P37

Optimal FWL Design of State-Space Digital Systems with Weighted Sensitivity Minimization and Sparseness Consideration

Gang Li, Brian D. O. Anderson, *Fellow, IEEE*, Michel Gevers, *Fellow, IEEE*, and Jane E. Perkins, *Student Member, IEEE*

Abstract—**The optimal finite word length (FWL) state-space digital system design problem is investigated. Instead of the usual sensitivity measure, it is argued that it may be desirable to minimize a frequency weighted sensitivity measure over all similarity transformations. The set of optimal realizations minimizing this weighted sensitivity is completely characterized, and an algorithm is proposed to find the optimal solution set. It is shown that a subset of the optimal realization set consists of sparse Schur realizations, whose actual sensitivity (taking into account the zero elements) is even smaller than the theoretical minimal sensitivity. Some nice properties of the Schur realizations are discussed. A numerical example that confirms the theoretical results is given.**

I. Introduction

MUCH attention has recently been paid to the finite word length (FWL) effects in digital system design. The optimal FWL state-space design has been considered as one of the most effective and elegant methods [1]–[5]. It is well known that any linear system can be represented by state-space equations, and that this state-space model is not unique. In the infinite precision case, all these realizations are equivalent since they yield one and the same transfer function. But different realizations have different numerical properties such as sensitivity and error propagation. This means that they are no longer equivalent in the finite precision case. The optimal FWL state-space design is to identify those realizations that minimize the degradation of the system performance due to the FWL effects. The deterioration of the performance of a realization of a digital filter due to the FWL effects can be separated into two components: one is due to the finite wordlength implementation of the coefficients of the filter, the other is due to roundoff of the signals after every arithmetical operation. The first effect is usually measured by a global sensitivity measure of the filter transfer function w.r.t. all the parameters [3]–[5], the other by the roundoff noise gain [1] and [2].

In [3], a global sensitivity measure of the transfer function w.r.t. the parameters of the state-space model was proposed by Tavsanoglu and Thiele, and a reasonable and easily computable upper bound for this measure was studied. It was shown in [5] that the realizations that minimize the upper bound also minimize the sensitivity measure itself and that, under a dynamic range constraint, this sensitivity measure and the roundoff noise gain are simultaneously optimized. The set of optimizing structures was characterized in [1]–[3] and [5].

In Tavsanoglu and Thiele's definition of sensitivity measure, the sensitivity behavior of a transfer function at one frequency point is considered to be as important as at another frequency point. From a practical point of view, we are usually interested in the performance of the transfer function within a specified frequency range—the bandwidth of the transfer function, for example. To achieve this, we define a weighted sensitivity function, and hence a corresponding measure in this paper. The optimal FWL design procedure for a frequency weighted sensitivity measure is given.

A frequency weighted measure has already been introduced by Thiele [5], but with a specific relationship between the weightings of the various terms of the measure. Under those constraints on the weightings, Thiele solved the sensitivity minimization problem using methods that are essentially the same as for the unweighted problem. Here we address the case of general unconstrained frequency weightings, and we solve the corresponding optimal state-variable design problem.

Optimal realizations are usually fully parametrized; see, e.g., [2] and [4]. In practice, it is desirable that the filter have a nice performance as well as a minimal number of coefficients to be implemented. Noting the fact that the optimal realizations minimizing this sensitivity measure are unique only up to an orthogonal similarity transformation, we further propose the use of Schur realizations within this class of optimal realizations. These Schur realizations have several advantages. They are sparse, and hence require fewer multi-

Manuscript received April 25, 1991; revised March 3, 1992. This work presents research results of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. This paper was recommended by Associate Editor J. Vandewalle.

G. Li and M. Gevers are with the Centre for Systems Engineering and Applied Mechanics, Universite Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium.

B. D. O. Anderson and J. E. Perkins are with the Department of Systems Engineering, Australian National University, Canberra ACT 2601, Australia.

IEEE Log Number 9200144.

plications. Some other useful properties of these realizations are discussed from a practical FWL implementation point of view.

This paper is organized as follows. In Section II we first set up the definitions of sensitivity functions and weighted sensitivity measure of a filter. The optimal FWL design problem is formulated. Our first new contribution is in Section III, where we study the optimal realization problem in terms of minimizing the frequency weighted sensitivity measure and establish the existence of optimal solutions. A recursive algorithm for solving the general minimization problem is given in Section IV, where an analytic solution is also given for the special case in which a proportionality relationship exists between certain weighted Gramians. Our second new contribution is in Section V: we propose the use of Schur realizations that belong to the optimal realization subset and are of a sparse form. An algorithm to search for further sparser realizations in this optimal realization set is also given. The Schur realizations are further investigated from a practical point of view in terms of pole sensitivity behavior. A numerical example is given in Section VI. Finally, some concluding remarks are given in Section VII.

II. Weighted Sensitivity Measure of a Realization

In this paper we consider the implementation of a discrete linear time-invariant single input, single output system having the following transfer function:

$$H(z) = \frac{\sum_{i=0}^{n} b_i z^{-i}}{1 + \sum_{i=1}^{n} a_i z^{-i}}. \tag{1}$$

This system can be implemented by a minimal state-space realization:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + du(t) \end{aligned} \tag{2}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n}$, $C^T \in \mathbb{R}^{n}$, and $d \in \mathbb{R}$. The transfer function can be expressed in terms of the state matrices as

$$H(z) = C(zI - A)^{-1}B + d. \tag{3}$$

We now define a realization set of S_H of this system as follows:

$$S_H = \{(A, B, C, d) : (A, B, C, d) \text{ satisfies (3)}\}.$$

Clearly, if (A, B, C, d) belongs to S_H, so does $(T^{-1}AT, T^{-1}B, CT, d)$ for any similarity transformation T. This means that S_H is an infinite set. In the infinite precision case, one realization is equivalent to another in this set since they all yield the same transfer function (3). The important point is, however, that different realizations have different numerical properties such as sensitivity to coefficient errors and propagation of signal roundoff errors. This signifies that different realizations behave differently in the finite precision case. In this sense they are no longer equivalent.

In practice it is impossible to realize the coefficients in (A, B, C, d) exactly due to FWL constraints. As a result, the actual system $H^*(z)$ has a transfer function given by (3) but with (A, B, C, d) replaced by their FWL version (A^*, B^*, C^*, d^*). Clearly, $H^*(z)$ and $H(z)$ will differ. In fixed-point implementation, the amount of this deviation can be measured by the sensitivity of the system transfer function $H(z)$ w.r.t. the coefficients of the matrices A, B, C, and d. Since different realizations have different sensitivities, as will be shown later, the optimal FWL state-space design is to search for those realizations that minimize the sensitivity in some proper measure.

There are of course several ways of defining an overall sensitivity measure. Here we present a measure proposed by Tavsanoglu and Thiele [3]. It is an absolute rather than relative sensitivity measure and is therefore based on a fixed-point arithmetical implementation; alternative floating-point implementations have been discussed in [6] and [7].

Definition 1: Let $M \in \mathbb{R}^{n\times m}$ be a matrix and let $f(M) \in \mathbb{C}$ be a scalar complex function of M, differentiable w.r.t. all the elements of M. We then define the sensitivity function of f w.r.t. M as

$$S_M \triangleq \frac{\partial f}{\partial M} \quad \text{with} \quad (S_M)_{ij} \triangleq \frac{\partial f}{\partial m_{ij}} \tag{4}$$

where m_{ij} denotes the (i, j)th element of the matrix M. ■

With these notations it is easy to show [3] that

$$\begin{aligned} S_A(z) &\triangleq \frac{\partial H(z)}{\partial A} = G(z)F^T(z) \\ S_B(z) &\triangleq \frac{\partial H(z)}{\partial B} = G(z) \\ S_C(z) &\triangleq \frac{\partial H(z)}{\partial C^T} = F(z) \end{aligned} \tag{5}$$

where

$$\begin{aligned} F(z) &\triangleq (zI - A)^{-1}B = \left[f_1(z) \cdots f_n(z)\right]^T \\ G^T(z) &\triangleq C(zI - A)^{-1} = \left[g_1(z) \cdots g_n(z)\right]. \end{aligned} \tag{6}$$

Note that the direct term d and the sensitivity function w.r.t. it are coordinate-independent, so they have nothing to do with the optimal realization problem, and hence they will be ignored in the subsequent analysis.

Definition 2: Let $f(z) \in \mathbb{C}^{n\times m}$ be any complex matrix valued function of the complex variable z. We then define the L_p-norm of $f(z)$ as

$$\|f\|_p \triangleq \left(\frac{1}{2\pi}\int_0^{2\pi} \|f(e^{j\omega})\|_F^p \, d\omega\right)^{1/p} \tag{7}$$

where $\|f(e^{j\omega})\|_F$ is the Frobenius norm of the matrix $f(e^{j\omega})$:

$$\begin{aligned} \|f(e^{j\omega})\|_F &\triangleq \left(\sum_{i=1}^{n}\sum_{k=1}^{m} |f_{ik}(e^{j\omega})|^2\right)^{1/2} \\ &= \left\{tr\left[f^T(e^{-j\omega})f(e^{j\omega})\right]\right\}^{1/2}. \end{aligned} \tag{8}$$

■

Tavsanoglu and Thiele [3] have proposed the following overall sensitivity measure of the transfer function $H(z)$ w.r.t. the parameters in the realization A, B, C:

$$M_a \triangleq \left\| \frac{\partial H}{\partial A} \right\|_1^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C^T} \right\|_2^2. \tag{9}$$

The mixing of different measures (L_1 and L_2) in the overall sensitivity measure above is motivated by the analytic properties of the first term on the right of (9), which allows one to derive an analytic minimization procedure for M_a; see [3] and [5]. The optimization of a more logical L_2 measure is much harder and has only recently been solved by the authors [19] and, independently, by Helmke and Moore [20].

Note that the measure in Definition 2 is in fact a frequency-independent mean value of a matrix function in the whole frequency range. Therefore, the sensitivity measure M_a defined in (9) considers the sensitivity behavior of the transfer function at one frequency point to be as important as at another frequency point. It is, however, usually the case that one is interested in the performance of the transfer function in a specified frequency band or even at some discrete frequency points. More precisely, one wants the transfer function to be less sensitive to the variations of the parameters in a certain frequency interval (the bandwidth, for example), and can allow a greater sensitivity in a frequency domain that one is not interested in. This observation leads to the definition of a weighted sensitivity and hence a weighted sensitivity measure.

Let $W_A(z)$, $W_B(z)$, and $W_C(z)$ be three integrable scalar functions of the complex variable z. Then the weighted sensitivity functions corresponding to those given in (5) are defined as

$$\begin{aligned} \frac{\delta H(z)}{\delta A} &\triangleq W_A(z) \frac{\partial H(z)}{\partial A} \\ \frac{\delta H(z)}{\delta B} &\triangleq W_B(z) \frac{\partial H(z)}{\partial B} \\ \frac{\delta H(z)}{\delta C^T} &\triangleq W_C(z) \frac{\partial H(z)}{\partial C^T}. \end{aligned} \tag{10}$$

Note that the notation is not meant to suggest that δ is a derivative operator. Now let

$$W_A(z) = W_1(z) W_2(z) \tag{11}$$

be any factorization of $W_A(z)$. With Definition 2, the overall weighted L_1/L_2 sensitivity measure is defined as

$$M_a^* \triangleq \left\| \frac{\delta H(z)}{\delta A} \right\|_1^2 + \left\| \frac{\delta H(z)}{\delta B} \right\|_2^2 + \left\| \frac{\delta H(z)}{\delta C^T} \right\|_2^2. \tag{12}$$

Now using (5), (10), and (11), M_a^* can be rewritten as

$$\begin{aligned} M_a^* = &\left\| W_1(z) G(z) \left(W_2(z) F(z) \right)^T \right\|_1^2 \\ &+ \left\| W_B(z) G(z) \right\|_2^2 + \left\| W_C(z) F(z) \right\|_2^2. \end{aligned} \tag{13}$$

A similarity transformation $x = Tz$ transforms $(A, B, C, F(z), G(z))$ into $(T^{-1}AT, T^{-1}B, CT, T^{-1}F(z), T^T G(z))$. This means that different realizations yield different sensitivity measures M_a^*. So an interesting problem is to find those realizations that minimize this sensitivity measure. The optimal FWL state-space design can then be formulated as follows:

$$\min_{(A, B, C) \in S_H} M_a^*. \tag{14}$$

In the next section, we will discuss how to solve the optimal FWL state-space design problem.

Comment: In [5], Thiele introduced a frequency weighted sensitivity measure similar to (12) but with

$$W_1(z) = W_B(z) \quad \text{and} \quad W_2(z) = W_C(z). \tag{15}$$

For this special choice of weighting functions, Thiele showed how to compute the weighted Gramians and solved a number of sensitivity minimization problems to which we shall return later. The choice (15) is justified for the case where colored noise with input spectrum $|W_C(z)|^2$ enters into the filter and where roundoff noise on the states is considered with a colored spectrum $|W_B(z)|^2$; see [5] for details. However, other choices can be motivated by other applications. To give but one example, consider the case where, in addition to minimizing the overall nonweighted sensitivity of the transfer function with respect to errors in the coefficients of A, B, C, it is desired to pay particular attention to the sensitivity of the poles of the realization with respect to errors in the coefficients of A. Note that the poles $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of A, so that the coefficients of B and C play no role in errors on the poles of the realization. In [5] Thiele proposed the following pole sensitivity measure

$$S_{A,\lambda} \triangleq \left\{ \sum_{k=1}^{n} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{\partial \lambda_k}{\partial a_{ij}} \right|^2 \right)^{1/2} \right\}^2. \tag{16}$$

For poles close to the unit circle, this measure can be approximated by

$$S_{A,\lambda} \approx \left\| \frac{\partial H(z)}{\partial A} \overline{W}_A(z) \right\|_1^2 \tag{17}$$

where

$$\overline{W}_A(z) = \sum_{k=1}^{n} \left| \frac{(1 - r_k)^2 D'(\lambda_k)}{N(\lambda_k)} \right| \delta(z - z_k).^1 \tag{18}$$

Here $N(z)$, $D(z)$ are defined by $H(z) \triangleq (N(z)/(D(z))$ and $D'(z) \triangleq (\partial D(z))/\partial z$, while r_k and z_k are defined by $\lambda_k = r_k e^{-j\omega_k}$ and $z_k = e^{-j\omega_k}$. In order to minimize the sensitivity of $H(z)$ with respect to the parameters of A, B, C with a particular emphasis on pole sensitivity, one might then

[1] Here δ denotes the Dirac function.

want to minimize the measure

$$S_A(z) = \left\| \frac{\partial H(z)}{\partial A} W_A(z) \right\|_1^2 + \left\| \frac{\partial H(z)}{\partial B} \right\|_2^2 + \left\| \frac{\partial H(z)}{\partial C} \right\|_2^2 \tag{19}$$

with $W_A(z) = 1 + \overline{W}_A(z)$. Note that in this case we do not have $W_A(z) = W_B(z)W_C(z)$ as in [5].

III. Optimal FWL Realizations

The difficulty in solving (14) is due to the fact that the first term on the right of (12) is a complicated function of the realization (A, B, C). To overcome this, note that by the Cauchy–Schwartz inequality

$$\left\| \frac{\delta H(z)}{\delta A} \right\|_1^2 = \| W_1(z)G(z)(W_2(z)F(z))^T \|_1^2 \le \| W_1(z)G(z) \|_2^2 \| W_2(z)F(z) \|_2^2 \tag{20}$$

where equality holds if and only if

$$\rho^2 G^H(z)G(z) \mid W_1(z) \mid^2 = F^H(z)F(z) \mid W_2(z) \mid^2 \quad \forall z \in \{ \mid z \mid = 1 \} \tag{21}$$

for some $\rho \neq 0 \in \Re$. We will study the following upper bound of M_a^*:

$$M_a^* \le \overline{M}_a^* \triangleq \| W_1(z)G(z) \|_2^2 \| W_2(z)F(z) \|_2^2 + \left\| \frac{\delta H(z)}{\delta B} \right\|_2^2 + \left\| \frac{\delta H(z)}{\delta C^T} \right\|_2^2. \tag{22}$$

We shall present methods for minimizing $\overline{M}_a^*$ and examine under which conditions realizations that minimize the upper bound $\overline{M}_a^*$ also minimize the measure itself, M_a^*. It is easy to show with (7) and (8) that

$$\overline{M}_a^* = tr(K_{o1})tr(K_{c2}) + tr(K_{oB}) + tr(K_{cC}) \tag{23}$$

where K_{o1}, K_{c2}, K_{oB}, and K_{cC} can be obtained by the following general expression:

$$K = \frac{1}{2\pi j} \oint_{\mid z \mid = 1} X(z) X^H(z) z^{-1} \, dz \tag{24}$$

with $X(z) = G(z)W_1(z)$, $F(z)W_2(z)$, $G(z)W_B(z)$, and $F(z)W_C(z)$, respectively. We call these four matrices K_{o1}, K_{c2}, K_{oB}, K_{cC} weighted Gramians. Several algorithms for computing a weighted Gramian are available in [5] and [8]. A similarity transformation $x = Tz$ transforms $(A, B, C, K_{cC}, K_{c2}, K_{oB}, K_{o1})$ into $(T^{-1}AT, T^{-1}B, CT, T^{-1}K_{cC}T^{-T}, T^{-1}K_{c2}T^{-T}, T^TK_{oB}T, T^TK_{o1}T)$. So, the optimal FWL design problem of (14) is replaced by the following upper bound minimization:

$$\min_{T:\det T \neq 0} \{ \overline{M}_a^* = tr(T^TK_{o1}T)tr(T^{-1}K_{c2}T^{-T}) + tr(T^TK_{oB}T) + tr(T^{-1}K_{cC}T^{-T}) \}. \tag{25}$$

Now, it is easy to see that

$$\overline{M}_a^* = tr(K_{o1}P)tr(K_{c2}P^{-1}) + tr(K_{oB}P) + tr(K_{cC}P^{-1}) \triangleq R(P) \tag{26}$$

where $P = TT^T$. Therefore,

$$\min_{T:\det T \neq 0} \overline{M}_a^* \iff \min_{P:\det P \neq 0} R(P). \tag{27}$$

In the remainder of this and the next section we shall study the minimization problem (27). Our developments will proceed as follows.

- First we show that a minimum of $R(P)$ exists and that it can only be achieved by nonsingular matrices P.
- We then show that this minimum is unique.
- We then proceed to the computation of the optimal solution P_{opt} by considering two different cases.
  We first consider the easier case where a proportionality relation exists between the weighted observability Gramians K_{o1} and K_{oB}, and similarly between the weighted controllability Gramians K_{c2} and K_{cC}. In this case, we shall compute an explicit solution of P_{opt}, and hence produce an explicit characterization of the optimal realization set.
  We then consider the general situation where no such relation exists. In such case, no explicit expression of the optimal solution can be given and an iterative algorithm is required for its computation.
- We shall also show that in the first case considered above, and for the special choice $W_1(z) = W_2(z)$, the optimal realization set minimizes not just the upper bound $\overline{M}_a^*$, but the sensitivity measure M_a^* itself.

Our first lemma shows that the minimum of $R(P)$ exists, and that it can be achieved by nonsingular P only. This means that (27) has solutions.

Lemma 1: With K_{oB} and K_{cC} nonsingular, the minimum of $R(P)$ defined in (26) exists and can be achieved only for nonsingular P.

Proof: By *SVD*, any (semi-)positive-definite matrix P can be decomposed into $P = \{p_{ij}\} = U^T\Sigma^2U$, where $U = \{u_{ij}\}$ is some orthogonal matrix and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_n)$, $\sigma_i \ge \sigma_{i+1} \ge 0$. So one has

$$p_{ij} = \sum_{k=1}^{n} u_{ki}u_{kj}\sigma_k^2 \qquad \forall i, j. \tag{28}$$

It is well known [16] that orthogonal matrices belong to a differentiable manifold of dimension $n(n-1)/2$. Therefore, locally, the elements u_{ij} of an orthogonal matrix U in $\Re^{n\times n}$ can be reparametrized as continuous functions of $n(n-1)/2$ parameters, i.e., of a vector

$$\bar{\theta} \triangleq (\theta_1\theta_2 \cdots \theta_N)^T \quad \text{with} \quad N = n(n-1)/2 \quad \text{for} \quad |\theta_l| \le \pi \quad \forall l. \tag{29}$$

Hence, p_{ij} is a continuous function of $\{\theta_i\}$ and $\{\sigma_k^2\}$ for all i and j. Now, we note that

$$R(P) = \sum_{i=1}^{n} m_{ii}(o1)\sigma_i^2 \sum_{i=1}^{n} m_{ii}(c2)\sigma_i^{-2} + \sum_{i=1}^{n} m_{ii}(oB)\sigma_i^2 + \sum_{i=1}^{n} m_{ii}(cC)\sigma_i^{-2}$$

where

$$M(o1) = UK_{o1}U^T \triangleq \{m_{ij}(o1)\}$$
$$M(c2) = UK_{c2}U^T \triangleq \{m_{ij}(c2)\}$$
$$M(oB) = UK_{oB}U^T \triangleq \{m_{ij}(oB)\}$$
$$M(cC) = UK_{cC}U^T \triangleq \{m_{ij}(cC)\}.$$

For any non-negative-definite matrix $M = \{m_{ij}\} \geq 0$, it is well known [15] that $\lambda_{\min}(M) \leq m_{ii}$ for $i = 1, 2, \cdots, n$, where $\lambda_{\min}(M)$ denotes the minimal eigenvalue of M. Clearly,

$$R(P) \geq m_{11}(oB)\sigma_1^2 \geq \lambda_{\min}[M(oB)]\sigma_1^2 = \lambda_{\min}[K_{oB}]\sigma_1^2 \tag{30}$$

and

$$R(P) \geq m_{nn}(cC)\sigma_n^{-2} \geq \lambda_{\min}[M(cC)]\sigma_n^{-2} = \lambda_{\min}[K_{cC}]\sigma_n^{-2}. \tag{31}$$

Also, for $P = I$, we get

$$R(I) = tr(K_{o1})tr(K_{c2}) + tr(K_{oB}) + tr(K_{cC}) \triangleq E_0 > 0.$$

Now choose any E in $\mathbb{R}_+$ such that $E \geq \max\{E_0, \lambda_{\min}(K_{oB}), \lambda_{\min}(K_{cC})\}$ and define

$$C_-^2(E) \triangleq \lambda_{\min}(K_{cC})/E \leq 1 \quad \text{and}$$
$$C_+^2(E) \triangleq E/\{\lambda_{\min}(K_{oB})\} \geq 1.$$

It then follows that $R(P) \geq E$ if either $\sigma_n^2 \leq C_-^2(E)$ or $\sigma_1^2 \geq C_+^2(E)$.

Since K_{oB} and K_{cC} are nonsingular by assumption and since E is finite, $C_-^2(E)$ and $C_+^2(E)$ as defined above satisfy $0 < C_-^2(E)$ and $C_+^2(E) < +\infty$. We now define the following closed set:

$$P_E \triangleq \{P: |\theta_i| \leq \pi, 0 < C_-^2(E) \leq \sigma_i^2 \leq C_+^2(E) < +\infty \quad i = 1, 2, \cdots, n\}. \tag{32}$$

For any P outside P_E (i.e., any P such that $\sigma_n^2 < C_-^2(E)$ or $\sigma_1^2 > C_+^2(E)$), we have $R(P) > E$. On the other hand, $P = I$ is in P_E with $R(I) \leq E$ and, using (31) and $\sigma_n^2 \leq C_+^2(E)$ together with the definition of $C_+^2(E)$ yields

$$R(P) \geq \frac{\lambda_{\min}(K_{oB})\lambda_{\min}(K_{cC})}{E} > 0 \quad \forall P \text{ in } P_E. \tag{33}$$

Since $R(P)$ is a continuous function of P in the closed set P_E with a strictly positive lower bound, it follows that $R(P)$ has its global minimum within this set, and since all the elements of P_E are nonsingular and bounded by the assumption on K_{oB} and K_{cC}, any element P for which $R(P)$ achieves the global minimum is nonsingular. ■

We have proved that (27) has solutions if K_{oB} and K_{cC} are both nonsingular. This condition is satisfied if the weighting functions $W_B(z)$ and $W_C(z)$ have no pole-zero cancellations with the system $H(z)$, i.e., if the system (A, B, C) is minimal and if the scalar weighting functions have no zeros at the poles of $H(z)$. In the sequel, this condition is assumed to be satisfied.

We shall now prove the uniqueness of the minimizing solution. To do so, we need the following lemma.

Lemma 2: Let M and X be two matrices of appropriate dimension; then

$$1) \quad \frac{d[tr(MX)]}{dX} = M^T$$
$$2) \quad \frac{d[tr(MX^{-1}]}{dX} = -(X^{-1}MX^{-1})^T.$$

Proof: The proof is fairly easy and can be found in [17]. ■

With Lemma 2, one has

$$\frac{\partial R(P)}{\partial P} = -tr(K_{o1}P)P^{-1}K_{c2}P^{-1} - P^{-1}K_{cC}P^{-1} + tr(K_{c2}P^{-1})K_{o1} + K_{oB}. \tag{34}$$

By letting $\partial R(P)/\partial P = 0$, one gets a necessary condition which the solution of (27) must satisfy

$$P[tr(K_{c2}P^{-1})K_{o1} + K_{oB}]P = tr(K_{o1}P)K_{c2} + K_{cC}. \tag{35}$$

Our next result shows the uniqueness of the solution of (35).

Theorem 1: With the four symmetric positive-definite matrices K_{o1}, K_{c2}, K_{oB} and K_{cC} defined by (24), (35) has a unique solution, and hence so does (27).

Proof: Let P_0 and P be two solutions of (35). From

$$P_0[tr(K_{c2}P_0^{-1})K_{o1} + K_{oB}]P_0 = tr(K_{o1}P_0)K_{c2} + K_{cC}$$

it follows that

$$I[tr(\tilde{K}_{c2}I^{-1})\tilde{K}_{o1} + \tilde{K}_{oB}]I = tr(\tilde{K}_{o1}I)\tilde{K}_{c2} + \tilde{K}_{cC}$$

where

$$\tilde{K}_{o1} = P_0^{1/2}K_{o1}P_0^{1/2}, \quad \tilde{K}_{c2} = P_0^{-1/2}K_{c2}P_0^{-1/2},$$
$$\tilde{K}_{oB} = P_0^{1/2}K_{oB}P_0^{1/2}, \quad \tilde{K}_{cC} = P_0^{-1/2}K_{cC}P_0^{-1/2}.$$

This means that by proper choice of the initial realization, (35) has the unit matrix as a solution. So, without loss of generality, one can assume $P_0 = I$. Therefore, one only needs to prove that if P is a solution of (35), then $P = I$ under the following constraint:

$$tr(K_{c2})K_{o1} + K_{oB} = tr(K_{o1})K_{c2} + K_{cC} \tag{36}$$

or equivalently,

$$tr(\tilde{K}_{c2})\tilde{K}_{o1} + \tilde{K}_{oB} = tr(\tilde{K}_{o1})\tilde{K}_{c2} + \tilde{K}_{cC} \quad (37)$$

where

$$\tilde{K}_{o1} = U^T K_{o1} U, \quad \tilde{K}_{c2} = U^T K_{c2} U,$$
$$\tilde{K}_{oB} = U^T K_{oB} U, \quad \tilde{K}_{cC} = U^T K_{cC} U \quad (38)$$

for an arbitrary orthogonal matrix U. In particular, one has

$$\left(\sum_{i=1}^{n} \tilde{K}_{c2}(i,i)\right)\tilde{K}_{o1}(j,j) + \tilde{K}_{oB}(j,j) = \left(\sum_{i=1}^{n} \tilde{K}_{o1}(i,i)\right)\tilde{K}_{c2}(j,j) + \tilde{K}_{cC}(j,j) \quad (39)$$

for all $j = 1, 2, \cdots, n$. Now, by SVD one has $P = UX^2U^T$ where U is some orthogonal matrix and $X^2 = \text{diag}(x_1^2, x_2^2, \cdots, x_n^2)$ with $x_1^2 \geq x_2^2 \geq \cdots \geq x_n^2 > 0$. Inserting $P = UX^2U^T$ in (35) and using (38) yields

$$\left\{\left(\sum_{i=1}^{n} \tilde{K}_{c2}(i,i)x_i^{-2}\right)\tilde{K}_{o1}(j,j) + \tilde{K}_{oB}(j,j)\right\}x_j^4 = \left(\sum_{i=1}^{n} \tilde{K}_{o1}(i,i)x_i^2\right)\tilde{K}_{c2}(j,j) + \tilde{K}_{cC}(j,j)$$

or

$$\left\{\left(\sum_{i=1}^{n} \tilde{K}_{c2}(i,i)x_i^{-2}\right)\tilde{K}_{o1}(j,j) + \tilde{K}_{oB}(j,j)\right\}x_j^2 = \left\{\left(\sum_{i=1}^{n} \tilde{K}_{o1}(i,i)x_i^2\right)\tilde{K}_{c2}(j,j) + \tilde{K}_{cC}(j,j)\right\}x_j^{-2} \quad (40)$$

for all $j = 1, 2, \cdots, n$. On the one hand, for $j = 1$, it follows from (40) that

$$\left(\sum_{i=1}^{n} \tilde{K}_{c2}(i,i)\right)\tilde{K}_{o1}(1,1) + \tilde{K}_{oB}(1,1)x_1^2 \leq \left(\sum_{i=1}^{n} \tilde{K}_{o1}(i,i)\right)\tilde{K}_{c2}(1,1) + \tilde{K}_{cC}(1,1)x_1^{-2}$$

since $x_1^2 \geq x_2^2 \geq \cdots \geq x_n^2 > 0$. On the other hand, taking $j = 1$ and the same U as in (38) it follows from (39) that

$$\left(\sum_{i=1}^{n} \tilde{K}_{c2}(i,i)\right)\tilde{K}_{o1}(1,1) + \tilde{K}_{oB}(1,1) = \left(\sum_{i=1}^{n} \tilde{K}_{o1}(i,i)\right)\tilde{K}_{c2}(1,1) + \tilde{K}_{cC}(1,1).$$

One concludes that $x_1^2 \leq 1$. Similarly, for $j = n$ one can obtain

$$\left(\sum_{i=1}^{n} \tilde{K}_{c2}(i,i)\right)\tilde{K}_{o1}(n,n) + \tilde{K}_{oB}(n,n)x_n^2 \geq \left(\sum_{i=1}^{n} \tilde{K}_{o1}(i,i)\right)\tilde{K}_{c2}(n,n) + \tilde{K}_{cC}(n,n)x_n^{-2}$$

and (39) yields

$$\left(\sum_{i=1}^{n} \tilde{K}_{c2}(i,i)\right)\tilde{K}_{o1}(n,n) + \tilde{K}_{oB}(n,n) = \left(\sum_{i=1}^{n} \tilde{K}_{o1}(i,i)\right)\tilde{K}_{c2}(n,n) + \tilde{K}_{cC}(n,n).$$

This implies that $x_n^2 \geq 1$. Since $x_1^2 \geq x_n^2$, one has $x_1^2 = x_2^2 = \cdots = x_n^2 = 1$, which leads to $P = I$. This completes the proof. ■

IV. Computation of the Optimal Realization Set

It appears difficult to find an explicit expression of the solution P of (35) when no particular relation exists between the frequency weightings. We will show later that in this general case P_{opt} can be computed as the limiting solution of a gradient algorithm. But first we show that an analytic solution of (35) can be computed for the case where

$$K_{oB} = \varrho_1 K_{o1}, \quad K_{cC} = \varrho_2 K_{c2} \quad (41)$$

with ϱ_1 and ϱ_2 two positive constants. We note that one particular case where this relationship holds is when $W_1(z) = W_B(z)$ and $W_2(z) = W_C(z)$ as in [5]; in that particular case, $\varrho_1 = \varrho_2 = 1$. To compute P_{opt} when (41) holds we need the following lemma.

Lemma 3: Let $W > 0$ and $M \geq 0$ be symmetric. The equation $PWP = M$ has a unique solution $P = P^T \geq 0$ and the solution is given by $P = W^{-1/2}[W^{1/2}MW^{1/2}]^{1/2}W^{-1/2}$, where for any $X \geq 0$, $X^{1/2}$ denotes the unique symmetric matrix satisfying $X^{1/2} \geq 0$ and $X^{1/2}X^{1/2} = X$.

Proof: Let $W^{1/2}$ be a positive-definite square root of W. Clearly, this square root is unique [15]. Then

$$PWP = M \Leftrightarrow W^{1/2}PW^{1/2}W^{1/2}PW^{1/2} = W^{1/2}MW^{1/2} \Leftrightarrow W^{1/2}PW^{1/2} = [W^{1/2}MW^{1/2}]^{1/2}$$

which leads to

$$P = W^{-1/2}[W^{1/2}MW^{1/2}]^{1/2}W^{-1/2}.$$

Evidently, P is unique. ■

Theorem 2: With four symmetric positive-definite matrices K_{o1}, K_{oB}, K_{c2}, and K_{cC} satisfying (41), there exists a unique solution P of (27) which is given by

$$P = \rho K_{o1}^{-1/2}[K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}]^{1/2}K_{o1}^{-1/2} \quad (42)$$

where

$$\varrho = (\varrho_2/\varrho_1)^{1/2}. \quad (43)$$

In addition, the optimal solutions of the optimization problem (25) are given by

$$T = \varrho^{1/2}K_{o1}^{-1/2}[K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}]^{1/4}V \quad (44)$$

where V is an arbitrary orthogonal matrix. The Gramians of the optimal realizations are characterized by

$$\tilde{K}_{o1} = \varrho^2\tilde{K}_{c2}. \quad (45)$$

Proof: With (35) and (41), it follows from Lemma 3 that P is given by (42), where ϱ is given by

$$\varrho = \sqrt{\frac{tr(K_{o1}P) + \varrho_2}{tr(K_{c2}P^{-1}) + \varrho_1}} = \sqrt{\frac{\varrho tr\left[\left(K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}\right)^{1/2}\right] + \varrho_2}{\varrho^{-1}tr\left[\left(K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}\right)^{1/2}\right] + \varrho_1}}$$

which yields (43). The uniqueness is evident. Since the minimum of $R(P)$ exists and since (35) is a necessary condition for achieving this minimum, (42) and (43) are the unique solution of (27) for the case (41). To obtain the set of optimal transformations T, note that P can be factored as

$$P = \varrho K_{o1}^{-1/2}\left[K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}\right]^{1/2}K_{o1}^{-1/2} = \left\{\varrho^{1/2}K_{o1}^{-1/2}\left[K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}\right]^{1/4}V\right\} \cdot \left\{\varrho^{1/2}K_{o1}^{-1/2}\left[K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}\right]^{1/4}V\right\}^T$$

where V is an arbitrary orthogonal matrix. Expression (44) then follows from $P = TT^T$. We can now compute the Gramians of the optimal realizations:

$$\tilde{K}_{o1} = T^TK_{o1}T = \varrho V^T\left(K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}\right)^{1/2}V$$

$$\tilde{K}_{c2} = T^{-1}K_{c2}T^{-T} = \varrho^{-1}V^T\left(K_{o1}^{1/2}K_{c2}K_{o1}^{1/2}\right)^{1/2}V$$

and hence

$$\tilde{K}_{o1} = \varrho^2\tilde{K}_{c2}. \tag{46}$$

Of course, by the relation (41), we also have

$$\tilde{K}_{oB} = \varrho_1\tilde{K}_{o1} \quad \text{and} \quad \tilde{K}_{cC} = \varrho_2\tilde{K}_{c2}. \tag{47}$$

Next we show that if $W_1(z) = W_2(z)$ holds in addition to (41), then the sensitivity measure itself is also minimized by the optimal realization set characterized by (46).

Corollary 1: Assume that, in addition to the relation (41) between the respective weighted Gramians, the two factors of the frequency weighting $W_A(z)$ are identical, i.e., $W_1(z) = W_2(z)$ $\forall z$. Then the realizations characterized by (46) minimize the frequency weighted sensitivity measure (12).

Proof: With the additional condition $W_1(z) = W_2(z)$, the optimality condition (46) can be written:

$$\frac{1}{2\pi j}\oint_{|z|=1}\tilde{G}(z)\tilde{G}^H(z)\,|W_1(z)|^2z^{-1}\,dz = \frac{1}{2\pi j}\oint_{|z|=1}\varrho^2\tilde{F}(z)\tilde{F}^H(z)\,|W_1(z)|^2z^{-1}\,dz. \tag{48}$$

It then immediately follows from [5, lemma 2] that

$$\tilde{G}^H(z)\tilde{G}(z) = \varrho^2\tilde{F}^H(z)\tilde{F}(z) \quad \forall z \in \{|z| = 1\}. \tag{49}$$

Therefore, the Cauchy–Schwartz inequality is satisfied with equality in (20), and the result follows by the same argument as in [5]. ■

We now turn to the general case where the relation (41) does not hold. In such a case, an explicit expression of the solution of (35) does not appear to be at hand. However, P can be obtained by an iterative procedure using a gradient algorithm:

$$P(k+1) = P(k) - \mu\left.\frac{\partial R(P)}{\partial P}\right|_{P=P(k)} \tag{50}$$

where $\partial R(P)/\partial P$ is given by (34) and μ is a positive step size. We have proved above that the function $R(P)$ has a unique (and hence global) minimum achieved by a nonsingular P. Therefore, the above algorithm will converge to P_{opt} for any positive definite initial condition.

As with almost any numerical minimization algorithm, applying to a problem for which no analytic solution is available, no hard and fast rules on the step size choice can be given. Our arguments have shown that the surface near the minimum is approximately quadratic, which gives a slight insight into the rate of convergence.

The choice of an appropriate initial condition will improve the convergence of the algorithm (50). We note that a necessary condition for (35) is $tr(K_{oB}P) = tr(K_{cC}P^{-1})$. So we use as initial condition a P_0 that minimizes $tr(K_{o1}P)tr(K_{c2}P^{-1})$ and ensures $tr(K_{oB}P) = tr(K_{cC}P^{-1})$ at the same time. We know that the P_0 minimizing the above trace product is not unique. In fact, suppose a P_1 has been found that minimizes this product; then so does kP_1 for any $k > 0$. Clearly, $P_0 = kP_1$ with

$$k^2 = tr(K_{cC}P_1^{-1})/tr(K_{oB}P_1) \tag{51}$$

will minimize $tr(K_{o1}P)tr(K_{c2}P^{-1})$ while at the same time producing $tr(K_{oB}P) = tr(K_{cC}P^{-1})$. It can be shown [8] that such a P_1 can be chosen to be $T_bT_b^T$ where T_b internally balances K_{o1} and K_{c2}, that is,

$$T_b^{-1}K_{c2}T_b^{-T} = T_b^TK_{o1}T_b \triangleq \text{diag}\,(\eta_1, \eta_2, \cdots, \eta_n).$$

This transformation matrix T_b can be obtained using a numerically well-conditioned algorithm due to Laub [10].

Finally, we note that for any optimal $P = TT^T$, the corresponding optimal transformation matrices can be constructed as

$$T = P^{1/2}V \tag{52}$$

for any orthogonal matrix V. All the arguments above can be summarized by the following theorem.

Theorem 3: The optimal transformation matrices, that is, the solutions of (25), are not unique and can be characterized by (52) where P is determined by the system and the weighting functions (it is the unique solution of (35)) while V is an arbitrary orthogonal matrix.

This means that there is a degree of freedom characterized by the set of orthogonal matrices V in this optimal transformation set. In the next section we will see how to exploit this freedom in order to simplify the implementation and to improve the computational performance of the system.

Comment: As we stated above, Thiele [5] has also considered the frequency weighted sensitivity measure (12) but only in the special case where $W_1(z) = W_B(z)$ and $W_2(z)$

$= W_C(z)$. He showed how to minimize the upper bound $\overline{M}_a^*$ for this case. He also showed that if, in addition, $W_1(z) = W_2(z) = W_B(z) = W_C(z)$, then the realizations that minimize the upper bound also minimize the weighted sensitivity measure itself. Our results extend Thiele's result in several ways. First we have explicitly characterized the realizations that minimize the upper bound in the case where the relation (15) is replaced by the weaker relation (41). Secondly, we have shown that if, in addition to (41), the factors of $W_A(z)$ are chosen to be identical, $W_1(z) = W_2(z)$, then this optimal realizaton set also minimizes the sensitivity measure M_a^* itself. Finally, we have shown how to compute a realization set that minimizes the upper bound even in the most general case where no special constraints hold between any of the frequency weightings or frequency weighted Gramians.

V. Schur Realizations in Digital Filter Design

From Theorem 3, one can see that the realizations determined by (25) form an optimal realization subset. This means that there exist some degrees of freedom in this equivalence subset. In [4], this freedom was used to find a Hessenberg realization in order to reduce the number of components to be implemented. Here, we investigate another realization, called Schur realization, which has some nice properties. We first define what we call a Schur realization.

Definition 3: Let $(A_s, B_s, C_s) \in S_H$ be a realization of $H(z)$. This realization is called Schur realization if and only if the matrix A_s is of the following real Schur form:

$$A_s = \begin{pmatrix} A_{11} & x & \cdot & \cdot & x & \cdot & x \\ 0 & A_{22} & \cdot & \cdot & x & \cdot & x \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & A_{ii} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 0 & \cdot & A_{mm} \end{pmatrix}$$

where each A_{ii} is either a real number or a real 2×2 matrix having complex conjugate eigenvalues.

Since for any matrix $M \in \mathbb{R}^{n \times n}$ there exists at least one orthogonal matrix U such that $U^T M U$ is of the real Schur form (see [11]), each realization can be transformed into a Schur realization by an orthogonal similarity transformation. In other words, there exists a subset of S_H, consisting of all the Schur realizations.

Comments:

1) A nice property of the Schur realization is its sparseness: A_s has at least $(1/2)n(n-1) - p$ zero elements where n is the dimension of A_s and p is the number of diagonal block-matrices of dimension 2×2 in A_s. This number can be increased by further orthogonal similarity transformations as we will see later.
2) We note that different realizations in this Schur realization subset have different numerical properties. It is well known that an orthogonal similarity transformation usually keeps the numerical properties of the realizations unchanged [12]. Therefore, it would be interesting to transform the realizations that already have some desirable numerical properties, such as the generically fully parametrized optimal realizations defined in Theorem 3, into a Schur realization by an orthogonal similarity transformation.

5.1. Sensitivity Consideration and Sparseness Improvement

In Section IV we have shown (see Theorem 3) that any two optimal realizations are related by an orthogonal similarity transformation. So from the preceding discussions, one can see that any non-Schur realization in this optimal realization subset can be transformed into a Schur realization with an orthogonal similarity matrix. This Schur realization evidently keeps the minimal sensitivity property. From a practical point of view, a general (fully parametrized) optimal realization and its corresponding Schur realization can have different sensitivity behaviors even though they yield the same theoretical minimal sensitivity measure value. Indeed, in the definition of sensitivity (see (12) and Definition 1), it has been assumed that any parameter in a realization (A, B, C) is allowed to vary and the sensitivity of the system transfer function w.r.t. every one of these parameters is taken into account in the overall sensitivity function and hence in the overall sensitivity measure. This can not exactly describe what happens in a real implementation procedure. In fact, whatever bit number that is used, parameters such as 0 and ± 1 can be exactly implemented. (Here, the parameters to be implemented are assumed to be normalized such that they are absolutely smaller than or equal to 1.) It is reasonable that the actual overall sensitivity measure should not take into account the sensitivity of the transfer function w.r.t. these trivial parameters. Keeping this in mind, one can argue that the Schur realizations will give a better actual sensitivity performance than the corresponding (fully parametrized) optimal realizations. We conclude that equivalent optimal realizations having the same theoretical minimal sensitivity measure could yield different actual sensitivity behaviors.

It follows from the above discussion that it is strongly desirable to find realizations, in the optimal realization subset, that have as many trivial parameters (i.e., 0 and ± 1) as possible. Since the optimal realizations are related by orthogonal similarity transformation matrices and since any unknown orthogonal matrix of order n has only $(1/2)n(n-1)$ degrees of freedom, generally speaking the sparsest realizations have at most that many trivial parameters (see the system Hessenberg forms in [4]). Now we will show that there exists an optimal Schur realization (A_s, B_s, C_s), in the optimal realization subset, that has at least $(1/2)n(n-1)$ zero parameters.

Theorem 4: For any transfer function of McMillan degree n, there exists a Schur realization (A_s, B_s, C_s) that belongs to the sensitivity optimal realization subset and has at least $(1/2)n(n-1)$ zero parameters.

Proof: First, from any optimal minimal realization, one can always find a Schur realization, which is obtained from this optimal realization by an orthogonal similarity transfor-

nation. So this Schur realization is optimal. Denote

$$U = \text{diag}\,(U_{11}, U_{22}, \cdots, U_{mm}) \tag{53}$$

where U_{ii} is either a 1×1 orthogonal matrix (i.e., ± 1) or a 2×2 orthogonal matrix, having the same dimension as A_{ii} in A_s. Let p be the number of 2×2 blocks. Notice that a similarity transformation by the matrix U will change neither the optimality nor the Schur structure of the realization, that is, if (A_s, B_s, C_s) is an optimal Schur realization, so is $(U^T A_s U, U^T B_s, C_s U)$. So this realization in the new coordinate system has at least $(1/2)(n-1)n - p$ zeros in $U^T A_s U$. Note that

$$U^T B_s = \left[B_1^T U_{11}, B_2^T U_{22}, \cdots, B_i^T U_{ii}, \cdots, B_m^T U_{mm}\right]^T. \tag{54}$$

Let A_{ii} be a block of dimension 2×2 in A_s. Then we denote

$$U_{ii} = \begin{pmatrix} \cos\delta_i & \sin\delta_i \\ -\sin\delta_i & \cos\delta_i \end{pmatrix}$$

and $B_i = [b_{1i}, b_{2i}]^T$. One can see that with the following choice of δ_i:

$$\delta_i = -\tan^{-1}\left(\frac{b_{2i}}{b_{1i}}\right)$$

the first element of $U_{ii}^T B_i$ will be zero. Therefore, with a series of p such 2×2 orthogonal matrices U_{ii}, the matrix B will have the following form:

$$B = U^T B_s = [0, x, 0, x, \cdots, 0, x]^T$$

which means that the realization in the new coordinate system will have at least $(1/2)n(n-1)$ zero parameters. ■

Comments:

1) In the same way, we can alternatively make matrix C, instead of B, have the same sparse form as above.
2) Besides the better sensitivity performance, the optimal Schur realization reduces the complexity of the implementation because of the sparseness of this realization. As a result, the processing will be faster.
3) We have shown that there exist optimal realizations that have at least $(1/2)n(n-1)$ zero parameters. This number corresponds precisely to the number of degrees of freedom in any orthogonal matrix. Clearly, the Schur realizations are not the unique sparse optimal realizations. We could place the $(1/2)n(n-1)$ trivial parameters in other positions. So an interesting problem is how to choose these positions in order to get the best sensitivity performance. This is a difficult problem since the actual sensitivity measure (i.e., discounting the sensitivity contribution of the trivial (i.e., fixed) parameters) is not tractable, and hence the minimization of this measure with these $(1/2)n(n-1)$ degrees of freedom is very hard, and we will not pursue this further.
4) In [18], Iwatsuki *et al.* have developed another kind of sparse optimal realizations using this freedom. Their development is based on a symmetrical property of the balanced realization, which is optimal only for the cases without frequency weighting. In this sense, our optimal Schur realizations are more general than the structures studied in [18].
5) Another approach to the realization of a digital filter comes from embedding it in an orthogonal filter [21]. This offers the advantage of guaranteed freedom from overflow, and automatic scaling. The noise gain is low, though not normally optimum. Sensitivity is not optimum, but may be attractive—at least in the nonfrequency weighted case. It seems improbable that one could make any systematic statement comparing the frequency-weighted sensitivity of such filters with the optimum structures of this paper.

5.2. Pole Sensitivity of Schur Realizations

Now we turn to the topic of how to implement some special parameters that seriously affect the performance of the system. In the preceding discussion, we mentioned that it is quite difficult to minimize the actual sensitivity measure w.r.t. the $(1/2)n(n-1)$ degrees of freedom in the optimal realization subset. We note that this measure is the average of a sensitivity function over the whole frequency domain. The frequency response characteristics of a filter are determined by its pole-zero positions. This is why in digital filter design, the pole and zero behaviors are also taken as an important design criterion. In [13], Mantey defined a pole sensitivity measure for a state-space realization and argued that in order to minimize this pole sensitivity measure the poles have to be implemented directly (in block diagonal forms). But this form evidently may yield a poor performance in terms of the sensitivity measure studied in this paper, because generally it does not belong to the optimal realization subset. In [14] Williamson defined a global pole sensitivity measure of a realization (A, B, C) of a filter $H(z)$ as

$$M_\lambda = \sum_{i=1}^{n} M_{\lambda_i} \tag{55}$$

with $M_{\lambda_i} = \|\partial\lambda_i / \partial A\|_F^2$ and λ_i the ith eigenvalue of A.

He showed that the pole sensitivity measure depends strongly on the chosen realization, that the partial sensitivity measure M_{λ_i} of the pole λ_i is larger than or equal to one, and that for a filter of order n the minimal value, n, of the global pole sensitivity M_λ is achieved if and only if the matrix A is normal [14].

An outstanding property of a Schur realization is that its poles are determined only by the main block diagonal elements of A_s. This property allows one to analyze its pole sensitivity behavior easily. By applying an orthogonal similarity transformation such as (53) to any optimal Schur realization, the *actual* pole sensitivity performance can be improved even though this orthogonal matrix will not change the *theoretical* pole sensitivity measure of the realization. We shall denote by Ψ the actual global pole sensitivity; in contrast to (55), the partial derivatives of the eigenvalues with respect to the fixed parameters of A are not taken into account in the definition of Ψ. Now, we first study the pole sensitivity of a Schur realization.

For any real pole λ_k corresponding to a 1×1 block matrix in A_s, it is easy to see that the *actual* partial sensitivity measure is

$$\Psi_{\lambda_k} \triangleq \left\| \frac{\partial \lambda_k}{\partial A_s} \right\|_F^2 = 1. \tag{56}$$

Indeed, the zero elements under the diagonal of A_s will have no implementation error at all, and hence this real pole will be unaffected by elements other than itself (since it is one of the diagonal elements of A_s). This means that for real poles the actual partial pole sensitivity measure of a Schur realization reaches the minimal possible value. Hence if A_s has all real eigenvalues, the actual global pole sensitivity measure $\Psi = n$ even though A_s is not normal.

Consider now a pair of complex conjugate poles corresponding to a 2×2 block matrix in A_s. Let $M = \{m_{ij}\}$ be a 2×2 real matrix, say a diagonal block matrix of A_s, having complex conjugate eigenvalues λ_1 and λ_2. Then the following condition is satisfied:

$$\Delta^2 = -4m_{12}m_{21} - (m_{11} - m_{22})^2 > 0. \tag{57}$$

By direct computation, we have the following sensitivity measures for $k = 1, 2$:

$$\left| \frac{\partial \lambda_k}{\partial m_{11}} \right|^2 = \left| \frac{\partial \lambda_k}{\partial m_{22}} \right|^2 = \frac{1}{4}\left[1 + \frac{(m_{11} - m_{22})^2}{\Delta^2} \right]$$

$$\left| \frac{\partial \lambda_k}{\partial m_{12}} \right|^2 = \frac{m_{21}^2}{\Delta^2}, \quad \left| \frac{\partial \lambda_k}{\partial m_{21}} \right|^2 = \frac{m_{12}^2}{\Delta^2}. \tag{58}$$

For this pair of complex poles the actual partial sensitivity measure with respect to M is given by

$$\Psi_{\lambda_1, \lambda_2} = \sum_{k=1}^{2} \left\| \frac{\partial \lambda_k}{\partial M} \right\|_F^2 = \frac{(m_{12} - m_{21})^2}{\Delta^2}. \tag{59}$$

Remark: For $m_{11} = m_{22}$, $m_{12} = -m_{21}$, the above partial sensitivity measure is 2. This means that the actual eigenvalue sensitivity measure achieves its minimum for a block matrix M having that structure.

Denote

$$U = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \quad \theta \in [0, \pi]. \tag{60}$$

Then,

$$R = U^T M U = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}$$

where

$$\begin{aligned} r_{11} &= m_{11}\cos^2\theta - (m_{12} + m_{21})\sin\theta\cos\theta + m_{22}\sin^2\theta \\ r_{12} &= m_{12}\cos^2\theta + (m_{11} - m_{22})\sin\theta\cos\theta - m_{21}\sin^2\theta \\ r_{21} &= m_{21}\cos^2\theta + (m_{11} - m_{22})\sin\theta\cos\theta - m_{12}\sin^2\theta \\ r_{22} &= m_{22}\cos^2\theta + (m_{12} + m_{21})\sin\theta\cos\theta + m_{11}\sin^2\theta. \end{aligned} \tag{61}$$

From (61) one can see that with a proper choice of θ, it is always possible to change one of the four r_{ij} to some desired value lying between bounds determined by the m_{ij}. Observing (58), we can see that the closer the two diagonal elements are, the more insensitive the measure is w.r.t. these elements because Δ^2 is unchanged by unitary transformation. So one can choose θ such that $r_{11} = r_{22}$, which makes $\partial\lambda_k / \partial r_{jj} = 1/4$ for $k, j = 1, 2$ (see (58)). This can be achieved by letting

$$\theta = \tan^{-1}\left[2\frac{m_{11} - m_{22}}{m_{12} + m_{21}} \right]. \tag{62}$$

Since the matrices M and R are both fully parametrized, the actual pole sensitivity measure is identical to the theoretical one. Hence, this reduction of the pole sensitivity measure with respect to the diagonal elements will correspondingly increase the measure with respect to the two off-diagonal elements, r_{12} and r_{21}, since the orthogonal transformation U does not change the theoretical pole sensitivity measure. So, what have we gained? The idea is that for a given bit number, B_c, for coefficient implementation, one can choose an FWL number v^* that is as close as possible to the off-diagonal element, say r_{21}, to which the pole sensitivity is highest. (This implies $r_{12}^2 > r_{21}^2$; see (58).) Using (61) one can then find a θ^* (and hence a U via (60)) such that this off-diagonal element is exactly equal to v^*. As a result, this parameter r_{21} having the highest sensitivity is implemented exactly with FWL, while the two diagonal elements r_{11} and r_{22} have near minimal sensitivity. Performing a similar orthogonal transformation for each diagonal block will improve the actual pole sensitivity behavior.

VI. Numerical Example

We now illustrate our previous theoretical results with a sixth-order narrow-band low-pass filter with a normalized sampling frequency $f_s = 1$, and filter design parameters $f_p = 0.03125$ (passband frequency), $F_s = 0.0390625$ (stopband frequency), and $\epsilon_p = 1$ dB (passband ripple). The upper limit of the frequency response is 0 dB, while the stopband attenuation is 46.68 dB. This example has been used in [14].

We present this filter in its controllable realization R_c:

$$A_c = \begin{pmatrix} 5.6526 & -13.3818 & 16.9792 & -12.1765 & 4.6789 & -0.7526 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$B_c = (1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)^T$$

$$C_c = (0.1511 \quad -0.4558 \quad 0.3855 \quad 0.1116 \quad -0.3074 \quad 0.1165) \times 10^{-2}$$

and $d = 0.4708 \times 10^{-2}$.

Taking this realization as the initial realization, we can compute the corresponding Gramians without weighting (that is $W_1(z) = W_2(z) = W_B(z) = W_C(z) = 1$). In this case, one optimal realization that minimizes the sensitivity measure is the balanced form R_b characterized by $K_{c2} = K_{cC} = K_{o1} = K_{oB}$ = diagonal.

6.1. Choices of Weighting Functions

As said before, the basic idea in using the weighting functions is to emphasize the behavior of the transfer function in some frequency domain of interest. So the choice of weighting functions depends completely on the specifications imposed on the implemented filter.

Taking $W_1(z) = W_B(z) = W_C(z) = H(z)$ and $W_2(z) = 1$ with $H(z)$ the transfer function of the filter itself, we can get the unique solution P of (35) by using the algorithm (50) and hence the optimal similarity transformation matrices T constructed by (52). We denote R_{opt} the realization determined by T in (52) with $V = I$. We truncate the fractional part of every coefficient of R_b and R_{opt}, which are of infinite precision, to 8 bits. We then compute the actual frequency response of these two FWL realizations and compare them with the ideal frequency response of the filter. The simulation results are given in Fig. 1, which shows the frequency responses of the exact transfer function, and those of the truncated versions of R_b and R_{opt}, in decibels as a function of the ratio between the frequency f and the sampling frequency f_s.

Comment: Since $H(z)$ is a low-pass filter with a very narrow band, taking the weighting functions equal to $H(z)$ means that we weigh the sensitivity of the filter within the filter bandwidth rather than distributing our accuracy efforts throughout the frequency range. So, the weighted optimal realization R_{opt} is expected to deliver a better performance than the optimal realization R_b without weighting within the filter bandwidth. This is confirmed by the simulation of Fig. 1. To illustrate the effect of a different choice of frequency weighting on the performance of the resulting optimal filter, we have performed a second optimal computation with a different frequency weighting. We have now taken $W_1(z) = W_B(z) = W_C(z) = W(z)$ and $W_2(z) = 1$, where the filter $W(z)$, shown in Fig. 2, is also low-pass but with an extra emphasis around the peak of the frequency response of $H(z)$, i.e., in the frequency range $[0.029, 0.032]$. The frequency responses of the exact system, and of the 8-bit truncations of the corresponding optimal realization, are shown in Fig. 3, together with the frequency response of the truncated optimal unweighted realization R_b, for comparison purposes. As expected, this second optimal realization has a better performance within the range where the frequency weighting is highest.

6.2. Schur Realization

From the realization R_b that belongs to the optimal realization subset for the case without weighting, one can get a corresponding Schur realization denoted by R_s. In the same way as described just above, we compute the actual frequency responses of the FWL realizations R_b and R_s with

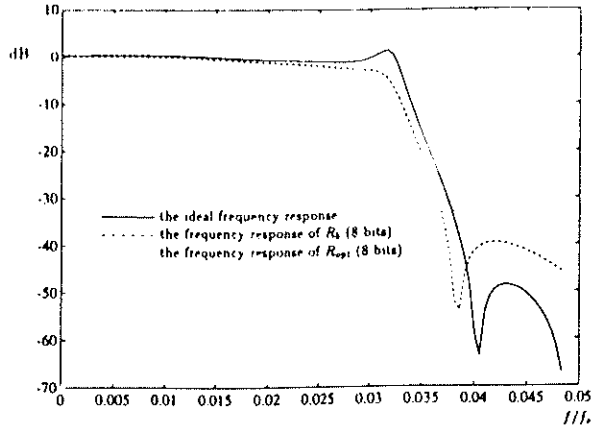

Fig. 1. Frequency responses of FWL weighted and nonweighted optimal realizations.

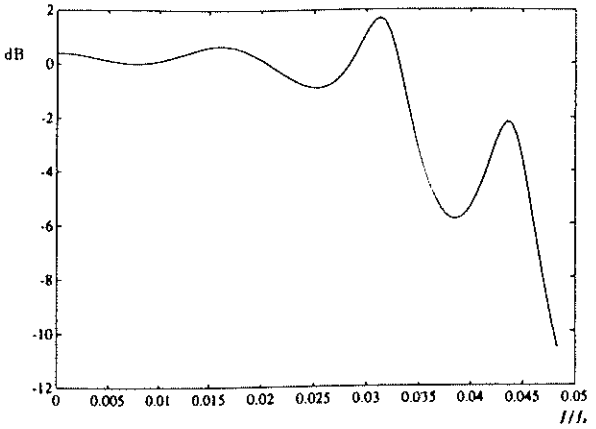

Fig. 2. Frequency response of weighting filter $W(z)$.

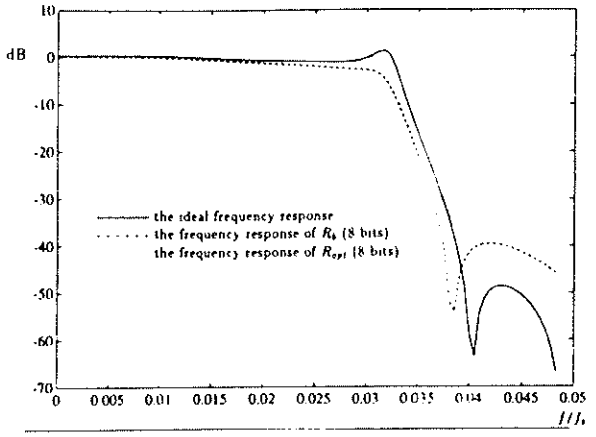

Fig. 3. Frequency responses of FWL weighted and nonweighted optimal realizations.

$p = 8$ bits. The results are shown in Fig. 4, together with the ideal frequency response of the filter. Recall that R_b and R_s are both optimal, but R_s is sparse while R_b is not.

We note that the result of these simulations is a reflection of the actual sensitivity, not just the theoretical one. We observe that the Schur realization R_s not only simplifies the actual implementation but, as expected, it also improves the sensitivity performance of the filter as compared to the

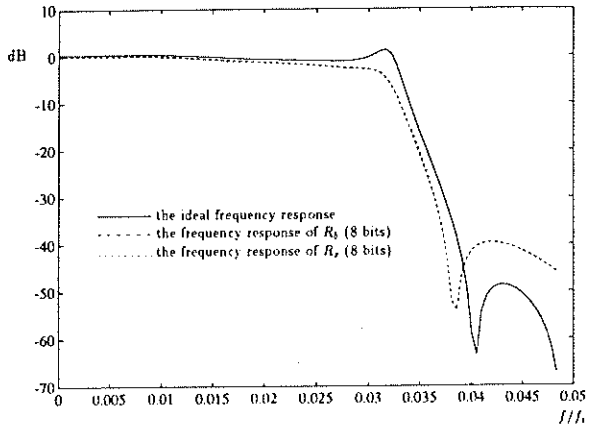

Fig. 4. Frequency responses of an FWL fully parametrized realization and an FWL optimal Schur realization.

balanced form R_b, particularly within the bandwidth of the filter.

VII. Conclusions

In this paper, we have considered the optimal FWL state-space digital system design problem. We have defined a weighted sensitivity and sensitivity measure, which are a generalization of those in [3] and [5]. For convenience of mathematical treatment, the optimal FWL design is performed in terms of minimizing an upper bound of this measure instead of the weighted sensitivity measure itself. Our first contribution in this paper has been to derive a necessary and sufficient condition that characterizes all the optimal similarity transformations and to give an algorithm for solving the minimization problem. The second one is to propose the use of Schur realizations obtained from the optimal realization subset. It is argued that these Schur realizations cannot only simplify the actual implementation but also improve the actual performance of the implemented filter. Some other properties of the Schur realizations have been discussed.

We have also illustrated with a numerical example the nice performance that can be achieved by the weighted optimal realization and the optimal Schur realization, respectively. The theoretical results have been confirmed.

Acknowledgment

The authors would like to thank U. Helmke for suggestions during the development of this paper.

References

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

[2] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273-281, Aug. 1977.

[3] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 884-888, Oct. 1984.

[4] W. J. Lutz and S. Louis Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 1114-1122, Sept. 1988.

[5] L. Thiele, "On the sensitivity of linear state-space systems," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 502-510, May 1986.

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

[7] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parametrizations," in *Proc. 29th Conf. Decision and Control*, pp. 954-959, Dec. 1990.

[8] G. Li, "Finite precision aspects in the parametrizations of control, estimation, and filtering problems," Ph.D. dissertation, Louvain Univ., Belgium, 1990.

[9] L. Thiele, "Design of sensitivity and roundoff noise optimal state-space discrete systems," *Int. J. Circuit Theory Appl.*, vol. 12, pp. 39-46, 1984.

[10] A. J. Laub, "Computing of balancing transformations," in *Proc. 29th IEEE Conf. JACC*, vol. 1, Paper FA 8-E, 1980.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*. North Oxford Academic, 1989, second ed.

[12] P. Van Dooren and M. Verhaegen, "On the use of unitary state-space transformations," *Contemporary Mathematics*, vol. 47, pp. 447-463, 1985.

[13] P. E. Mantey, "Eigenvalue sensitivity and state variable selection," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 263-269, June 1968.

[14] D. Williamson, "Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using feedback," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1210-1220, Oct. 1986.

[15] R. Bellman, *Introduction to Matrix Analysis*. New York: McGraw-Hill, 1970, second ed.

[16] F. Neiss, *Determinanten und Matrizen*. Berlin, Germany: Springer-Verlag, 1967.

[17] J. E. Perkins, U. Helmke, and J. B. Moore, "Balanced realizations via gradient flow techniques," *Syst. Contr. Lett.*, vol. 14, pp. 369-380, 1990.

[18] M. Iwatsuki, M. Kawamata, and T. Higuchi, "Statistical sensitivity and minimum sensitivity structures with fewer coefficients in discrete time linear systems," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 72-80, Jan. 1989.

[19] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. New York: Springer-Verlag, Communication and Control Engineering Series, to be published.

[20] U. Helmke and J. B. Moore, "L^2 sensitivity minimization of linear system representations via gradient flows," submitted to *J. Math. Syst., Estimation and Control*, 1991.

[21] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*. Reading, MA: Addison Wesley, 1987.

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 from Louvain University, Belgium, in 1988 and 1990, respectively.

Since 1991 he has been with the Control Group at Louvain University as a Researcher. His research interests include digital system design, optimal control and filtering, numerical problems in estimation and control theory, and digital signal processing.

Brian D. O. Anderson (S'62-M'66-SM'74-F'75) received his undergraduate education at the University of Sydney and Stanford University. He received the doctorate (honoris causa) from the Université Catholique de Louvain, Belgium.

Following completion of his education, he worked in industry in Silicon Valley and served as an assistant professor in the department of electrical engineering at Stanford. He was foundation professor of electrical engineering at the University of Newcastle, Australia from 1967 till 1981 and is now professor of Systems Engineering at the Australian National University. His interests are in control and signal processing.

Dr. Anderson is a Fellow of the Royal Society, the Australian Academy of Science, Australian Academy of Technological Sciences and Engineering, and the Institute of Electrical and Electronic Engineers, and an Honorary Fellow of the Institution of Engineers, Australia. He is serving a term as President of the International Federation of Automatic Control from 1990 to 1993.

Michel Gevers (S'66–M'72–SM'86–F'90) received the electrical engineering degree from Louvain University, Belgium, in 1968, and the 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. 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 co-author with R. R. Bitmead and V. Wertz of *Adaptive Optimal Control—The Thinking Man's GPC* (Prentice Hall, 1990), and with G. Li of *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects* (Springer-Verlag, Communication and Control Engineering Series, 1992).

Jane Perkins (S'91) received the B.Sc. (hons) degree at the University of Western Australia in 1988.

She is currently working towards a Ph.D. degree in the Department of Systems Engineering at the Australian National University.