \Delta B_\delta \\ C_z = C_\delta. \end{cases} \tag{2.7}$$

Besides the theoretically interesting property that when the sampling time goes to zero the δ-operator models of sampled data systems approach the continuous-time models (see [4]), the main potential advantage of δ-operator models is numerical, in the case of finite arithmetic under fast sampling. In [8] we have compared absolute and relative sensitivities of z- and δ-operator state space models w.r.t. the parameters of the state space matrices. Our purpose here will be to compare their respective roundoff noise gains.

We first recall a few basic facts concerning roundoff noise propagation in state-variable realizations when the quantizations are carried out *before multiplication*. We refer to [3] for more details. Assuming that the coefficients of $(A_\varrho, B_\varrho, C_\varrho)$ are represented exactly with B_c fractional bits, that the state and the output have B fractional bits ($B > B_c$), and that the input signal has $B - B_c$ fractional bits, then the finite word length (FWL) implementation of (2.3a) is

$$\begin{cases} \varrho x_t^* = A_\varrho Q(x_t^*) + B_\varrho u_t \\ y_t^* \quad = C_\varrho Q(x_t^*). \end{cases} \tag{2.8}$$

Here Q represents the quantizer: it rounds the B bit fraction x_t^* to $(B - B_c)$ bits before multiplication. The roundoff noise

$$e_t \triangleq x_t^* - Q(x_t^*) \tag{2.9}$$

is usually modeled as white noise of zero mean with covariance $q^2 I$, with $q^2 = (1/12)2^{-2(B-B_c)}$.

Comment 2.1: Expression (2.8) represents the FWL implementation of (2.3a) under the assumption that the operator ϱ is implemented exactly. This is the case when $\varrho = z$, or $\varrho = \delta$ with $\Delta = 1$. If $\varrho = (z-1)/\Delta$, the actual implementation of the δ-operator model (2.8) is as follows [see (2.5)]:

$$\begin{cases} x_{t+1}^* = x_t^* + \Delta(A_\delta Q(x_t^*) + B_\delta u_t) \\ y_t^* \quad = C_\delta Q(x_t^*). \end{cases} \tag{2.10}$$

We have shown in [8] that, in order to minimize the sensitivity of the transfer function to coefficient errors, Δ should be chosen as small as possible, but compatible with the dynamic range constraints on the coefficients of A_δ, B_δ, and C_δ. This often allows one to choose values $\Delta < 1$ yielding a minimal sensitivity. Clearly, when $\Delta < 1$, an additional noise is introduced by the multiplication of Δ by $w_t \triangleq A_\delta Q(x_t^*) + B_\delta u_t$ in (2.10). Indeed, if the implementation of Δ requires B_Δ bits, then w_t must be rounded off to $B - B_\Delta$ bits to produce a B-bit number in (2.10). A complete analysis of this additional roundoff noise can easily be performed. In practice $B_\Delta << B_c$ (for example, $B_\Delta = 1$, $B_c = 8$ is typical), and the analysis then shows that this additional noise introduced by the multiplication with Δ in (2.10) can be neglected.

Comment 2.2: One procedure that has been advocated to reduce the effect of roundoff noise in digital filter realizations is the use of integer residue feedback [3], [9], and [10]. In such case, the FWL shift-operator state-space realization (2.8) is modified according to

$$\begin{cases} x_{t+1}^* = A_z Q(x_t^*) + B_z u_t + J e_t \\ y_t^* \quad = C_z Q(x_t^*) + h e_t \end{cases} \tag{2.11}$$

where all components of J and h are integers (see [3]). We note that with the choices $J = I$ and $h = 0$ and using (2.7), (2.11) becomes identical to the δ-operator implementation (2.10) with $\Delta = 1$. We conclude that the FWL δ-realization is a special case of the residue feedback realization for the choices $J = I$ and $h = 0$.

Denoting $\epsilon y_t \triangleq y_t - y_t^*$, then the roundoff noise gain is usually defined as (see e.g., [3])

$$G = \frac{1}{q^2} \lim_{t \to \infty} E[(\epsilon y_t)^2]. \tag{2.12}$$

To compute G in the special cases of z- and δ-operator realizations, we first write a state equation for the error $E_t \triangleq x_t - x_t^*$. It follows from (2.8) and (2.9) that

$$\begin{cases} \varrho E_t = A_\varrho E_t + A_\varrho e_t \\ \epsilon y_t \ = C_\varrho E_t + C_\varrho e_t. \end{cases} \tag{2.13}$$

Replacing ϱ, respectively, by z and δ in (2.13), it then follows that

$$G_z = \operatorname{tr}(W_0) \tag{2.14}$$

$$G_\delta = \operatorname{tr}(W) \tag{2.15}$$

where W_0 is the observability Gramian of the realization (A_z, B_z, C_z)

$$W_0 \triangleq \sum_{i=0}^{\infty} (A_z^i)^T C_z^T C_z A_z^i \tag{2.16}$$

and W is defined as

$$W \triangleq \Delta A_\delta^T W_0 \Delta A_\delta + C_z^T C_z. \tag{2.17}$$

W_0 is the solution of the Lyapunov equation

$$W_0 = A_z^T W_0 A_z + C_z^T C_z. \tag{2.18}$$

Using (2.18) and (2.7), alternative expressions for W can be obtained

$$\begin{aligned} W &= (A_z - I)^T W_0 (A_z - I) + C_z^T C_z \\ &= (I - A_z)^T W_0 + W_0 (I - A_z) \\ &= 2W_0 - A_z^T W_0 - W_0 A_z. \end{aligned} \tag{2.19}$$

It should be clear that the previous expressions hold for matrices (A_δ, C_δ) and (A_z, C_z) that are related by (2.7).

III. Minimization of the Roundoff Noise Gain

For any realization $(A_z, B_z, C_z) \in S_z$, whose corresponding realization in S_δ is $(A_\delta, B_\delta, C_\delta) \in S_\delta$ through (2.7), one has

$$A_z' = T^{-1} A_z T, \qquad B_z' = T^{-1} B_z, \qquad C_z' = C_z T \tag{3.1a}$$

and

$$A_\delta' = T^{-1} A_\delta T, \qquad B_\delta' = T^{-1} B_\delta, \qquad C_\delta' = C_\delta T \tag{3.1b}$$

where T is any nonsingular matrix of proper dimension. It is therefore clear that if (W_0, W) defined in (2.16) and (2.17) correspond to (A_z, B_z, C_z) [equivalently to $(A_\delta, B_\delta, C_\delta)$], the corresponding (W_0', W') in the new coordinates satisfy the following transformation:

$$W_0' = T^T W_0 T, \qquad W' = T^T W T. \tag{3.2}$$

It follows from (2.14) and (2.15) that different realizations in S_ϱ ($\varrho = z, \delta$) yield different roundoff noise gains. The interesting problem is to find the optimal realizations in S_ϱ, which minimize the roundoff noise gain

$$\min_{(A_z', B_z', C_z') \in S_z} G_z' = \min_{T:\, \det T \neq 0} \operatorname{tr}(T^T W_0 T) \tag{3.3a}$$

$$\min_{(A_\delta, B_\delta, C_\delta) \in S_\delta} G_\delta' = \min_{T:\, \det T \neq 0} \operatorname{tr}(T^T W T). \tag{3.3b}$$

Note that the problem (3.3) does not make sense unless a scaling of the states is introduced since "smaller" T yielding smaller G_ϱ' would make the states larger. In order to maintain the amplitude of the states within an acceptable range, and hence to reduce the probability of overflow, an l_2-norm scaling on the states is introduced in practice [1], [2], which is equivalent to the following constraint on the realizations in the new coordinates:

$$(W_c')_{ii} = (T^{-1} W_c T^{-T})_{ii} = 1 \quad \forall i \tag{3.4}$$

where

$$W_c \triangleq \sum_{i=0}^{\infty} A_z^i B_z B_z^T (A_z^i)^T \tag{3.5}$$

is the controllability Gramian of the system under the realization $(A_z, B_z, C_z) \in S_z$. So, now the problem of minimizing the roundoff noise gain G_ϱ' under l_2-scaling can be formulated by combining (3.3) and (3.4), (3.5). The minimum achievable gain in S_z was originally given in [1] and [2]

$$G_z^{\min} = \frac{1}{n} \left(\sum_{i=1}^{n} \sigma_i \right)^2 \tag{3.6}$$

where $\{\sigma_i\}$ is the Hankel singular value set of the system defined by

$$\{\sigma_i^2\} = \lambda(W_c W_0) = \lambda(W_c' W_0'). \tag{3.7}$$

This minimum is achieved by a set of realizations in S_z, all of which satisfy the dynamic range constraint (3.4). A constructive procedure for computing this optimal realization set has been given by Hwang [1].

The noise gain in S_δ is given by the exact same form as the noise gain in S_z, with W_0 replaced by W and with the same dependence on T [see (3.3a) and (3.3b)], while the same l_2-norm scaling (3.4) applies. Therefore the procedure of [1] and [2] applies identically to this case. The minimum noise gain in S_δ is thus given by

$$G_\delta^{\min} = \frac{1}{n} \left(\sum_{i=1}^{n} \nu_i \right)^2 \tag{3.8}$$

where $\{\nu_i\}$ is called the residue mode set [3] defined by

$$\{\nu_i^2\} = \lambda(W_c W) = \lambda(W_c' W'). \tag{3.9}$$

For the same reason, the optimal realizations in S_δ that achieve $G_\delta^{\min}$ are obtained in the same way as in [1].

Comment 3.1: Since the residue feedback realization of [3], in the case $J = I$ and $h = 0$, is identical to the δ-realization with $\Delta = 1$ (see Comment 2.2), it follows that the optimal residue feedback realizations in this special case are identical to the optimal δ-realizations, and hence are also obtained by Hwang's procedure. This result was rederived by Williamson [3, theorem 5.2].

Now a reasonable question is under what conditions do we have

$$G_\delta^{\min} < G_z^{\min}. \tag{3.10}$$

Clearly (3.10) holds if and only if the sum of the residue modes is less than the sum of the Hankel singular values. It would be interesting to produce simple conditions—on $H(z)$ or on some realization of $H(z)$—under which (3.10) holds, without having to compute the Hankel singular values δ_i and the Residue Modes ν_i. This problem was addressed by Williamson [3] who gave some sufficient conditions for (3.10) to hold. In the next section, we will give some new conditions and compare them with Williamson's.

IV. Some New Conditions for Superiority of δ-Realizations

In this section, we give some conditions under which δ-operator implementations yield a smaller roundoff noise gain than shift-operator realizations, that is conditions on the transfer function under which (3.10) can be achieved. In order to do so, we first present the following lemma.

Lemma 4.1: Let $\{\rho_i^2, \rho_i \geq 0\}$ and $\{\theta_i^2, \theta_i \geq 0\}$ be the diagonal element and eigenvalue sets, respectively, of a semipositive definite symmetric matrix M. Then

$$\sum_{i=1}^{n} \rho_i \geq \sum_{i=1}^{n} \theta_i \tag{4.1}$$

and equality is achieved if and only if the matrix M is diagonal, i.e., $\rho_i = \theta_i \; \forall i$.

Proof: see [1].

With this lemma we can prove the following theorem, which gives a first new set of sufficient conditions under which (3.10) holds.

Theorem 4.1: For any stable minimal SISO system (2.1), (3.10) holds if all the diagonal elements $\{a_{ii}\}$ of A_z^{in} satisfy

$$\tfrac{1}{2} \leq a_{ii} \quad \forall i \tag{4.2}$$

where $(A_z^{\text{in}}, B_z^{\text{in}}, C_z^{\text{in}})$ is the input-balanced realization of $H(z)$, characterized by its Gramian matrices

$$W_c^{\text{in}} = I,\; W_0^{\text{in}} = \text{diag}\,(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2) \triangleq \Sigma^2 \tag{4.3}$$

Proof: First, we note that the Hankel singular values $\{\sigma_i\}$ and residue modes $\{\nu_i\}$ are invariant under a coordinate transformation. So

$$\{\nu_i^2\} = \lambda(W_c W) = \lambda(W_c^{\text{in}} W^{\text{in}}) = \lambda(W^{\text{in}}) \tag{4.4}$$

where W^{in} is defined in (2.19) for the input balanced realization characterized by (4.3)

$$\begin{aligned} W^{\text{in}} &= (A_z^{\text{in}} - I)^T \Sigma^2 (A_z^{\text{in}} - I) + (C_z^{\text{in}})^T C_z^{\text{in}} \\ &= (I - A_z^{\text{in}})^T \Sigma^2 + \Sigma^2 (I - A_z^{\text{in}}). \end{aligned} \tag{4.5}$$

Denote $W^{\text{in}} \triangleq \{w_{ij}\}$ and $A_z^{\text{in}} \triangleq \{a_{ij}\}$. It is clear that

$$w_{ii} = 2(1 - a_{ii})\sigma_i^2 \quad \forall i.$$

Since W^{in} is positive definite and symmetric, it follows that $w_{ii} > 0$. According to Lemma 4.1, one has

$$\sum_{i=1}^{n} \nu_i \leq \sum_{i=1}^{n} w_{ii}^{1/2} = \sum_{i=1}^{n} \sqrt{2(1 - a_{ii})}\,\sigma_i.$$

The theorem follows from the fact that (3.10) holds if (4.2) is satisfied. ■

In [3], Williamson has given another sufficient condition under which (3.10) holds. This condition is on the internally balanced realization $(A_z^{ib}, B_z^{ib}, C_z^{ib}) \in S_z$, which is characterized by its Gramians

$$W_c^{ib} = W_0^{ib} = \Sigma = \text{diag}\,(\sigma_1, \sigma_2, \cdots, \sigma_n). \tag{4.6}$$

We now give a sharper result, also on the internally balanced realization.

Theorem 4.2: For any stable minimal discrete linear time invariant SISO system, there exists an internally balanced form $(A_z^{ib}, B_z^{ib}, C_z^{ib}) \in S_z$ such that

$$1)\; A_z^{ib} = \begin{pmatrix} A_{11} & A_{12} \\ -A_{12}^T & A_{22} \end{pmatrix} \quad \text{with } A_{ii} = A_{ii}^T, \quad i = 1, 2 \tag{4.7a}$$

2) If $\{\theta_i\} \triangleq$ union of $\lambda(A_{11})$ and $\lambda(A_{22})$, then

$$\min_i \theta_i \leq \text{Re}\,(\lambda_k) \leq \max_i \theta_i \tag{4.7b}$$

for any $\lambda_k \in \lambda(A_z^{ib}) \quad \forall k$[1].

3) if $\min \theta_i \geq 1/2$, then (3.10) holds. (4.7c)

Proof: see the Appendix.

We now compare our results with those of Williamson [3].

1. The existence of (4.7a) is always guaranteed while, in [3], the Hankel singular values are assumed to be distinct for (4.7a) to exist.

2) Williamson gave the following sufficient condition for (3.10): $\min_i \theta_i \geq 1 - (1/2n)$ with n the order of system. Here we need only $\min_i \theta_i \geq 1/2$, which is a sharper result.

Comment 4.1: Theorems 4.1 and 4.2 yield some sufficient conditions under which (3.10) holds. These conditions require the computation of (input or internally) balanced forms. Numerically well-conditioned algorithms to compute balanced forms can be found in [11] and [12].

Now, we will give another sufficient condition for (3.10) to hold. This condition is on the poles of the system.

Theorem 4.3: For any stable minimal discrete SISO linear time-invariant system, if the poles λ_i of the system satisfy the following condition

$$\sum_{i=1}^{n} \lambda_i \geq n - \frac{1}{2} \tag{4.8}$$

then (3.10) holds.

Proof: The internally balanced form defined by (4.6) satisfies the following Lyapunov matrix equation

$$\Sigma = A_z^{ib} \Sigma A_z^{ibT} + B_z^{ib} B_z^{ibT}. \tag{4.9}$$

It follows that the diagonal elements $\{a_{ii}\}$ of A_z^{ib} satisfy

$$|a_{ii}| < 1. \tag{4.10}$$

From $\Sigma_{i=1}^n a_{ii} = \Sigma_{i=1}^n \lambda_i$ one obtains

$$\min_i a_{ii} + (n - 1) > \sum_{i=1}^{n} \lambda_i$$

or

$$\min_i a_{ii} > 1 - \left(n - \sum_{i=1}^{n} \lambda_i\right).$$

[1]We use $\lambda(A)$ to denote the set of all eigenvalues of A.

Therefore, using (4.8), we have $\min_i a_{ii} > 1/2$. The theorem follows directly by applying Theorem 4.1. ■

The sufficient condition (4.8) can be rewritten as

$$\bar{\lambda} \triangleq \frac{1}{n}\sum_{i=1}^{n} \lambda_i \geq 1 - \frac{1}{2n} \triangleq \bar{\lambda}_{\min}. \qquad (4.11)$$

Clearly, $\bar{\lambda}$ is the mean value of the poles.

Example 1: for $n = 4$, $\bar{\lambda}_{\min} = 1 - 1/2n = 0.875$. So, for any system of order 4, the optimal δ-operator implementation will be superior to the shift operator implementation in terms of roundoff noise gain if the mean pole value $\bar{\lambda}$ is larger than 0.875.

Example 2: in [12], a sixth-order narrow-band low-pass filter is considered, whose poles are $0.9723 \pm j0.1989$, $0.9389 \pm j0.1623$, $0.9152 \pm j0.0646$. For this filter, one has

$$\bar{\lambda} = 0.9441, \qquad \text{and } \bar{\lambda}_{\min} = 0.9167.$$

So, for this system, the optimal δ-operator implementation will have a better performance in terms of roundoff noise gain than the optimal shift-operator implementation.

Comment 4.2: Theorem 4.3 yields a sufficient condition for (3.10) that is very easy to test, since the system is normally given by its transfer function from which the poles can be obtained easily.

Comment 4.3: This theorem guarantees the superiority of implementation in δ-operator over shift operator for a class of systems. In fact, it implies that for systems whose poles are clustered around $z = +1$, the δ-operator implementation will yield a better performance in terms of minimizing the roundoff noise gain. The often used narrow-band low-pass filters belong to this class [7]. In modern control, fast sampling is used in order to keep enough information [4]. The discrete time models used in practice come from the corresponding continuous time systems sampled with high frequency. With the sampling frequency chosen between 5 and 50 times the maximal frequency of interest as proposed by Middleton and Goodwin [4], the poles of the corresponding discrete time models and controllers are clustered around $z = +1$, and so here again the δ-operator models will typically perform better.

Comment 4.4: The optimal realizations in either S_z or S_δ yield a system matrix $(A, B, C)_{\text{opt}}$ full of non-one-or-zero elements, which is not very desirable since it maximizes the number of arithmetic operations. For those reasons, sparser realizations are preferred and some efforts have been made in this direction [13]–[16]. We note that for the class of systems discussed just before, $G_\delta^{\min}$ is smaller than $G_z^{\min}$. This implies that some sparse realizations in S_δ could have a noise gain G_δ near $G_z^{\min}$ (of course, larger than $G_\delta^{\min}$). For example, the companion form (direct form) realization in S_δ can give a very nice performance [7]. In the next section, we will give another sparse realization in S_δ based on a polynomial parametrization approach. With the same numerical example as in [7] we will see that this realization yields a roundoff noise gain G_δ smaller than $G_z^{\min}$.

V. A Sparser Realization in S_δ

In this section, we first give a brief introduction of the polynomial parametrization concept and present some results without proofs. These polynomial parametrizations are fully developed and exploited in [17] and [18]. Based on Chebyshev polynomials, a sparse structure will be given which, in fact, is a realization in S_δ. This structure will be seen to have better performance than the optimal realization in S_z in terms of minimizing the roundoff noise gain for the numerical example to be given in the next section.

We start here with the representation (2.2), which we recall for convenience

$$H_\delta(\delta) = \frac{\beta_1\delta^{n-1} + \cdots + \beta_n}{\delta^n + \alpha_1\delta^{n-1} + \cdots + \alpha_n}. \qquad (5.1)$$

Using vector notations

$$\bar{\alpha} \triangleq (1 \quad \alpha_1 \quad \cdots \quad \alpha_n)^T, \qquad \bar{\beta} \triangleq (0 \quad \beta_1 \quad \cdots \quad \beta_n)^T \qquad (5.2)$$

and

$$\bar{\delta} \triangleq (\delta^n \quad \delta^{n-1} \quad \cdots \quad \delta \quad 1)^T \qquad (5.3)$$

(5.1) can be written as

$$H_\delta(\delta) = \frac{\bar{\beta}^T\bar{\delta}}{\bar{\alpha}^T\bar{\delta}} = \frac{\bar{\beta}^T T^{-1} T\bar{\delta}}{\bar{\alpha}^T T^{-1} T\bar{\delta}} = \frac{\bar{\theta}^T\bar{p}(\delta)}{\bar{\eta}^T\bar{p}(\delta)} \qquad (5.4)$$

where

$$\bar{\eta} = T^{-T}\bar{\alpha} \triangleq (1 \quad \eta_1 \quad \cdots \quad \eta_n)^T \qquad (5.5a)$$

$$\bar{\theta} = T^{-T}\bar{\beta} \triangleq (0 \quad \theta_1 \quad \cdots \quad \theta_n)^T \qquad (5.5b)$$

$$\bar{p}(\delta) = T\bar{\delta} \triangleq [p_0(\delta) \cdots p_{n-1}(\delta)\, p_n(\delta)]^T \qquad (5.5c)$$

with T an $(n + 1) \times (n + 1)$ nonsingular matrix, whose first column is $(1 \quad 0 \quad 0 \quad \cdots \quad 0)^T$ but that is otherwise arbitrary. So

$$H_\delta(\delta) = \frac{\theta_1 p_1(\delta) + \cdots + \theta_n p_n(\delta)}{p_0(\delta) + \eta_1 p_1(\delta) + \cdots + \eta_n p_n(\delta)}. \qquad (5.6)$$

It is clear that now the system is parametrized by $(\bar{\eta}, \bar{\theta})$ under the polynomial operator $\bar{p}(\delta)$ where $p_0(\delta)$ is a monic polynomial of degree n, and $p_i(\delta)$, $i = 1, \cdots, n$, are polynomials of degree less than n. The transformation from (5.1) to (5.6) is uniquely determined by the choice of the polynomial set $p_i(\delta)$ or, equivalently, the matrix T. These two quantities are related, in a one-to-one way, as follows:

$$T = \begin{pmatrix} 1 & p_{01} & p_{02} & \cdots & p_{0n} \\ 0 & p_{11} & p_{12} & \cdots & p_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \triangleq \begin{pmatrix} 1 & p_{01} & \cdots & p_{0n} \\ 0 & & & \\ \vdots & & T_p & \\ 0 & & & \end{pmatrix} \qquad (5.7)$$

where the p_{ij} are the coefficients of the polynomials $p_i(\delta)$: $p_i(\delta) = \Sigma_{j=1}^{n} p_{ij}\delta^{n-j}$.

If $(A_\delta^c, B_\delta^c, C_\delta^c) \in S_\delta$ is the controllable realization in S_δ corresponding to the transfer function model (5.1), then a realization $(A_\delta, B_\delta, C_\delta) \in S_\delta$ corresponding to (5.6) can be obtained by (see [17] and [18]):

$$A_\delta = T_p A_\delta^c T_p^{-1}, \qquad B_\delta = T_p B_\delta^c, \qquad C_\delta = C_\delta^c T_p^{-1} \tag{5.8}$$

where $T_p \in \mathbb{R}^{n\times n}$ is defined in (5.7) and depends only on the last n polynomials $p_1(\delta), \cdots, p_n(\delta)$ of degree $n-1$. Clearly, by proper choice of the polynomial set (i.e., of T_p), the realization $(A_\delta, B_\delta, C_\delta)$ can be put in a desired form through (5.8).

We illustrate the use of polynomial basis functions with Chebyshev polynomials of the first type. Without going into details, consider polynomials generated in the following recursive way (with $\Delta = 1$):

$$p_{i-1}(\delta) = \delta p_i(\delta) + c_i p_{i+1}(\delta) \tag{5.9}$$

with $p_n(\delta) = 1$, $p_{n-1}(\delta) = \delta$. In state-space form this corresponds to

$$A_\delta = \begin{pmatrix} -\eta_1 & -\eta_2 - c_1 & -\eta_3 & \cdots & -\eta_{n-1} & -\eta_n \\ 1 & 0 & -c_2 & \cdots & 0 & 0 \\ \vdots & & & & \vdots & \vdots \\ & & & & 0 & -c_{n-1} \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \quad B_\delta = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad C_\delta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix}^T \tag{5.10}$$

$\{\eta_i\}$, $\{\theta_i\}$ can be obtained through (5.5) since T is determined from the polynomial set.

With the following choice for the c_i, (5.9) generates polynomials $p_i(\delta)$ related to the Chebyshev polynomials of the first type

$$c_i = \begin{cases} -(4k^2)^{-1}, & i = 1, 2, \cdots, n-2 \\ -(2k^2)^{-1}, & i = n-1. \end{cases} \tag{5.11}$$

We will say that the corresponding $(A_\delta, B_\delta, C_\delta)$ is in Chebyshev form. Another special case of (5.9) and (5.10) is when $c_1 = c_2 = \cdots = c_n = 0$. This corresponds to the delay replaced direct form of [7]. It can also be seen as a special case of a Chebyshev form with the choice $k = \infty$.

The realization (5.10) should be scaled before it is implemented. This can be done by simply applying a diagonal transformation matrix T, which leaves the zeroes unchanged in the form (5.10). The l_2-scaled version of (5.10) requires $(2n-3)$ more multipliers than the l_2-scaled controllable realization $(A_z^c, B_z^c, C_z^c) \in S_z$ and $(n-2)$ more multipliers than the l_2-scaled controllable realization $(A_\delta^c, B_\delta^c, C_\delta^c) \in S_\delta$. Since the c_i in (5.10) are free, one could think of minimizing the roundoff noise gain over all sparse realizations of the form (5.10) by optimizing over the c_i. This is a very hard problem. Instead, one can restrict oneself to Chebyshev forms, where the c_i obey (5.11), and use the scalar factor k, called adaptive factor, to make the roundoff noise gain G_δ of (5.10) as small as possible. In the special case where the optimal k would be found to be infinite, this would indicate that the delay replaced direct form is optimal among all realizations of the form (5.10). In the next section we give a numerical example wherein the structure (5.10) yields a G_δ that is an order of magnitude smaller than the roundoff noise gain of $(A_\delta^c, B_\delta^c, C_\delta^c)$ and five times smaller than $G_z^{\min}$.

VI. Numerical Example

For ease of comparison, we will use the same example as in [7]. This is a sixth-order narrow-band low-pass filter with a normalized sampling frequency $f_s = 1$. The corresponding transfer function and l_2-scaled direct form realization $(A_\delta^c, B_\delta^c, C_\delta^c) \in S_\delta$, called delay replaced direct form (DRDF), can be found in [7] and corresponds to $\Delta = 1$ and $k = \infty$. We shall compare it with the l_2-scaled realization $(A_\delta, B_\delta, C_\delta)$ in (5.10) with $\Delta = 1$ and $k = 4$

$$A_\delta = \begin{pmatrix} -0.3474 & -0.2780 & 0.2167 & 0.1148 & -0.2829 & -0.1607 \\ 0.1460 & 0 & -0.0912 & 0 & 0 & 0 \\ 0 & 0.1713 & 0 & -0.0820 & 0 & 0 \\ 0 & 0 & 0.1905 & 0 & -0.1898 & 0 \\ 0 & 0 & 0 & 0.0823 & 0 & -0.1779 \\ 0 & 0 & 0 & 0 & 0.1757 & 0 \end{pmatrix}$$

$$B_\delta = (0.3562 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)^T$$

$$C_\delta = (0.0042 \quad 0.0576 \quad 0.0683 \quad 0.1462 \quad 0.0824 \quad 0.2085). \tag{6.1}$$

TABLE I
PERFORMANCE COMPARISON OF FIVE DIFFERENT REALIZATIONS

l_2-scaled	G_ϱ	M_ϱ
$(A_z^c, B_z^c, C_z^c) \in S_z$	1.973×10^{10}	1.3814×10^{11}
$(A_z, B_z, C_z)_{\text{opt}} \in S_z$	1.3329	15.3306
$(A_\delta^c, B_\delta^c, C_\delta^c) \in S_\delta$(DRDF)	2.6985	1.1514×10^3
$(A_\delta, B_\delta, C_\delta)_{\text{opt}} \in S_\delta$	0.0646	18.3936
$(A_\delta, B_\delta, C_\delta)$ in (6.1)	0.2876	73.9616

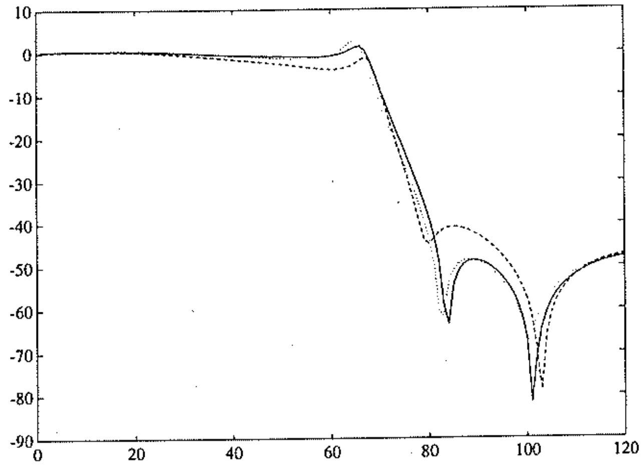

Fig. 1. Frequency responses for the three realizations.

It has been shown [19] that the sensitivity measure M_z of any l_2-scaled realization (A_z, B_z, C_z) is given by

$$M_z = (n + 1) \operatorname{tr}(W_0) + n. \tag{6.2}$$

In [8] it is shown that with $\Delta = 1$ the sensitivity measure M_δ of any $(A_\delta, B_\delta, C_\delta) \in S_\delta$ is also given by (6.2), where $(A_\delta, B_\delta, C_\delta)$ and (A_z, B_z, C_z) are related by (2.7). The theoretical results are given in Table I.

Comments:

1) That $G_\delta^{\min}$ (= 0.0646) $<$ $G_z^{\min}$ (= 1.3329) was already known through the calculations in Example 2 of Section IV and the application of Theorem 4.3. Both $(A_z, B_z, C_z)_{\text{opt}}$ and $(A_\delta, B_\delta, C_\delta)_{\text{opt}}$ yield fully parametrized realizations.

2) The delay replaced direct form $(A_\delta^c, B_\delta^c, C_\delta^c)$ has a performance in terms of G_δ that is not much worse than $G_z^{\min}$, which corresponds to a fully parametrized realization. By adding $(n - 2)$ elements to $(A_\delta^c, B_\delta^c, C_\delta^c)$, the realization $(A_\delta, B_\delta, C_\delta)$ of (6.1) yields a roundoff noise gain (0.2876) that is 10 times smaller than that of the DRDF $(A_\delta^c, B_\delta^c, C_\delta^c)$ and almost five times smaller than $G_z^{\min}$.

3) To confirm the computation of M_ϱ, we give some simulations based on the l_2-scaled FWL (coefficient) implementations for three realizations: (A_z^c, B_z^c, C_z^c), the DRDF $(A_\delta^c, B_\delta^c, C_\delta^c)$ and the Chebyshev form $(A_\delta, B_\delta, C_\delta)$ of (6.1). For each of these realizations, we round the coefficients to p bits, then compute the magnitude of the corresponding frequency response and compare it with that of the ideal frequency response (corresponding to $p = \infty$). The results are given in Figs. 1 and 2.

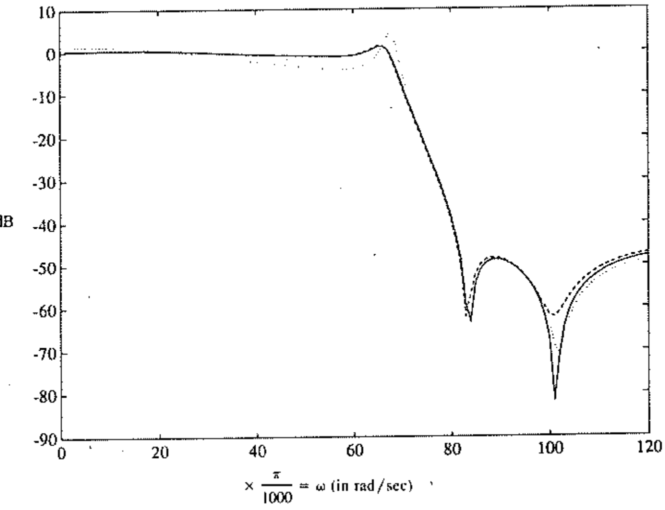

Fig. 2. Frequency responses for the three realizations.

The figures show that the δ-operator implementation in the form (5.10) gives an excellent fit to the ideal frequency response, particularly in the lower frequency range: with the same number of bits, eight, it yields a much better result than the direct δ-form, and with 10 bits it is almost indistinguishable from the ideal frequency response. The superiority of the δ-operator implementation (5.10) over the direct form shift-operator realization is evident: the 10-bit implementation of $(A_\delta, B_\delta, C_\delta)$ in the form (5.10) yields an even better fit than the 18-bit shift-operator form.

VII. CONCLUSION

In this note, we have analyzed the FWL implementation of the state-space model of a discrete system in δ-operator form. The expression of the roundoff noise gain has been derived. It has been shown that the δ-operator implementation is, in fact, a special case of the FWL implementation with residue feedback [3], [9]. We have then examined the problem of minimizing the roundoff noise gain over the realization set of δ-operator models, S_δ. Some new conditions for the superiority of optimal δ-operator implementations over optimal z-operator implementations have been given, where the optimality is in terms of minimizing the roundoff noise gain. The superiority of the optimal δ-operator implementations over the usual shift operator implementations allows one to find some sparser realizations in S_δ, which often yield almost the same performance as the fully parametrized optimal realizations in S_z. Our theoretical results have been confirmed by a numerical example and by simulations.

Note that the conditions given in this paper are sufficient for guaranteeing $G_\delta^{\min} < G_z^{\min}$. The numerical example shows that one often has $G_\delta^{\min} \ll G_z^{\min}$. One open problem is how to sharpen those sufficient conditions further. It is believed that the answer to this problem depends on the exploration of new properties of balanced forms for discrete time systems. Some investigations along this line are being carried out.

Appendix
Proof of Theorem 4.2

First, note that Kung [20] has shown that for any stable minimal SISO discrete linear system, there always exists an internally balanced form that satisfies the following symmetry property

$$A_z^{ib} = QA_z^{ib^T}Q, \qquad B^{ib} = QC_z^{ib^T} \tag{A.1}$$

where Q is a sign matrix

$$Q = \operatorname{diag}(u_1, u_2, \cdots, u_n) \qquad u_i = \pm 1 \quad \forall i$$

Clearly, with a series of permutations, Q can be transformed to

$$\begin{pmatrix} I_1 & 0 \\ 0 & -I_2 \end{pmatrix} \text{ with}$$

$$I_i = \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}_{n_i \times n_i} \qquad i = 1, 2 \tag{A.2}$$

where $n_1 + n_2 = n$, while preserving the structure of (A.1). So, without loss of generality, Q in (A.1) can be assumed to be of the form (A.2). It follows from (A.1) that

$$A_z^{ib}Q = QA_z^{ib^T} = (A_z^{ib}Q)^T = \begin{pmatrix} X_{11} & X_{12} \\ X_{12}^T & X_{22} \end{pmatrix}$$

with X_{ii}, $i = 1, 2$, real symmetric. It then follows that

$$A_z^{ib} = \begin{pmatrix} X_{11} & X_{12} \\ X_{12}^T & X_{22} \end{pmatrix} Q = \begin{pmatrix} X_{11} & X_{12} \\ X_{12}^T & X_{22} \end{pmatrix} \begin{pmatrix} I_1 & O \\ O & -I_2 \end{pmatrix}$$

$$= \begin{pmatrix} X_{11} & -X_{12} \\ X_{12}^T & -X_{22} \end{pmatrix}$$

So, (4.7a) is proved with the following identifications:

$$A_{11} = X_{11}, \qquad A_{22} = -X_{22}, \qquad A_{12} = -X_{12}$$

Now consider (4.7b). It is well known that for any square real matrix A, the following decomposition holds

$$A = A_s + A_{sk} \qquad \text{with } A_s = A_s^T,\ A_{sk}^T = -A_{sk}. \tag{A.3}$$

If λ is an eigenvalue, $\lambda \in \lambda(A)$, with X the corresponding eigenvector, then

$$AX = \lambda X \Rightarrow X^H AX = \lambda \|X\|_2^2 \Rightarrow X^H A_s X = \operatorname{Re}(\lambda)\|X\|_2^2.$$

By SVD decomposition, A_s can be written as

$$A_s = U \begin{pmatrix} \theta_1 & & 0 \\ & \ddots & \\ 0 & & \theta_n \end{pmatrix} U^T \text{ with } U \text{ orthogonal.}$$

Denote $Y = UX$, then

$$\sum_{i=0}^{n} \theta_i |y_i|^2 = \operatorname{Re}(\lambda) \|X\|_2^2.$$

Since $\|Y\|_2^2 = \|X\|_2^2$, or $\Sigma_{i=1}^n |y_i|^2 = \Sigma_{i=1}^n \|X_i\|^2$, it follows that

$$\min_i \theta_i \le \operatorname{Re}(\lambda) \le \max_i \theta_i$$

(4.7b) follows with

$$A_s = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}, \qquad A_{sk} = \begin{pmatrix} 0 & A_{12} \\ -A_{12}^T & 0 \end{pmatrix}.$$

Finally, recall that $\{\theta_i\} = \lambda(A_{11}) \cup \lambda(A_{22})$ with A_{11}, A_{22} real symmetric. If $\min_i \theta_i > 0$, then A_{ii}, $i = 1, 2$ are also positive definite. It is well known (see [21 p. 134]) that

$$\min_i \theta_i \le \min_i a_{ii}$$

where $\{a_{ii}\}$ is the diagonal element set of A_s (or A_z^{ib}). We note that the matrix A_z^{in} of the input-balanced form has the same diagonal elements as A_z^{ib}. Clearly, (4.7c) then follows from Theorem 4.1 with $\min_i \theta_i \ge 1/2$.

Acknowledgment

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility rests with its authors. The authors would like to thank the reviewers for helpful comments that helped improve the quality of this paper.

References

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

[2] C. T. Mullis and R. A. Roberts, "Filter structures which minimize roundoff noise in fixed-point digital filters," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, 1976, pp. 505–508.

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

[4] R. H. Middleton and G. C. Goodwin, *Digital Control and Estimation: A Unified Approach.* Englewood Cliffs, NJ: Prentice Hall, 1990.

[5] R. C. Agarwal and C. S. Burrus, "New recursive digital filter structures having very low sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-22, no. 12, pp. 921–927, Dec. 1975.

[6] G. Orlandi and G. Martinelli, "Low-sensitivity recursive digital filters obtained via the delay replacement," *IEEE Trans. Circuits Syst.*, vol. CAS-31, no. 7, pp. 654–657, July 1984.

[7] D. Williamson, "Delay replacement in direct form structures," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 4, pp. 453–460, Apr. 1988.

[8] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parametrization," in *Proc. 29th IEEE Conf. Decision and Control*, Hawaii, vol. 2, Dec. 1990, pp. 954–959.

[9] A. I. Abu-El-Haija and A. M. Petersen, "An approach to eliminate roundoff errors in digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 195–198, 1979.

[10] D. Williamson and S. Sriharan, "Residue feedback in digital filters using fractional feedback coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 477–483, Apr. 1985.
[11] A. J. Laub, "Computation of balancing transformations," in *Proc. JACC*, San Francisco, CA, vol. 1, 1980.
[12] B. C. Moore, "Principal component analysis in linear systems: Controllability, observability and model reduction," *IEEE Trans. Automat. Control*, vol. AC-26, no. 1, pp. 17–32, Feb. 1981.
[13] L. M. Smith and B. W. Bomar, "An algorithm for constrained roundoff noise minimization in digital filters with application to two-dimensional filters," *IEEE Trans. Circuits Syst.*, vol. 35, no. 11, pp. 1359–1368, Nov. 1988.
[14] W. J. Lutz and S. L. Hakimi, "Design of multi-output systems with minimum sensitivity," *IEEE Trans. Circuits Syst.*, vol. 35, no. 9, pp. 1114–1123, Sept. 1988.
[15] G. Amit and U. Shaked, "Small roundoff noise realization of fixed-point digital filters and controllers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 880–891, June 1988.
[16] G. Li, B. D. O. Anderson, M. Gevers, and J. E. Perkins, "Optimal FWL design of state-space digital systems with weighted sensitivity minimization and sparseness consideration," *IEEE Trans. Circuits Syst.*, vol. 39, no. 5, pp. 365–377, May 1992.
[17] G. Li, "Finite precision aspects in the parametrization of control, estimation, and filtering problems," Ph.D. Dissertation, Université Catholique de Louvain, Nov. 1990.
[18] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. London: Springer-Verlag, Communication and Control Engineering Series, 1993.
[19] V. Tavšanoglu and L. Thiele, "Optimal design of state-space digital filter by simultaneous minimization of sensitivity," *IEEE Trans. Circuits Syst.*, vol. 35, no. 9, pp. 1114–1122, Sept. 1984.
[20] S. Y. Kung, "A new identification and model reduction algorithm via SVD," in *IEEE Proc. 12th Asilomar Conf. on Circuits, Syst., Comput.*, Pacific Grove, CA, 1978, pp. 705–714.
[21] R. Bellman, *Introduction to Matrix Analysis*, 2nd ed. New York: McGraw-Hill, 1970.

Gang Li received the B.S. degree in electrical engineering from Beijing Institute of Technology, Beijing, China, in 1982, and the M.S. and Ph.D. degrees both from Louvain University, Belgium, in 1988 and 1990, respectively.

In 1991 he was with the Control Group at Louvain University as a Postdoctoral Researcher. He is now a Lecturer at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include digital system design, optimal control and filtering, numerical problems in estimation and control theory, and digital signal processing.

Michel Gevers (S'66–M'72–SM'90–F'90) was born in Antwerp, Belgium, in 1945. He received the Electrical Engineering degree from Louvain University, Belgium, in 1968, and a Ph.D. degree from Stanford University, California, in 1972.

He is now Professor and Head of the Laboratoire d'Automatique, Dynamique et Analyse des Systèmes at Louvain University in Louvain la Neuve, Belgium. He has spent long-term visits in several universities, including the University of Newcastle, Australia, the Technical University of Vienna, and a three-year term at the Australian National University. His research interests are in system identification, adaptive estimation and control, multivariable system theory, optimal control and filtering, and the numerical aspects of filter and controller design.

Dr. Gevers is a Fellow of the Belgian American Educational Foundation. He has been Associate Editor of *Automatica* and the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and is presently Associate Editor of *Mathematics of Control, Signals, and Systems*. He is a coauthor with R. R. Bitmead and V. Wertz of *Adaptive Optimal Control—The Thinking Man's GPC*, published by Prentice Hall in 1990, and with G. Li of *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects*, to be published by Springer-Verlag (Communication and Control Engineering Series) in February 1993. He was awarded a Harkness Fellowship and an ESRO/NASA International Fellowship for his studies in the U.S.