A Unifying Framework for Finite Wordlength Realizations

Thibault Hilaire, Philippe Chevrel, and James F. Whidborne, *Member, IEEE*

***Abstract*—A general framework for the analysis of the finite wordlength (FWL) effects of linear time-invariant digital filter implementations is proposed. By means of a special implicit system description, all realization forms can be described. An algebraic characterization of the equivalent classes is provided, which enables a search for realizations that minimize the FWL effects to be made. Two suitable FWL coefficient sensitivity measures are proposed for use within the framework, these being a transfer function sensitivity measure and a pole sensitivity measure. An illustrative example is presented.**

***Index Terms*—Coefficient sensitivity, digital filter implementation, digital filter wordlength effects, finite wordlength (FWL) effects, implicit systems, optimal realization.**

I. Introduction

WHEN digital filters are implemented, they are implemented with finite precision due to the finite wordlength (FWL) of the representation of numbers within the computing machine. There are two FWL effects. The first is the addition of noise into the system resulting from the rounding of variables before and after each arithmetic operation—the "roundoff noise." The second is the degradation in the performance and/or the stability resulting from rounding of the filter coefficients—the "coefficient sensitivity." The FWL problem is hence to analyze the effects to ensure that they do not cause significant deterioration in the performance of an implemented filter. The effects are obviously dependent upon the chosen wordlength and on the chosen arithmetic format (floating-point, fixed-point, etc.). Slightly less obvious is the fact that the FWL effects are very dependent upon the particular realization, (direct form, cascade, etc.), and upon the chosen operator (shift operator, δ operator, etc.). Thus, in seeking to alleviate the FWL effects, the realization must also be considered.

The FWL effects have been studied for many years. Although many of the early works were motivated by problems in control systems [1], [2], the analysis of the effects were often considered in the open loop. See [3] for a comprehensive review of early work. Further reviews can be found in [4]–[6]. There has been a large amount of work that considers the problem of roundoff noise (e.g., [7]–[9]), however the emphasis of this paper is on the coefficient sensitivity problem.

Early consideration of the transfer function sensitivity to rounding errors in the coefficients can be found in [10], [11]. The work of Thiele [12]–[14] is particularly important in defining a norm on the input–output sensitivity that is tractable. This sensitivity measure provides the foundation for much of the subsequent work. Solutions for other similar measures can be found in [5], [15] and further developed in, for example, [16]–[18]. A related measure using a statistical analysis of the input–output sensitivity has been developed [19]. An extension to the multivariable system case is provided in [20]. The closed-loop control case has also been considered, for example in [21]. Methods for the simultaneous minimization of a sensitivity measure with roundoff noise [22] and subject to scaling requirements [23] have also been developed recently.

The sensitivity of the poles (and zeros) is also a commonly used measure of the coefficient rounding effect. An early analysis appears in [24]. Mantey [25] showed that the poles/eigenvalues are dependent on the state-space realization. It is well-known that an eigenvalue sensitivity is minimized if the system is normal [26]. However Gevers and Li [5] subsequently determined the realization that would minimize a pole sensitivity measure combined with a zero sensitivity measure proposed in [27]. Much subsequent work (see [28]–[31], for example) has considered various similar eigenvalue sensitivity measures for closed-loop control systems.

Most of the significant results have expressed the filter in the state space form. Although most realizations can be transformed into the state-space form, this form is not completely general and has several limitations. Firstly, the analysis of the rounding effect of a specific coefficient in a particular realization form can become very difficult after transformation to the state space form. Secondly, many realization forms require the computation of intermediate variables that cannot be expressed in the state-space form. Furthermore, the state space form is specific to the chosen operator. In reality all implementable operators are actually implemented using the shift operator. For example, a realization expressed in the form of a δ-operator is actually implemented using a shift operator in combination with an intermediate variable.

Thus, a description that includes intermediate variables is required. This paper proposes a particular implicit state-space description that is not subject to these limitations. The proposed specialized implicit form provides a generalized description of any realization in a form that allows a straightforward analysis of the FWL effects as will be shown later in this paper. The description is macroscopic in that it does not require coding

Manuscript received July 10, 2006; revised November 27, 2006. This paper was recommended by Associate Editor T. Hinamoto.

T. Hilaire is with the Institut National de Recherche en Informatique et en Automatique (INRIA), Lannion 22300, France (e-mail: thibault.hilaire@irisa.fr).

P. Chevrel is with the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), Nantes 44321, France.

J. F. Whidborne is with the Department of Aerospace Sciences, Cranfield University, Cranfield MK43 0AL, U.K.

Digital Object Identifier 10.1109/TCSI.2007.902408

1549-8328/$25.00 © 2007 IEEE

details and is platform independent but gives a direct relationship between the description and the implementation algorithm. Note that the idea of representing the intermediate variables in the description has been considered previously [32] (see also [33], [34]), but the description form is less general than the implicit form considered in this paper. For example, δ-realizations cannot be described using this form.

The paper is organized as follows. In the next section, the specialized implicit form is proposed, and a number of definitions given. The idea of a set of structured realizations, or structuration is introduced and several examples of structurations provided. In Section III, the equivalence classes of a realization in the specialized implicit form are provided. This is necessary to enable the determination of realizations that are relatively impervious to FWL effects. In Section IV, several coefficient sensitivity measures are proposed for use with the specialized implicit form. Some examples are given in Section V. Note that although the emphasis in this paper is on the coefficient sensitivity problem, the proposed implicit form can be just as useful to the analysis and solution of the roundoff noise problem, and this will be done in future works.

II. Unifying Framework

A. Specialized Implicit Form

To show the utility of the implicit realization, we consider an example of the implementation of a δ-operator state-space realization. It is well-known [5], [35], [36] that the δ-operator is numerically superior to the usual shift operator generally resulting in less sensitive implementations with less rounding noise.

For a realization expressed with the δ-operator, the input/output relation is

$$\begin{cases} \delta[X(k)] = A_\delta X(k) + B_\delta U(k) \\ Y(k) = C_\delta X(k) + D_\delta U(k) \end{cases} \tag{1}$$

with $\delta = (q-1)/\Delta$, where Δ is a strictly positive constant and q is the delay operator [5]. This is equivalent, in infinite precision, to the classical state-space realization

$$\begin{cases} q[X(k)] = A_q X(k) + B_q U(k) \\ Y(k) = C_q X(k) + D_q U(k) \end{cases} \tag{2}$$

with $A_q = \Delta A_\delta + I$, $B_q = \Delta B_\delta$, $C_q = C_\delta$ and $D_q = D_\delta$.

With these two equivalent realizations, the parametrization is different, therefore when the parameters are subjected to FWL rounding, the two realizations are no longer equivalent, and the impact of the quantization is different. In addition, in order to implement the δ-operator, intermediate variables are necessary. These are also subject to FWL quantization. So the following algorithm:

$$\begin{aligned} T &\leftarrow A_\delta X(k) + B_\delta U(k) \\ X(k+1) &\leftarrow X(k) + \Delta T \\ Y(k) &\leftarrow C_\delta X(k) + D_\delta U(k) \end{aligned} \tag{3}$$

implements (1) where T is an intermediate variable vector.

There are many other possible implementation forms, such as direct form I or II, cascade/parallel decomposition, lattice filters, mixed q/δ, etc., and many of these also require intermediate variables. In order to consider all of them within a general unifying framework, we propose a description, in a single equation, of the filter implementation. The equation provides an explicit description of the parametrization, and allows the analysis of the FWL effects, but is still a macroscopic description. Furthermore, the description is given within a formalism such that the description takes the form of an implicit state-space system. This specialized implicit form is given by

$$\begin{pmatrix} J & 0 & 0 \\ -K & I_n & 0 \\ -L & 0 & I_p \end{pmatrix} \begin{pmatrix} T(k+1) \\ X(k+1) \\ Y(k) \end{pmatrix} = \begin{pmatrix} 0 & M & N \\ 0 & P & Q \\ 0 & R & S \end{pmatrix} \begin{pmatrix} T(k) \\ X(k) \\ U(k) \end{pmatrix} \tag{4}$$

where the following are true.

- $J \in \mathbb{R}^{l\times l}$, $K \in \mathbb{R}^{n\times l}$, $L \in \mathbb{R}^{p\times l}$, $M \in \mathbb{R}^{l\times n}$, $N \in \mathbb{R}^{l\times m}$, $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times m}$, $R \in \mathbb{R}^{p\times n}$, $S \in \mathbb{R}^{p\times m}$, $T(k) \in \mathbb{R}^l$, $X(k) \in \mathbb{R}^n$, $U(k) \in \mathbb{R}^m$ and $Y(k) \in \mathbb{R}^p$.
- Matrix J is lower triangular with 1's on the diagonal, i.e.

$$J = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \star & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ \star & \star & \dots & 1 \end{pmatrix}. \tag{5}$$

- $T(k+1)$ is the intermediate variable in the calculations of step k (the column of 0's in the second matrix shows that $T(k)$ is not used for the calculation at step k—this characterizes the concept of an intermediate variable).
- $X(k+1)$ is the stored state-vector ($X(k)$ is effectively stored from one step to the next, in order to compute $X(k+1)$ at step k).

$T(k+1)$ and $X(k+1)$ form the descriptor-vector: $X(k+1)$ is stored from one step to the next, while $T(k+1)$ is computed and used within one time step.

It is implicitly assumed throughout the paper that the computations associated with the realization (4) are executed in row order giving the following algorithm:

[i] $JT(k+1) \leftarrow MX(k) + NU(k)$

[ii] $X(k+1) \leftarrow KT(k+1) + PX(k) + QU(k)$

[iii] $Y(k) \leftarrow LT(k+1) + RX(k) + SU(k)$.

Note that in practice, steps [ii] and [iii] could be exchanged to reduce the computational delay. Also note that because the computations are executed in row order and J is lower triangular with 1's on the diagonal, there is no need to compute J^{-1}. The example in Section V-B shows how to exploit this particularity that gives an extra degree of freedom (49).

This form is a special case of the generalized implicit form [37]

$$E\begin{pmatrix} X(k+1) \\ Y(k) \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} X(k) \\ U(k) \end{pmatrix} \tag{6}$$

but it is not singular. Equation (4) is equivalent in infinite precision to the classical state-space form

$$\left(\begin{array}{c} T(k+1) \\ X(k+1) \\ \hline Y(k) \end{array}\right) = \left(\begin{array}{cc|c} 0 & J^{-1}M & J^{-1}N \\ 0 & A & B \\ \hline 0 & C & D \end{array}\right) \left(\begin{array}{c} T(k) \\ X(k) \\ \hline U(k) \end{array}\right) \tag{7}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$ where

$$A = KJ^{-1}M + P, \quad B = KJ^{-1}N + Q \tag{8}$$
$$C = LJ^{-1}M + R, \quad D = LJ^{-1}N + S. \tag{9}$$

Note that (7) corresponds to a different parametrization than (4). The system transfer function is given by

$$H(z) = C(zI_n - A)^{-1}B + D. \tag{10}$$

B. Definitions

To complete the framework, the following definitions are required.

Definition 1: A **realization**, $\mathcal{R}$, is defined by the specific set of matrices J, K, L, M, N, P, Q, R, and S used to describe a realization with the implicit form of (4):

$$\mathcal{R} :\triangleq (J, K, L, M, N, P, Q, R, S). \tag{11}$$

Remark 1: $\mathcal{R}$ can also be defined by the matrix $Z \in \mathbb{R}^{(l+n+p)\times(l+n+m)}$

$$Z \triangleq \begin{pmatrix} -J & M & N \\ K & P & Q \\ L & R & S \end{pmatrix} \tag{12}$$

and the dimensions l, m, n and p, so $\mathcal{R}$ could be defined by $\mathcal{R} := (Z, l, m, n, p)$.

Definition 2: $\mathscr{R}_H$ denotes the set of realizations with transfer function H. These realizations are said to be equivalent.

In order to encompass realizations with some special structure (q-operator state-space, δ-operator state-space, direct form, cascade, lattice filters, etc.), we define a set of realizations that possess a particular structure.

Definition 3: A **structuration**[1] $\mathscr{S}$ is a set of realizations having a common structure: some coefficients or some dimensions are fixed *a priori*.

Some examples of common structurations are given in the next section.

Definition 4: $\mathscr{R}_H^{\mathscr{S}}$ is the set of equivalent structured realizations. Realizations from $\mathscr{R}_H^{\mathscr{S}}$ are structured according to $\mathscr{S}$ and have a transfer function H. Hence, $\mathscr{R}_H^{\mathscr{S}} \triangleq \mathscr{R}_H \cap \mathscr{S}$.

Definition 5: A **parametrization** of a realization $\mathcal{R}$ is the set of coefficients of Z that are significant for the realization.

For example, with the q-operator state-space realization of (2), the parametrization is given by the matrices A_q, B_q, C_q and D_q. But for the δ-operator state-space realization of (1), the parametrization is given by the matrices A_δ, B_δ, C_δ, D_δ and the parameter Δ. We will see in the next section that the δ-operator state-space realization includes some additional parameters that are always set to unity or zero. These are not 'significant coefficients' and hence are not included in the parametrization.

[1]This is a useful French word that we have purloined. It is also used in the field of social sciences. Here it means the set of structured realizations.

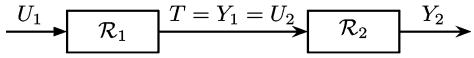

Fig. 1. Cascade form.

C. Some Examples

The δ-realization given by (1) is equivalent, in both finite and infinite precision, to algorithm (3). So this corresponds to the following specialized implicit form:

$$\begin{pmatrix} I_n & 0 & 0 \\ -\Delta I_n & I_n & 0 \\ 0 & 0 & I_p \end{pmatrix} \begin{pmatrix} T(k+1) \\ X(k+1) \\ Y(k) \end{pmatrix} = \begin{pmatrix} 0 & A_\delta & B_\delta \\ 0 & I_n & 0 \\ 0 & C_\delta & D_\delta \end{pmatrix} \begin{pmatrix} T(k) \\ X(k) \\ U(k) \end{pmatrix}. \tag{13}$$

So, the δ-structuration $\mathscr{S}_\delta$ is formally defined by

$$\begin{aligned} \mathscr{S}_\delta = \{\mathcal{R} := (&I_n, \Delta I_n, 0, A_\delta, B_\delta, I_n, 0, C_\delta, D_\delta) \\ &\forall m \in \mathbb{N}, n \in \mathbb{N}, p \in \mathbb{N} \\ &\forall \Delta \in \mathbb{R}^+, A_\delta \in \mathbb{R}^{n\times n}, B_\delta \in \mathbb{R}^{n\times m} \\ &\forall C_\delta \in \mathbb{R}^{p\times n}, D_\delta \in \mathbb{R}^{p\times m}\}. \end{aligned} \tag{14}$$

The cascade form is a common realization for filter implementation. It generally has good FWL properties compared to the direct forms. For cascade form, the filter is decomposed into a number of lower order (usually first- and second-order) transfer function blocks connected in series. For the next example, we consider two standard q-operator filter blocks connected in series as shown in Fig. 1.

If the two state-space realizations $\mathcal{R}_1$ and $\mathcal{R}_2$ are defined by (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2), then cascading $\mathcal{R}_1$ with $\mathcal{R}_2$ leads to the following realization:

$$Z = \left(\begin{array}{c|cc|c} -I & C_1 & 0 & D_1 \\ \hline 0 & A_1 & 0 & B_1 \\ B_2 & 0 & A_2 & 0 \\ \hline D_2 & 0 & C_2 & 0 \end{array}\right) \tag{15}$$

from which definition of the structuration $\mathcal{S}$ immediately follows. The output of $\mathcal{R}_1$ is computed in the intermediate variable, and used as the input of $\mathcal{R}_2$.

The main point is that if we consider the equivalent state-space realization, with parameters

$$\begin{aligned} A &= \begin{pmatrix} A_1 & 0 \\ B_2C_1 & A_2 \end{pmatrix} \\ B &= \begin{pmatrix} B_1 \\ B_2D_1 \end{pmatrix} \\ C &= (D_2C_1 \quad C_2) \\ D &= D_2D_1 \end{aligned} \tag{16}$$

the parametrization is not the one used in the computations.

Remark 2: The cascade structuration can be easily extended to a series of specialized implicit forms and to general multiple cascaded systems.

For a given form, it is generally straightforward to define the structuration. A number of other examples are given in [38] and [39].

III. Equivalent Classes

In order to exploit the potential offered by the specialized implicit form in improving implementations, it is necessary, to describe sets of equivalent system realizations. However, nonminimal realizations may provide better implementations (the δ-form can be seen as a nonminimal realization when expressed in the implicit state-space form with the shift operator. Hence, the notion of equivalence needs to be extended so that the system state dimension does not need to be preserved. The *Inclusion Principle*, introduced by Šiljak and Ikeda [40], [41] in the context of decentralized control, is useful here as it allows the formalization of the *equivalence* and *inclusion* relations between two system realizations.

Definition 6: Consider two systems $\mathcal{S}$ and $\tilde{\mathcal{S}}$, with state dimension n and $\tilde{n} \geqslant n$ respectively, described in classical state-space form by matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $\tilde{A} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$, $\tilde{B} \in \mathbb{R}^{\tilde{n}\times m}$ and $\tilde{C} \in \mathbb{R}^{p\times\tilde{n}}$. System $\mathcal{S}$ is said to be **included** in system $\tilde{\mathcal{S}}$ (denoted by $\mathcal{S} \subset \tilde{\mathcal{S}}$) if there exists $(\mathcal{U},\mathcal{V}) \in \mathbb{R}^{n\times\tilde{n}} \times \mathbb{R}^{\tilde{n}\times n}$ such that $\mathcal{U}\mathcal{V} = I_n$ and, for any initial state $X(0) = X_0$ of $\mathcal{S}$ and any input $(U(k))_{k\geqslant 0}$, the choice of the initial state $\tilde{X}(0) = \mathcal{V}X_0$ of $\tilde{\mathcal{S}}$ implies

$$\begin{cases} X(k) = \mathcal{U}\tilde{X}(k) \\ Y(k) = \tilde{Y}(k) \end{cases} \quad \forall k \geqslant 0. \tag{17}$$

Remark 3: Equation (17) implies that system $\tilde{\mathcal{S}}$ contains all the necessary information to describe the behavior of $\mathcal{S}$.

The principle is extended here to the specialized implicit form in order to characterize equivalence classes. An equivalence class is defined by a certain minimal realization and all the realizations that include this realization. They can be built using the following proposition:

Proposition 1: Consider a realization $\mathcal{R} := (J,K,L,M,N,P,Q,R,S)$ with dimensions l,m,n,p. A realization $\tilde{\mathcal{R}}$ that includes $\mathcal{R}$ can be constructed as follows.

- Choose $\tilde{n}$ and $\tilde{l}$ such that $\tilde{n} + \tilde{l} \geqslant n + l$.
- Choose $(\mathcal{U},\mathcal{V}) \in \mathbb{R}^{n\times\tilde{n}} \times \mathbb{R}^{\tilde{n}\times n}$ such that $\mathcal{U}\mathcal{V} = I_n$, $(\mathcal{W},\mathcal{T}) \in \mathbb{R}^{l\times\tilde{l}} \times \mathbb{R}^{\tilde{l}\times l}$ such that $\mathcal{W}\mathcal{T} = I_l$ and $(\mathcal{X},\mathcal{Y}) \in \mathbb{R}^{l\times\tilde{l}} \times \mathbb{R}^{\tilde{l}\times l}$ such that $\mathcal{X}\mathcal{Y} = I_l$.
- Choose complementary matrices[2] $\mathcal{M}_{\tilde{J}-1} \in \mathbb{R}^{\tilde{l}\times\tilde{l}}$, $\mathcal{M}_{\tilde{K}} \in \mathbb{R}^{\tilde{n}\times\tilde{l}}$, $\mathcal{M}_{\tilde{L}} \in \mathbb{R}^{p\times\tilde{l}}$, $\mathcal{M}_{\tilde{M}} \in \mathbb{R}^{\tilde{l}\times\tilde{n}}$, $\mathcal{M}_{\tilde{N}} \in \mathbb{R}^{\tilde{l}\times m}$, $\mathcal{M}_{\tilde{P}} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$, $\mathcal{M}_{\tilde{Q}} \in \mathbb{R}^{\tilde{n}\times m}$, $\mathcal{M}_{\tilde{R}} \in \mathbb{R}^{p\times\tilde{n}}$ and $\mathcal{M}_{\tilde{S}} \in \mathbb{R}^{p\times m}$ such that, if we denote $\tilde{J}^{-1} = \mathcal{T}J^{-1}\mathcal{X} + \mathcal{M}_{\tilde{J}-1}$, $\tilde{K} = \mathcal{V}K\mathcal{W} + \mathcal{M}_{\tilde{K}}$, $\tilde{L} = L\mathcal{W} + \mathcal{M}_{\tilde{L}}$, $\tilde{M} = \mathcal{Y}M\mathcal{U} + \mathcal{M}_{\tilde{M}}$, $\tilde{N} = \mathcal{Y}N + \mathcal{M}_{\tilde{N}}$, $\tilde{P} = \mathcal{V}P\mathcal{U} + \mathcal{M}_{\tilde{P}}$, $\tilde{Q} = \mathcal{V}Q + \mathcal{M}_{\tilde{Q}}$, $\tilde{R} = R\mathcal{U} + \mathcal{M}_{\tilde{R}}$, $\tilde{S} = S + \mathcal{M}_{\tilde{S}}$ and,

$$\begin{pmatrix} \mathcal{M}_{\tilde{A}} & \mathcal{M}_{\tilde{B}} \\ \mathcal{M}_{\tilde{C}} & \mathcal{M}_{\tilde{D}} \end{pmatrix} = \begin{pmatrix} K \\ L \end{pmatrix} J^{-1}(M \quad N) + \begin{pmatrix} P & Q \\ R & S \end{pmatrix} - \begin{pmatrix} \tilde{K} \\ \tilde{L} \end{pmatrix} \tilde{J}^{-1}(\tilde{M} \quad \tilde{N}) + \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix}$$

then $\mathcal{U}(\mathcal{M}_{\tilde{A}})^i\mathcal{V} = 0\ \forall i \geqslant 1$, $\mathcal{U}(\mathcal{M}_{\tilde{A}})^i\mathcal{M}_{\tilde{B}} = 0\ \forall i \geqslant 0$, $\mathcal{M}_{\tilde{C}}(\mathcal{M}_{\tilde{A}})^i\mathcal{V} = 0\ \forall i \geqslant 0$, $\mathcal{M}_{\tilde{C}}(\mathcal{M}_{\tilde{A}})^i\mathcal{M}_{\tilde{B}} = 0\ \forall i \geqslant 0$ and $\mathcal{M}_{\tilde{D}} = 0$ are satisfied.

[2]These matrices are called *complementary matrices*. $\mathcal{M}_{\tilde{X}}$ is complementary in that it fills the gap between $\bar{X}$ and the similarity on X : $\bar{X} = \mathcal{T}_1 X \mathcal{T}_2 + \mathcal{M}_{\tilde{X}}$.

If so, the realization $\tilde{\mathcal{R}} := (\tilde{J},\tilde{K},\tilde{L},\tilde{M},\tilde{N},\tilde{P},\tilde{Q},\tilde{R},\tilde{S})$ includes the realization $\mathcal{R}$.

Proof: The proof can be derived directly from the characterization of the Inclusion Principle [40], [42], [43]. The details are omitted here but can be found in [38]. ■

Although this proposition gives the formal description of equivalent classes, it is of practical interest to consider realizations of the same dimensions ($\tilde{l} = l$ and $\tilde{n} = n$) where transformations from one realization to another is only a similarity transformation.

Proposition 2: Consider a realization $\mathcal{R} := (Z,l,m,n,p)$. All the realizations $\tilde{\mathcal{R}} := (\tilde{Z},l,m,n,p)$ with

$$\tilde{Z} = \begin{pmatrix} \mathcal{Y} & & \\ & \mathcal{U}^{-1} & \\ & & I_p \end{pmatrix} Z \begin{pmatrix} \mathcal{W} & & \\ & \mathcal{U} & \\ & & I_m \end{pmatrix} \tag{18}$$

and $\mathcal{U},\mathcal{W},\mathcal{Y}$ are nonsingular matrices, are equivalent to $\mathcal{R}$.

It is also possible to just consider a subset of similarity transformations that preserve a particular structure, say cascade or delta. For example, if an initial δ-structured realization $\mathcal{R} := (Z_0,n,m,n,p)$ is given, the subset of equivalent δ-structured realization is defined by

$$\mathscr{R}_H^{\mathscr{S}_\delta} = \left\{ \begin{array}{l} \mathcal{R} := (Z,n,m,n,p) \backslash \\ Z = \begin{pmatrix} \mathcal{U}^{-1} & & \\ & \mathcal{U}^{-1} & \\ & & I_p \end{pmatrix} Z_0 \begin{pmatrix} \mathcal{U} & & \\ & \mathcal{U} & \\ & & I_m \end{pmatrix} \\ \forall \mathcal{U} \in \mathbb{R}^{n\times n} \text{ nonsingular} \end{array} \right\} \tag{19}$$

In addition to a description of the various existing realizations with the exact parametrization, this formalism gives an algebraic characterization of equivalent classes. These classes can be used to search for an optimal structured realization (see Section V-A).

IV. Sensitivity Measures

In order to be able to accurately assess the suitability of a particular realization in the specialized implicit form, some measure of the coefficient sensitivity are required. The measures need to be computationally tractable but also need to account for the fact that common structured realizations (e.g., δ) contain many coefficients that are either zero or unity and hence do not contribute to the FWL effects. Such coefficients are known as trivial parameters. Furthermore, it is useful if the measures can take into account the choice of arithmetic format (floating point or fixed point) of the coefficients' representation.

A. Coefficient Quantization

A coefficient's quantization depends both on their value and their representation.

Firstly if the value of a coefficient is such that it will be quantized without error, then that parameter makes no contribution to the overall coefficient sensitivity. Hence, we introduce weighting matrices W_J to W_S (and also W_Z) respectively associated with matrices J to S of a realization, such that

$$(W_X)_{i,j} \triangleq \begin{cases} 0, & \text{if } X_{i,j} \text{ is exactly implemented} \\ 1, & \text{otherwise.} \end{cases} \tag{20}$$

Secondly, different representation schemes may be considered. Here, we consider both fixed-point and floating-point representations of coefficients expressed using β bits.

A fixed-point coefficient x is represented by $(-1)^s.N.2^{-\beta_f}$, where $s \in \{0,1\}$, N is an integer coded with β_g bits and β_f an integer (not stored in the representation) such that $\beta_g+\beta_f+1 = \beta$. The quantized $x^\dagger$ of x is such that $|x^\dagger - x| < 2^{-(\beta_f+1)}$.

A floating-point coefficient is represented by $(-1)^s.w.2^e$ where $w \in [0,1[$ (or $w \in [0.5,1[$ for a normalized floating-point representation) and e is an integer coded with β_e bits[3] ($\beta_e + \beta_w + 1 = \beta$). The quantized $x^\dagger$ of x is, in this case, such that $|x^\dagger - x| < x.2^{-(\beta_w+1)}$.

The choice of β_f and e can be unique for each coefficient ($e = \lceil \log_2 |x| \rceil$ and $\beta_f = \beta - 1 - \lceil \log_2 |x| \rceil$, where $\lceil \cdot \rceil$ is the *ceiling* operator). Alternatively, β_f and e are defined for a group of coefficients (in order to reduce the required bit-shifts and the subsequent computational cost). This defines the *block-fixed-point* and *block-floating-point* schemes. Following [44], we introduce the *generalized* dynamic range bit β_r ($\beta_r = \beta_g$ or β_e) and the *precision* bit length β_p ($\beta_p = \beta_f$ or β_w).

Usually, the blocks used in block-representation correspond to the matrices J to S, but there is no necessity for this, and blocks can be chosen at will. To define the blocks of a realization $\mathcal{R} := (Z, l, m, n, p)$, we introduce the matrix η_Z such that

$$(\eta_Z)_{i,j} \triangleq \left\{ \begin{array}{l} \text{the largest absolute value of} \\ \text{the block in which } Z_{i,j} \text{ resides.} \end{array} \right. \quad (21)$$

This allows a completely general definition of the blocks. Thus, there could be just a single unique block, or every block could consist of only one coefficient. For example, denoting $\mathcal{E}_{a,b} \in \mathbb{R}^{a \times b}$ as a matrix of 1's and

$$\|X\|_{\max} \triangleq \max_{i,j} |X_{i,j}| \quad (22)$$

then using a block-representation corresponding to the matrices J to S gives

$$\eta_Z = \begin{pmatrix} \|J\|_{\max}\mathcal{E}_{l,l} & \|M\|_{\max}\mathcal{E}_{l,n} & \|N\|_{\max}\mathcal{E}_{l,m} \\ \|K\|_{\max}\mathcal{E}_{n,l} & \|P\|_{\max}\mathcal{E}_{n,n} & \|Q\|_{\max}\mathcal{E}_{n,m} \\ \|L\|_{\max}\mathcal{E}_{p,l} & \|R\|_{\max}\mathcal{E}_{p,n} & \|S\|_{\max}\mathcal{E}_{p,m} \end{pmatrix}.$$

With a single unique block for Z we get $\eta_Z = \|Z\|_{\max}\mathcal{E}_{l+n+p,l+n+m}$, and for one block per coefficient we get $(\eta_Z)_{i,j} = |Z_{i,j}|$.

Proposition 3: During the quantization process, Z is perturbed to $Z + r_Z \times \Delta$ where

$$r_Z \triangleq \left\{ \begin{array}{ll} W_Z, & \text{for fixed-point representation} \\ 2\eta_Z \times W_Z, & \text{for floating-point representation} \end{array} \right. \quad (23)$$

Δ is a matrix dependant on the β_p *precision* bit length, and $\times$ denotes the Schur product. If $\beta_{p_{i,j}}$ is the precision bit-length of $Z_{i,j}$, then $|\Delta_{i,j}| < 2^{-(\beta_{p_{i,j}}+1)}$.

Remark 4: With this formalism for the different representation schemes, note that the choice of the scale parameter (e or β_f) is defined for each coefficient ($e_{i,j} = \lceil \log_2 |\eta_{Z_{i,j}}| \rceil$ and $\beta_{f_{i,j}} = \beta - 1 - \lceil \log_2 |\eta_{Z_{i,j}}| \rceil$) and that it is also possible to define the minimum bit length β to code each coefficient without overflow or underflow [38].

[3]The difference with fixed-point is that e is coded with β_e bits and can be changed. With fixed-point, β_f is fixed and implicit.

B. Transfer Function Sensitivity Measure

The sensitivity measure proposed here extends the measure proposed by Gevers and Li [5] to the specialized implicit state-space form as well as accounting for trivial parameters and the coefficient representation.

Let $H^\dagger \triangleq H|_{Z+r_Z\times\Delta}$ denote the transfer function H perturbed by the quantization process. Then, in the single-input single-output (SISO) case, $\forall z \in \mathbb{C}$:

$$H^\dagger(z) - H(z) = \sum_{i,j} \Delta_{i,j} \left.\frac{\partial H^\dagger(z)}{\partial \Delta_{i,j}}\right|_{\Delta=0} + o\left(\|\Delta\|_{\max}\right). \quad (24)$$

Then

$$\|H^\dagger - H\|_2 \leqslant \|\Delta\|_{\max} \left\| \left.\frac{\partial H^\dagger}{\partial \Delta}\right|_{\Delta=0} \right\|_2 \quad (25)$$

where $\|\cdot\|_2$ is the L_2-norm. It is easy to see that

$$\left.\frac{\partial H^\dagger}{\partial \Delta}\right|_{\Delta=0} = \frac{\partial H}{\partial Z} \times r_Z \quad (26)$$

and so (25) leads to the following transfer function sensitivity measure:

Definition 7 (SISO Transfer Function Sensitivity): Consider a realization $\mathcal{R} := (Z, l, m, n, p)$ with an associated matrix r_Z. The sensitivity of the realization's transfer function H with respect to all the nontrivial coefficients of $\mathcal{R}$, is defined in the SISO case by

$$M_{L_2}^W \triangleq \left\| \frac{\partial H}{\partial Z} \times r_Z \right\|_2^2. \quad (27)$$

Remark 5: The measure $M_{L_2}^W$ differs from the measure proposed by Gevers and Li [5] for SISO classical state-space fixed-point realizations which is defined by

$$M_{L_2} \triangleq \left\| \frac{\partial H}{\partial A} \right\|_2^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C} \right\|_2^2. \quad (28)$$

For the definition of M_{L_2}, the $\|\partial H/\partial D\|_2^2 = 1$ term is ignored because it is invariant to the possible realization. However, for the specialized implicit form, $D = LJ^{-1}N + S$ and so the sensitivity of the direct feed-through term, $\partial D/\partial Z$, is dependent on the realization.

This measure can be extended to the multiple-input multiple output (MIMO) case. However it is also useful to be able to consider the contribution of each coefficient to the overall sensitivity. The *transfer function sensitivity matrix*, denoted by $\frac{\delta H}{\delta Z}$, is the matrix of the L_2-norm of the sensitivity of the transfer function H with respect to each coefficient $Z_{i,j}$. It is defined by

$$\left(\frac{\delta H}{\delta Z}\right)_{i,j} \triangleq \left\| \frac{\partial H}{\partial Z_{i,j}} \right\|_2 \quad (29)$$

and allows the evaluation of the overall impact of each coefficient. It can be used to evaluate the overall sensitivity. From the properties of the L_2-norm, we have

$$\left\|\frac{\delta H}{\delta Z}\right\|_F = \left\|\frac{\partial H}{\partial Z}\right\|_2 \tag{30}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Definition 8: The MIMO transfer function sensitivity is defined by

$$M_{L_2}^W = \left\|\frac{\delta H}{\delta Z} \times r_Z\right\|_F^2. \tag{31}$$

Next, we introduce a new operator that simplifies the subsequent expressions for $M_{L_2}^W$ and the transfer function sensitivity matrix.

Definition 9: The operator $\circledast$ is defined by

$$A \circledast B \triangleq \mathrm{Vec}(A).\left[\mathrm{Vec}(B^\top)\right]^\top \tag{32}$$

where $A \in \mathbb{C}^{m\times p}$ and $B \in \mathbb{C}^{l\times n}$ and where $\mathrm{Vec}(\cdot)$ is the classical operator that transforms a matrix to a column vector. So $A \circledast B \in \mathbb{R}^{mp\times ln}$.

Lemma 1: Consider two matrices (or transfer functions), $G \in \mathbb{C}^{m\times p}$ and $H \in \mathbb{C}^{q\times n}$, that are assumed to be independent of a matrix $X \in \mathbb{R}^{p\times q}$. Then

$$\frac{\partial(GXH)}{\partial X} = G \circledast H \tag{33}$$

and

$$\frac{\partial(GX^{-1}H)}{\partial X} = -(GX^{-1}) \circledast (X^{-1}H). \tag{34}$$

Proof: The proof can be found in [38] and comes from

$$\frac{\partial(GXH)}{\partial X} = (I_p \otimes G)\frac{\partial X}{\partial X}(I_q \otimes H). \tag{35}$$

■

Proposition 4:

$$\frac{\partial H}{\partial Z} = (H_3 \quad H_1 \quad I_p) \circledast \begin{pmatrix} H_4 \\ H_2 \\ I_m \end{pmatrix} \tag{36}$$

where

$$\begin{aligned} H_1 &: z \mapsto C(zI_n - A)^{-1} \\ H_2 &: z \mapsto (zI_n - A)^{-1}B \\ H_3 &: z \mapsto C(zI_n - A)^{-1}KJ^{-1} + LJ^{-1} \\ H_4 &: z \mapsto J^{-1}M(zI_n - A)^{-1}B + J^{-1}N \end{aligned}$$

and the dimensions of the transfer functions H_1 to H_4 are $p\times n$, $n\times m$, $p\times l$ and $l\times m$ respectively.

Proof: From the application of Lemma 1 with (8) and (9) to (10). Details are given in [38]. ■

Proposition 5: Consider a matrix X and three transfer function matrices Y, F and G of appropriate dimensions such that

$$\frac{\partial Y}{\partial X} = F \circledast G. \tag{37}$$

The sensitivity matrix is given by

$$\left(\frac{\delta Y}{\delta X}\right)_{i,j} = \|F_{\bullet,i}G_{j,\bullet}\|_2 \quad \forall i,j. \tag{38}$$

Proof: The proof is straightforward and can be found in [38]. ■

C. Pole Sensitivity Measure

The transfer function sensitivity does not explicitly consider the stability of the system. To ensure that the implementation is stable, the sensitivity of the poles needs to be considered. Let $(\lambda_k)_{1\leqslant k\leqslant n}$ denote the poles of a realization $\mathcal{R} := (Z, l, m, n, p)$. These poles are perturbed during the quantization process to $(\lambda_k^\dagger)_{1\leqslant k\leqslant n}$ with

$$\left|\left|\lambda_k^\dagger\right| - |\lambda_k|\right| \leqslant \sum_{i,j} \Delta_{i,j} \left.\frac{\partial\left|\lambda_k^\dagger\right|}{\partial\Delta_{i,j}}\right|_{\Delta=0} + o\left(\|\Delta\|_{\max}\right). \tag{39}$$

Clearly, this expression provides a means by which the minimum bit-length to preserve stability can be determined *a priori*. This is explored in [38] and [45].

We can define the following pole sensitivity measure.

Definition 10: Consider a realization $\mathcal{R} := (Z, l, m, n, p)$ and associated quantization description matrix r_Z. The pole sensitivity measure of $\mathcal{R}$ is defined by

$$\Psi = \sum_{k=1}^{n} \left\|\frac{\partial|\lambda_k|}{\partial Z} \times r_Z\right\|_F^2. \tag{40}$$

The following lemma is required prior to providing a means of evaluating Ψ.

Lemma 2: ([28]) Consider a differentiable function $f : \mathbb{R}^{m\times n} \to \mathbb{C}$, and two matrices $M \in \mathbb{R}^{m\times n}$ and $X \in \mathbb{R}^{p\times q}$. Let M_0, M_1 and M_2 be constant matrices with appropriate dimensions, then the following results hold.

- If $M = M_0 + M_1XM_2$, then
$$\frac{\partial f(M)}{\partial X} = M_1^\top \frac{\partial f(M)}{\partial M} M_2^\top.$$
- If $M = M_0 + M_1X^{-1}M_2$, then
$$\frac{\partial f(M)}{\partial X} = -(M_1X^{-1})^\top \frac{\partial f(M)}{\partial M}(X^{-1}M_2)^\top.$$

Proposition 6:

$$\frac{\partial|\lambda_k|}{\partial Z} = (KJ^{-1} \quad I \quad 0)^\top \frac{\partial|\lambda_k|}{\partial A} \begin{pmatrix} J^{-1}M \\ I \\ 0 \end{pmatrix}^\top. \tag{41}$$

Proof: Apply Lemma 2 to (8). ■

The term $\partial|\lambda_k|/\partial A$ can be evaluated using the following lemma.

Lemma 3: Let $M \in \mathbb{R}^{n\times n}$ be diagonalisable. Let $(\lambda_k)_{1\leqslant k\leqslant n}$ be its eigenvalues, and $(x_k)_{1\leqslant k\leqslant n}$ the corresponding right eigenvectors. Denote $M_x \triangleq (x_1x_2\ldots x_n)$ and $M_y = (y_1y_2\ldots y_n) \triangleq M_x^{-H}$. Then

$$\frac{\partial\lambda_k}{\partial M} = y_k^* x_k^\top \quad \forall k = 1,\ldots,n \tag{42}$$

and

$$\frac{\partial|\lambda_k|}{\partial M} = \frac{1}{|\lambda_k|} Re\left(\lambda_k^* \frac{\partial \lambda_k}{\partial M}\right) \tag{43}$$

where $\cdot^*$ denotes the conjugate operation, $Re(\cdot)$ the real part and $\cdot^H$ the transpose conjugate operator.

Proof: The procedure for the proof can be found in [30].■

Remark 6: In a similar manner to the transfer function sensitivity matrix, (29), a *pole sensitivity matrix* can be constructed to evaluate the overall impact of each coefficient. Let $\frac{\delta|\lambda|}{\delta Z}$ denote the pole sensitivity matrix defined by

$$\left(\frac{\delta|\lambda|}{\delta Z}\right)_{i,j} \triangleq \sqrt{\sum_{k=1}^{n}\left(\frac{\partial|\lambda_k|}{\partial Z_{i,j}}\right)^2}. \tag{44}$$

The pole sensitivity measure is then given by

$$\Psi = \left\| \frac{\delta|\lambda|}{\delta Z} \times r_Z \right\|_F^2. \tag{45}$$

V. Illustrative Example

A. Optimal Realization Problem

The problem of determining the best realization can be posed as follows:

Problem 1 (Optimal Realization Problem): Consider a transfer function H and a sensitivity measure $\mathcal{J}$. The *optimal design problem* is to find the best realization $\mathcal{R}_{\text{opt}}$ with transfer function H according to the criteria $\mathcal{J}$, that is

$$\mathcal{R}_{\text{opt}} = \arg\min_{\mathcal{R}\in\mathscr{R}_H} \mathcal{J}(\mathcal{R}). \tag{46}$$

Due to the size of $\mathscr{R}_H$, this problem cannot be solved practically. Indeed, a solution may even have infinite dimension. Hence, the following problem is introduced to restrict the search to a particular structuration.

Problem 2 (Optimal Structured Realization Problem): The problem to find the optimal structured realization $\mathcal{R}_{\text{opt}}^{\mathscr{S}}$, that is

$$\mathcal{R}_{\text{opt}}^{\mathscr{S}} = \arg\min_{\mathcal{R}\in\mathscr{R}_H^{\mathscr{S}}} \mathcal{J}(\mathcal{R}). \tag{47}$$

The *Inclusion Principle* (Propositions 1 and 2) provides the means to search over the structured realizations set $\mathscr{R}_H^{\mathscr{S}}$. Since the measure $\mathcal{J}$ could be nonsmooth and/or nonconvex, the Adaptive Simulated Annealing (ASA) [46] method has been chosen to solve Problem 2. This method has worked well for other optimal realization problems [30]. The resulting optimal sensitivities for the measures proposed in Section IV are presented next.

B. Example

To illustrate the use of the proposed measures and the optimal design problem, we consider a sixth-order narrowband low-pass filter from [27] given by $H(z) = (0.0047 - 0.0251z^{-1} + 0.0584z^{-2} - 0.0761z^{-3} + 0.05844z^{-4} - 0.0251z^{-5} + 0.0047z^{-6})/(1 - 5.6526z^{-1} + 13.3818z^{-2} - 16.9792z^{-3} + 12.1765z^{-4} - 4.6789z^{-5} + 0.7526z^{-6})$. The poles are $\lambda_1, \lambda_2 = 0.97226 \pm 0.19890$, $\lambda_3, \lambda_4 = 0.93887 \pm 0.16232$ and $\lambda_5, \lambda_6 = 0.91518 \pm 0.064645$ (all the computations are performed with double floating-point precision, but the results are quoted only to 4 significant digits). The weighting matrix W_Z is constructed according to (20) (only 0, ± 1 are considered as exactly implemented). Note that the optimizations are performed using the pole sensitivity measure Ψ, but could also be done using the transfer function sensitivity measure or a mixed measure.

TABLE I
Measures for Different Realizations

realization	$M_{L_2}^W$	Ψ
Z_1	1.9071e+10	2.5514e+4
Z_2	1.6621e+2	1.9723
Z_3	1.4247e+8	1.7346
Z_4	1.7329e+7	1.7321
Z_5	6.3340e+6	0.4335

The following realizations are considered.

- Z_1 Direct form II with shift-operator q: it corresponds to a canonical state-space form: the transfer function coefficients directly appear in Z.
- Z_2 Balanced classical state-space realization.
- Z_3 Ψ-optimal classical state-space realization: the set of equivalent classical state-space realizations is searched using (48), shown at the bottom of the next page, where $(A_q^0, B_q^0, C_q^0, D_q^0)$ is one classical state-space realization of H and $\mathcal{U}$ is a nonsingular matrix.
- Z_4 Ψ-optimal implicit state-space realization: we consider all the equivalent realizations described by (49), shown at the bottom of the next page, where E is a lower triangular matrix [as for J in (5)]. This can be described in the implicit state-space framework by (50), shown at the bottom of the next page, and equivalent realizations can be searched with the similarity shown in (51) at the bottom of the next page, where $\mathcal{U}$ is a nonsingular matrix and $\mathcal{Y}$ is chosen so that $\mathcal{Y}E\mathcal{U}$ is still lower triangular (in practice, the coefficients of the new matrix E are chosen by the optimization algorithm, and $\mathcal{Y}$ is then deduced).
- Z_5 Ψ-optimal state-space δ-realization ($\Delta = 2^{-2}$): the δ-structured equivalent set $\mathscr{R}_H^{\mathscr{S}_\delta}$ [see (19)] is searched, with an initial realization given as for (13).

The values of the sensitivity measures for the various realizations are given in Table I. Note that the balanced realization Z_2 has the best transfer function sensitivity. This is not surprising since it is known that balanced realizations are optimal in another transfer function sensitivity sense [5]. The pole sensitivity of Z_2 is fairly good, but is not minimal and can be reduced by a factor of 4 by using the state-space δ realization. However, realizations Z_3, Z_4 and Z_5 have a poor transfer function sensitivity.

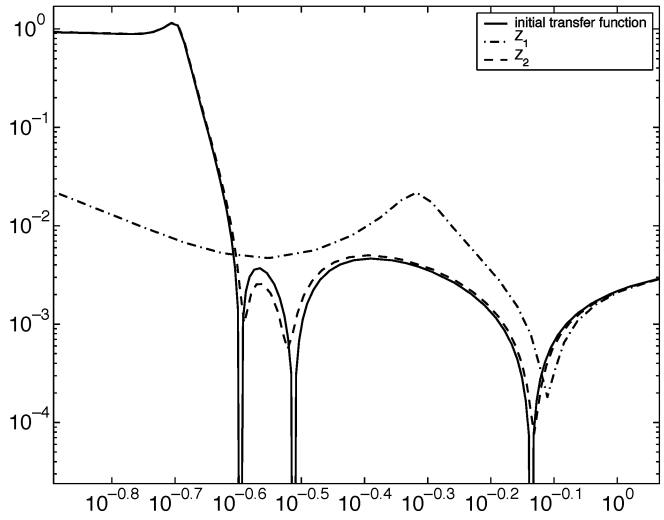

Fig. 2. Transfer function frequency response after quantization (11 bits fixed-point with natural blocks).

This is because only pole sensitivity measure Ψ has been minimized. This indicates that minimizing the pole sensitivity alone is not good enough for this application, and some weighted combination of both sensitivities would be preferable.

Fig. 2 shows the *a posteriori* degradation of the transfer function frequency response when implemented in fixed-point, with natural blocks and with 11 bits, for the realizations Z_1 and Z_2 (compared to the initial transfer function H). It shows the very poor sensitivity of the direct form, Z_1. This realization is actually unstable, and in Table II the *a posteriori* perturbation of the transfer function poles are shown. Poles $\lambda_5^\dagger$ and $\lambda_6^\dagger$ are highly sensitive for the direct form and have moved onto the real axis.

TABLE II
POLES DISPLACEMENT AFTER QUANTIZATION (11 bits FIXED-POINT WITH NATURAL BLOCKS)

	Z_1 $\times 10^{-1}$	Z_2 $\times 10^{-4}$	Z_3 $\times 10^{-4}$	Z_4 $\times 10^{-4}$	Z_5 $\times 10^{-5}$
$\lambda_1^\dagger - \lvert\lambda_1\rvert$	2.7871	4.4861	2.6428	2.5198	8.4470
$\lambda_2^\dagger - \lvert\lambda_2\rvert$	2.7871	4.4861	2.6428	2.5198	8.4470
$\lambda_3^\dagger - \lvert\lambda_3\rvert$	3.4274	1.4895	1.1344	2.5644	0.7647
$\lambda_4^\dagger - \lvert\lambda_4\rvert$	3.4274	1.4895	1.1344	2.5644	0.7647
$\lambda_5^\dagger - \lvert\lambda_5\rvert$	8.2677	1.1139	0.9968	0.3943	3.4414
$\lambda_6^\dagger - \lvert\lambda_6\rvert$	2.0237	1.1139	0.9968	0.3943	3.4414

VI. CONCLUSION

A unifying framework for describing FWL implementations is presented. The tools for utilizing this framework to obtain optimal realizations are provided. The characterization of the equivalent classes provides the means for searching over equivalent structurations. Previously proposed measures are extended to the framework and these enable the coefficient sensitivity to be analyzed and minimized in a unified way. Of course there is no guarantee that one particular structuration will result in a better implementation than another, but the framework allows the systematic comparison of realizations with different structures.

The problem for closed-loop control systems has been studied [38], [47] but there is still further work to be done. Consideration of the quantization noise remains, as does consideration of the sparseness and the dynamic range. The optimization problem is hard, and although the ASA method works well, there are no guarantees that the globally optimal realization is found.

$$Z = \left(\begin{array}{c|c|c} & & \\ \hline & \mathcal{U}^{-1} & \\ \hline & & I_p \end{array}\right) \left(\begin{array}{c|c|c} & & \\ \hline & A_q^0 & B_q^0 \\ \hline & C_q^0 & D_q^0 \end{array}\right) \left(\begin{array}{c|c|c} & & \\ \hline & \mathcal{U} & \\ \hline & & I_m \end{array}\right) \tag{48}$$

$$\begin{cases} EX(k+1) & \text{=AX(k)+BU(k)} \\ Y(k) & \text{=CX(k)+DU(k)} \end{cases} \tag{49}$$

$$Z_0 = \left(\begin{array}{c|c|c} -E & A & B \\ \hline I_n & 0 & 0 \\ \hline 0 & C & D \end{array}\right) \tag{50}$$

$$Z = \left(\begin{array}{c|c|c} \mathcal{Y}^{-1} & & \\ \hline & \mathcal{U}^{-1} & \\ \hline & & I_p \end{array}\right) Z_0 \left(\begin{array}{c|c|c} \mathcal{U} & & \\ \hline & \mathcal{U} & \\ \hline & & I_p \end{array}\right) \tag{51}$$

REFERENCES

[1] J. E. Bertram, "The effects of quantization in sampled-feedback systems," *Trans. Amer. Inst. Elect. Eng.*, vol. 77, pp. 177–182, 1958.

[2] J. B. Slaughter, "Quantization errors in digital control systems," *IEEE Trans. Autom. Contr.*, vol. AC-9, no. 1, pp. 70–74, Jan. 1964.

[3] B. Liu, "Effects of finite word-length on the accuracy of digital filters—A review," *IEEE Trans. Circuit Theory*, vol. 18, no. 6, pp. 670–677, Jun. 1971.

[4] D. Williamson, *Digital Control and Implementation, Finite Wordlength Considerations*. London, U.K.: Prentice-Hall, 1992.

[5] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Probems*. London, U.K.: Springer-Verlag, 1993.

[6] R. S. H. Istepanian and J. F. Whidborne, "Finite-precision computing for digital control systems: Current status and future paradigms," in *Digital Controller Implementation and Fragility: A Modern Perspective*, R. S. H. Istepanian and J. F. Whidborne, Eds. London, U.K.: Springer-Verlag, 2001, ch. 1, pp. 1–12.

[7] I. W. Sandberg, "Floating-point-roundoff accumulation in digital-filter realizations," *Bell Syst. Tech. J.*, vol. 46, no. 8, pp. 1775–1791, 1967.

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

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

[10] J. Knowles and E. Olcayto, "Coefficient accuracy and digital filter response," *IEEE Trans. Circuits Syst.*, vol. CAS-15, no. 1, pp. 31–41, Mar. 1968.

[11] 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.

[12] V. Tavşanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensibility and roundoff noise," *IEEE Trans. Acoust., Speech, Signal, Process.*, vol. CAS-31, no. 10, pp. 884–888, Oct. 1984.

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

[14] L. Thiele, "On the sensitivity of linear state space systems," *IEEE Trans. Circuits Syst.*, vol. CAS-33, no. 5, pp. 502–510, 1986.

[15] W.-Y. Yan and J. B. Moore, "On L_2-sensitivity minimization of linear state-space systems," *IEEE Trans. Circuits Syst., Fundam. Theory Appl.*, vol. 39, no. 8, pp. 641–648, Aug. 1992.

[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. II, Exp. Briefs*, vol. 39, no. 5, pp. 365–377, May 1992.

[17] C. Xiao, "Improved L_2-sensitivity for state-space digital system," *IEEE Trans. Signal Process.*, vol. 45, no. 4, pp. 837–840, Apr. 1997.

[18] T. Hinamoto, S. Yokoyama, T. Inoue, W. Zeng, and W. Lu, "Analysis and minimization of L_2-sensitivity for linear systems and two-dimensional state-space filters using general controllability and observability gramians," *IEEE Trans. Circuits Syst., Fundam. Theory Appl.*, vol. 49, no. 9, pp. 1279–1289, Sep. 2002.

[19] 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, no. 1, pp. 72–80, Jan. 1989.

[20] W. J. Lutz and S. L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circuits Syst.*, vol. 35, no. 9, pp. 1114–1122, Sep. 1988.

[21] A. G. Madievski, B. D. O. Anderson, and M. Gevers, "Optimum realizations of sampled-data controllers for FWL sensitivity minimization," *Automatica*, vol. 31, no. 3, pp. 367–379, 1995.

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

[23] T. Hinamoto, H. Ohnishi, and W.-S. Lu, "Minimization of L_2 sensitivity for state-space digital filters subject to L_2-dynamic-range scaling constraints," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 52, no. 10, pp. 641–645, 2005.

[24] J. F. Kaiser, "Digital filters," in *System Analysis by Digital Computer*, F. F. Kuo and J. F. Kaiser, Eds. New York: Wiley, 1966, pp. 218–285.

[25] P. E. Mantey, "Eigenvalue sensitivity and state-variable selection," *IEEE Trans. Autom. Contr.*, vol. AC-13, no. 3, pp. 263–269, Mar. 1968.

[26] R. Skelton and D. Wagie, "Minimal root sensitivity in linear systems," *J. Guidance, Contr. Dyn.*, vol. 7, no. 5, pp. 570–574, 1984.

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

[28] G. Li, "On the structure of digital controllers with finite word length consideration," *IEEE Trans. Autom. Contr.*, vol. 43, no. 5, pp. 689–693, May 1998.

[29] J. F. Whidborne, R. Istepanian, and J. Wu, "Reduction of controller fragility by pole sensitivity minimization," *IEEE Trans. Autom. Contr.*, vol. 46, no. 2, pp. 320–325, Feb. 2001.

[30] J. Wu, S. Chen, G. Li, R. Istepanian, and J. Chu, "An improved closed-loop stability related measure for finite-precision digital controller realizations," *IEEE Trans. Autom. Contr.*, vol. 46, no. 7, pp. 1162–1166, Jul. 2001.

[31] H.-J. Ko and W. S. Yu, "Guaranteed robust stability of the closed-loop systems for digital controller implementations via orthogonal hermitian transform," *IEEE Trans. Syst., Man, Cybern.*, vol. 34, no. 4, pp. 1923–1932, Aug. 2004.

[32] D. S. K. Chan, "The structure of recursible multidimensional discrete systems," *IEEE Trans. Autom. Contr.*, vol. AC-25, no. 4, pp. 663–673, Aug. 1980.

[33] D. S. K. Chan, "Constrained minimization of roundoff noise in fixed-point digital filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '79)*, Washington, DC, Apr. 1979, pp. 335–339.

[34] P. Moroney, A. S. Willsky, and P. K. Houpt, "The digital implementation of control compensators: The coefficient wordlength issue," *IEEE Trans. Autom. Contr.*, vol. AC-25, no. AC-4, pp. 621–630, Aug. 1980.

[35] R. Middleton and G. Goodwin, *Digital Control and Estimation, a Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1990.

[36] R. Goodall, "Perspectives on processing for real-time control," *Ann. Rev. Contr.*, vol. 25, pp. 123–131, 2001.

[37] J. Aplevich, *Implicit Linear Systems*. New York: Springer-Verlag, 1991.

[38] T. Hilaire, "Analyse et synthèse de l'implémentation de lois de contrôle-commande en précision finie (étude dans le cadre des applications automobiles sur calculateur embarquée)," Ph.D. dissertation, Dept. Comp. Sci., Université de Nantes, Nantes, France, Jun. 2006.

[39] T. Hilaire, P. Chevrel, and Y. Trinquet, "Implicit state-space representation: A unifying framework for FWL implementation of LTI systems," in *Proc. 16th IFAC World Congr.*, Jul. 2005, CDROM.

[40] M. Ikeda, D. Šiljak, and D. White, "An inclusion principle for dynamic systems," *IEEE Trans. Autom. Contr.*, vol. 29, no. AC-3, pp. 244–249, Mar. 1984.

[41] D. Šiljak, *Decentralized Control of Complex Systems*. New York: Academic, 1991.

[42] M. Ikeda and D. Šiljak, "Overlapping decompositions, expansions, and contractions of dynamic systems," *Large Scale Syst.*, vol. 1, pp. 29–38, 1980.

[43] L. Bakule, J. Rodellar, and J. Rossell, "Structure of expansion-contraction matrices in the inclusion principle for dynamic systems," *SIAM Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1136–1155, 2000.

[44] J. Wu, S. Chen, and J. Chu, "Comparative study on finite-precision controller realizations in different representation schemes," in *Proc. 9th Annual Conf. Chinese Autom. Comput. Soc.*, Luton, U.K., Sep. 2003, pp. 257–262.

[45] J. Wu, S. Chen, J. F. Whidborne, and J. Chu, "A unified closed-loop stability measure for finite-precision digital controller realizations implemented in different representation schemes," *IEEE Trans. Autom. Contr.*, vol. 48, no. 5, pp. 816–823, May 2003.

[46] L. Ingber, "Adaptive simulated annealing (ASA): Lessons learned," *Contr. Cybern.*, vol. 25, no. 1, pp. 33–54, 1996.

[47] T. Hilaire, P. Chevrel, and J. Whidborne, "Low parametric closed-loop sensitivity realizations using fixed-point and floating-point arithmetic," in *Proc. Eur. Contr. Conf. (ECC'07)*, Kos, Greece, Jul. 2–5, 2007, to be published.

Thibault Hilaire received the M.Sc. and Ph.D. degrees in control and applied computing sciences from University of Nantes, Nantes, France, in 2003 and 2006, respectively.

He is currently a Postdoctoral Researcher at the IRISA Laboratory (R2D2), Institut National de Recherche en Informatique et en Automatique (INRIA), Lannion, France. He is working on finite word length implementation (sensitivity measures, roundoff noise, optimal design, etc.).

Philippe Chevrel received the Ph.D. degree from the University of Paris XI, in 1993.

He is currently a Professor at Ecole des Mines de Nantes, Nantes, France, and is a member of the control team at Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), Nantes France. He is author or coauthor of nearly 100 research publications and reports, including patents and book chapters. His research interests include robust control, structured control and control implementation, from theoretical to practical, with applications to automotive systems, power systems, and vibration control.

James F. Whidborne (M'96) received the B.A. degree in engineering from Cambridge University, Cambridge, U.K., in 1982, and the M.Sc. and Ph.D. degree in systems and control from UMIST, Manchester, U.K., in 1987 and 1992, respectively.

He is currently a Senior Lecturer in the Department of Aerospace Sciences, Cranfield University, Cranfield, U.K. He has over 100 research publications, including three books and his research interests include optimal finite-precision controller implementations, multi-objective robust control design, fluid flow control and control of unmanned air vehicles.