Designing Low Parametric Sensitivity FWL Realizations of LTI Controllers/Filters within the Implicit State-Space Framework

Thibault Hilaire, Philippe Chevrel, Yvon Trinquet

Designing Low Parametric Sensitivity FWL Realizations of LTI Controllers/Filters within the Implicit State-Space Framework

T. Hilaire*‡,P. Chevrel, *IEEE Member**† and Y. Trinquet*
*IRCCyN, UMR CNRS 6597, 1 rue de la Noë
44321 NANTES, FRANCE
†EMN, 4 rue A. Kaster, La Chantrerie
44307 NANTES, FRANCE
‡PSA Peugeot Citroën, 18 rue des Fauvelles
92256 La GARENNE, FRANCE

***Abstract*— The problem of Finite Word Length (FWL) implementation of Linear Time Invariant (LTI) filters or controllers is considered in this paper. A specialized implicit state-space representation enabling a macroscopic description of the algorithm to be implemented is exhibited. It constitutes a unifying framework to encompass various implementation form, such as q, δ, the observer-based and other realizations. This paper formalizes especially the problem consisting to analyze the parametric sensitivity of such realizations, and then to optimize them in order to limit deteriorations along the process of FWL implementation. The sensitivity of some structured realizations with respect to the coefficients involved and the computation effort are compared on an example.**

I. Introduction

The majority of modern control systems are digitaly implemented with general micro-controller or with specific computing device, like DSP[1]. Since the processor cannot compute with infinite precision (except in some special cases), the Finite World Length (FWL) implementation of a control algorithms leads to a deterioration of the input/output relationship. The deterioration so induced has two separate origins : the quantization of the coefficients involved and the roundoff errors in the numerical computations[1]. They can respectively be formalized as parametric errors and numerical noises. Both depends on the realization, the structure of the control algorithm and the software technics used.
Numerous works in the filtering and control community have studied ([2], [3], [4], [1]) the implementation problem. They often look for the *best* realization, relatively to different criteria : the computations saving, the parametric sensitivity or roundoff noise gain.

The implementation problem is an important one in the modern automotive industry. In a single car, hundreds of filters and controllers are embedded in digital processors. Most of them, for cost reason, are fixed-point processors (in opposition to floating-point processors) and the implementation of effective controllers (from engine control to vehicule dynamics) requires a specific attention to avoid numerical difficulties.
Even for more advanced processors, there is a real need for a methodology in order to manage better the compromise between the readability of the code, the on-line computation effort (computation time, memory size, ...) and the quality of the implementation (minimization of the FWL degradation).

This paper attempt to provide tools and a formalism allowing to get, for a given system, a parametrization consistent with the implementation stage. Precisely, the macroscopic description of the algorithm to be implemented, represented here within the implicit state-space framework, will be analysed and designed so as to prevent important deterioration (due to FWL implementation).
The paper is organized as follows. Section II presents a survey of classical optimal low-parametric sensitivity design. Section III exhibits the way to use an implicit state-space realization (first presented in [11]) in order to encapsulate, in a single framework directly related to implementation, various parametrizations usually studied separately (such as q, δ or observer-based realizations). Section IV, the main contribution of this paper, generalizes the use of the parametric sensitivity to this implicit description for the purpose to design optimal sensitivity realizations, and section V exhibits the generalized optimal realization design problem. Finally, section VI presents the results obtained with different realizations while section VII conclude.

II. The classical low sensitivity realization problem

In this paper, only the deterioration induced by the quantization of the coefficients, is considered. Pratically, during the implementation process, the coefficients can be approximated in different ways, depending on the wordlength and the representation method : undoubtedly, a coefficient will be truncated more using a 8-bits fixed point representation than a 32-bits floating point one.
Moreover, the quantization may have different impact on the

This work was supported by PSA Peugeot Citroën.
Email: *thibault.hilaire@irccyn.ec-nantes.fr* and *philippe.chevrel@emn.fr*
[1]Digital Signal Processor

deterioration of the control law. In this context, the theory of parametric sensitivity[5] is very useful in order to evaluate the deterioration caused when implementing LTI discrete time systems ([6], [3]). Some works consider the zero-pole sensitivity ([6], [7], [1]) of the filter considered or even of the closed-loop when dealing with a controller. Others consider the input/ouput parametric sensitivity defined by

$$\frac{\partial H}{\partial \rho} \tag{1}$$

where H is a transfer function and ρ one of the parameters from which it is defined. ρ could also be a parameter matrix, and the derivation of H with respect to a matrix ρ is then defined by the matrix of the derivation with respect to each element of ρ.

Let's consider a transfer function $H(z)$ and one of its realization $(A_q^0, B_q^0, C_q^0, D_q^0)$ linked through (2) in the time domain using the shift operator q.

$$H(z) = C_q^0(zI - A_q^0)^{-1}B_q^0 + D_q^0 \tag{2}$$

The input/output relationship may be obtained, for zero initial conditions, from

$$\left\{ \begin{array}{rcl} qX_k & = & A_q^0 X_k + B_q^0 U_k \\ Y_k & = & C_q^0 X_k + D_q^0 U_k \end{array} \right. \tag{3}$$

with $qX_k \triangleq X_{k+1}$.

The realizations of the form $(T^{-1}A_q^0T, T^{-1}B_q^0, C_q^0T, D_q^0)$, with T a non-singular matrix, are all equivalent in infinite precision. Obviously they are no more in finite precision : different realizations of the same system could lead to different performances when implemented in finite wordlength.
The optimal FWL implementation problem is often associated to the problem of finding, in the equivalent realizations set, those that optimize the norm of the parametric sensitivity.
Tavşanoğlu and Thiele[8] have first proposed a L_1/L_2 sensitivity measure, that mixes L_1 and L_2 norms.

$$M_{L_{12}} \triangleq \left\| \frac{\partial H}{\partial A} \right\|_1^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C} \right\|_2^2 \tag{4}$$

Gevers and Li[3] have shown that the realizations minimizing this measure are nothing else than the *internally balanced*[2] realizations. They alternatively proposed to consider the measure

$$M_{L_2} \triangleq \left\| \frac{\partial H}{\partial A} \right\|_2^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C} \right\|_2^2 \tag{5}$$

with some possible additionnal frequency weights. They also show how to take into account sparsity constrains.
According to (5), one problem considered by Gevers and Li consists in finding one realization that minimize M_{L_2}

$$\min_{(A,B,C,D)\in\Omega_q} M_{L_2}(A, B, C, D) \tag{6}$$

[2]A realization is called internally balanced if its controllability and observability Gramians are identical and diagonal

with

$$\Omega_q = \left\{ (T^{-1}A_q^0T, T^{-1}B_q^0, C_q^0T, D_q^0) \backslash T \text{ non-singular} \right\} \tag{7}$$

It is important at this stage to notice that Ω_q, the set of equivalent realizations, contains minimal[3] realizations only.

This problem has also being considered when using the δ-operator ([2], [9], [3], [10]) defined by

$$\delta \triangleq \frac{q-1}{\Delta} \tag{8}$$

where Δ is a positive real constant[4]. The problem consists then in finding the matrices $(A_\delta, B_\delta, C_\delta, D_\delta)$ such that the measure M_{L_2} of $H(\delta) = C_\delta(\delta I - A_\delta)^{-1}B_\delta + D_\delta$ is minimized.

III. Macroscopic representation of algorithm through the implicit state-space framework

Various implementation forms have to be taken into consideration (shift, δ-realization, observer-state-feedback, direct form I or II, cascade realization, etc...), in order to evaluate their parametric sensitivity.
[11] shows that a specialized implicit state-space form could be used as a unifying framework, able to describe, in a single equation, various classical implementation. This form is also directly connected to the in-line calculations to be performed and allow a more detailed (still macroscopic) description of a FWL implementation.

Equation (9) recalls the specialized implicit form proposed to make explicit the parametrization and the intermediate variables used.

$$\begin{pmatrix} J & 0 & 0 \\ -K & E & 0 \\ -L & 0 & I \end{pmatrix} \begin{pmatrix} T_{k+1} \\ X_{k+1} \\ Y_k \end{pmatrix} = \begin{pmatrix} 0 & M & N \\ 0 & P & Q \\ 0 & R & S \end{pmatrix} \begin{pmatrix} T_k \\ X_k \\ U_k \end{pmatrix} \tag{9}$$

where

- $J \in \mathbb{R}^{q\times q}$, $E \in \mathbb{R}^{n\times n}$, $K \in \mathbb{R}^{n\times q}$, $L \in \mathbb{R}^{p\times q}$, $M \in \mathbb{R}^{q\times n}$, $N \in \mathbb{R}^{q\times m}$, $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times m}$, $R \in \mathbb{R}^{p\times n}$, $S \in \mathbb{R}^{p\times m}$, $T_k \in \mathbb{R}^q$, $X_k \in \mathbb{R}^n$, $U_k \in \mathbb{R}^m$ and $Y_k \in \mathbb{R}^p$.
- the matrix J is lower triangular with 1 on the diagonal
- E is nonsingular, and, most often to be taken equal to identity
- T_{k+1} is the intermediate variable in the calculations of step k (the column of 0 in the second matrix shows that T_k is not used for the calculation at step k : that characteizes the concept of intermediate variables)
- 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 state-vector : X_{k+1} is stored from one step to the next, while T_{k+1} is computed and used

[3]the term *minimal* refers to the fact that the transfer function $H(z)$ cannot be represented by a realization whose state vector X has a smaller dimension

[4]In [2], Δ corresponds to the sampling period, but this contraint is removed in [3]

inside one time step.

It is implicitly considered throught the paper that the computations associated to the realization (9) are ordered from top to bottom. So the following algorithm is associated in a one to one manner to (9) :

[1] $J.T_{k+1} = M.X_k + N.U_k$:
calculation of the intermediate variables. J is lower triangular, so $T_{k+1}^{(0)}$ is first calculated, and then $T_{k+1}^{(1)}$ using $T_{k+1}^{(0)}$ and so on ... (There's no need to compute J^{-1})

[2] $EX_{k+1} = K.T_{k+1} + P.X_k + Q.U_k$

[3] $Y_k = L.T_{k+1} + R.X_k + S.U_k$

J and E beeing nonsingular, equation (9) is equivalent in infinite precision to the classical state-space form :

$$\left(\begin{array}{c} T_{k+1} \\ X_{k+1} \\ \hline Y_k \end{array}\right) = \left(\begin{array}{cc|c} 0 & J^{-1}M & J^{-1}N \\ 0 & A & B \\ \hline 0 & C & D \end{array}\right) \left(\begin{array}{c} T_k \\ X_k \\ \hline U_k \end{array}\right) \tag{10}$$

with

$$A = E^{-1}KJ^{-1}M + E^{-1}P \tag{11}$$

$$B = E^{-1}KJ^{-1}N + E^{-1}Q \tag{12}$$

$$C = LJ^{-1}M + R \tag{13}$$

$$D = LJ^{-1}N + S \tag{14}$$

However, (10) corresponds to a different parametrization than the one in (9).
The transfer function considered may be then defined by

$$H(z) = C(zI - A)^{-1}B + D \tag{15}$$

Although Gevers and Li make no difference between the terms realization, parametrization or representation, it becomes important from now to distinguish and precise them. It is also important, according to the unifying characteristic of the implicit state-space framework, to define the common terms of all specialized implementations.

Definition 1 *A **realization** $\mathcal{R}$ is defined by the specific set of matrices used for the internal description (see (9))*

$$\mathcal{R} \triangleq (E, J, K, L, M, N, P, Q, R, S) \tag{16}$$

It is said singular if its matrices J or E are singular. A non-singular realization is said to be a realization of the transfer function $H(z)$ if it has the same input-output mapping (assuming null initial conditions)

Definition 2 *$\mathcal{R}_H$ denotes the set of all non-singular realizations of the transfer function $H(z)$. These realizations are said to be equivalent.*

Definition 3 *A **structuration** $\mathcal{S}$ is a subset of realizations having a special structure : some coefficients or some dimensions of the realization matrices $(E, J, K, L, M, N, P, Q, R, S)$ are then* a priori *fixed.*

For example, the δ-structuration, which is the set of realizations with the δ-operator, is defined by

$$\mathcal{S}_\delta \triangleq \left\{ \mathcal{R} \;\middle\backslash\; \begin{array}{l} \mathcal{R} = (I, I, \Delta I, 0, A_\delta, B_\delta, I, 0, C_\delta, D_\delta) \\ \forall (\Delta, A_\delta, B_\delta, C_\delta, D_\delta) \end{array} \right\} \tag{17}$$

Definition 4 *A **structured realization** $\mathcal{R}_H^{\mathcal{S}}$ is the subset of realizations of a transfer function $H(z)$ according to a structuration $\mathcal{S}$*

$$\mathcal{R}_H^{\mathcal{S}} \triangleq \mathcal{R}_H \cap \mathcal{S} \tag{18}$$

Different structured realizations (δ-operator, state-feedback control, cascade realization, etc...) are considered with more details in [11].

IV. The transfer function sensitivity measure, and computations volume

One way to determine the impact of the coefficients' quantization is to compute the sensitivity of the realization considered according to each coefficient involved. The L_2 "measure" (equation (5)) considered by Gevers and Li may be generalized by considering realization (9) as :

$$M_{L_2}^1 \triangleq \sum_{X \in \{E,J,K,L,M,N,P,Q,R,S\}} \left\| \frac{\partial \tilde{H}}{\partial X} \right\|_2^2 \tag{19}$$

with $\tilde{H}(z) \triangleq H(z) - D = C(zI - A)^{-1}B$.
It sums up the possible impact of the quantization of all the coefficients involved in the implicit state-space form. it is preferable to consider $\tilde{H}(z)$ instead of $H(z)$, because $\frac{\partial \tilde{H}}{\partial X} = \frac{\partial H}{\partial X} - \frac{\partial D}{\partial X}$ is strictly proper whatever X and have always a L_2-norm. Moreover, D is independent of the state-space coordinate and have not to be consider here.

As seen in the different implicit structuration examples [11], for a given form (a realization with the δ-operator for example) an important part of the coefficients in $(E, J, K, L, M, N, P, Q, R, S)$ are null or equal to unity. They will not have to be quantized during the implementation process. Therefore, they don't contribute to deteriorate the input/ouput relationship [12]. To take this into account, the measure M_1 may be enriched by introducing *weighting matrices*, allowing to specify that a coefficient will be implemented exactly or not.
Consider X one of the realization matrices (E, J, K, L, M, N, P, Q, R or S). The associated weighting matrice W_X is then defined as :

$$(W_X)_{i,j} = \begin{cases} 0 & \text{if } X_{i,j} \text{ could be } \textit{exactly} \text{ implemented} \\ 1 & \text{if not} \end{cases} \tag{20}$$

The following **new sensitivity** measure, more consistent with FWL implementation analysis, will be used from now :

$$M_{L_2}^W \triangleq \sum_{X \in \{E,J,K,L,M,N,P,Q,R,S\}} \left\| \frac{\partial \tilde{H}}{\partial X} \times W_X \right\|_2^2 \tag{21}$$

where $\times$ denotes the *Schur* (or *Hadamard*) product.

For clarity of the presentation, the SISO transfer function sensitivity measure is considered next (without loss of generality).

Proposition 1 *The sensitivity function of $H(z)$ with respect to each matrix of the considered realization are given by*

$$\frac{\partial \tilde{H}}{\partial S} = 0 \quad (22)$$

$$\frac{\partial \tilde{H}}{\partial R} = \left((zI-A)^{-1}B\right)^{\top} \quad (23)$$

$$\frac{\partial \tilde{H}}{\partial Q} = \left(C(zI-A)^{-1}E^{-1}\right)^{\top} \quad (24)$$

$$\frac{\partial \tilde{H}}{\partial P} = \frac{\partial \tilde{H}}{\partial Q}\frac{\partial \tilde{H}}{\partial R} \quad (25)$$

$$\frac{\partial \tilde{H}}{\partial L} = \left(J^{-1}M(zI-A)^{-1}B\right)^{\top} \quad (26)$$

$$\frac{\partial \tilde{H}}{\partial N} = \left(C(zI-A)^{-1}E^{-1}KJ^{-1}\right)^{\top} \quad (27)$$

$$\frac{\partial \tilde{H}}{\partial M} = \left(\frac{\partial \tilde{H}}{\partial N} + \left(LJ^{-1}\right)^{\top}\right)\frac{\partial \tilde{H}}{\partial R} \quad (28)$$

$$\frac{\partial \tilde{H}}{\partial K} = \frac{\partial \tilde{H}}{\partial Q}\left(\frac{\partial \tilde{H}}{\partial L} + \left(J^{-1}N\right)^{\top}\right) \quad (29)$$

$$\frac{\partial \tilde{H}}{\partial E} = -\frac{\partial H}{\partial Q}\left(A(zI-A)^{-1}B+B\right)^{\top} \quad (30)$$

$$\frac{\partial \tilde{H}}{\partial J} = -\left(LJ^{-1}\right)^{\top}\frac{\partial \tilde{H}}{\partial L} - \frac{\partial \tilde{H}}{\partial N}\left(J^{-1}N\right)^{\top} \quad (31)$$

$$-\frac{\partial \tilde{H}}{\partial N}\frac{\partial \tilde{H}}{\partial L} \quad (32)$$

Proof: The demonstrations are omitted for lack of place, but these expressions derive from equations (15) and (11) to (14). ■

This measure is applied to some various realizations in section VI, equations (37), (38), (40) and (41).

Other criteria may be used in order to evaluate and compare the pertinence of a realization over another : computation time, roundoff noise gain, and the readibility of the code (physical meaning associated to coefficients and state variable, which make the code easier to maintain and adopt). Concerning the first point, we will focus on the number of additions and multiplications rather than the computation time, wich depends also on hardware and the way to perform the operations. For example, the number of operations depends on the coefficients representation (fixed-point, integer, choice of the wordlength and scaling adjustment, ...) and the software tricks used (integer error feedback[13], quantization method during the calculations, ...).

As the specialized implicit state-space realization (9) allows a judicious (although macroscopic) way to describe the code to be implemented, it makes the number of additions and multiplications to be generically performed easy to evaluate.

Proposition 2 *Assuming that E is equal to identity, and $n_{\mathcal{R}}^0$ and $n_{\mathcal{R}}^1$ being respectively the number of null elements of the matrix J, K, L, M, N, P, Q, R, S and the number of trivial elements (0, 1 or -1) of these matrices, then the associated algorithm requires $(m+n+q)(m+n+p-1)-m-n_{\mathcal{R}}^0$ additions and $(m+n+q)(m+n+p)-n_{\mathcal{R}}^1$ multiplications.*

Proof: If Y is a constant in $\mathbb{R}^{a\times b}$ and Z some value $\mathbb{R}^{1\times b}$, the product $Y.Z$ needs at least $a(b-1)-n_Y^0$ additions and $ab-n_Y^1$ multiplications where n_Y^0 denotes the number of the matrix Y's null elements and n_Y^1 the number of trivial elements of Y. ■

V. Generalized Optimal Realization design

It is now important to construct and characterize, for a given transfer function $H(z)$, some subsets of $\mathcal{R}_H$.

Proposition 3 *Let's consider a realization $\mathcal{R}$ and a transfer function $H(z)$.*
$\mathcal{R} \in \mathcal{R}_H$ iff $H(z) = C(zI-A)^{-1}B + D$ with A,B,C and D defined by equations (11) to (14)

Proposition 4 *Let's give us a transfer function $H(z)$ and one of its realizations $\mathcal{R} = (E, J, K, L, M, N, P, Q, R, S)$ with sizes m, n, p and q (m is the size of the inputs, n the dimension of the stored state, p the size of the outputs and q the dimension of the intermediate variable).*
Let's now consider the realization $\bar{\mathcal{R}} = (\bar{E}, \bar{J}, \bar{K}, \bar{L}, \bar{M}, \bar{N}, \bar{P}, \bar{Q}, \bar{R}, \bar{S})$ with same sizes m, n, p and q and with

$$\begin{array}{llll} \bar{E} = WET & \bar{J} = UJV \\ \bar{K} = WKV & \bar{L} = LV \\ \bar{M} = UMT & \bar{N} = UN \\ \bar{P} = WPT & \bar{Q} = WQ \\ \bar{R} = RT & \bar{S} = S \end{array} \quad (33)$$

where $T \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{q\times q}$, $V \in \mathbb{R}^{q\times q}$ and $W \in \mathbb{R}^{n\times n}$ are non-singular matrices.

*With this construction, $\bar{\mathcal{R}}$ **is equivalent to** $\mathcal{R}$.*

Proof: According to equations (15), (11) to (14) and (33), the realizations $\mathcal{R}$ and $\bar{\mathcal{R}}$ share the same transfer function $H(z)$. ■

It is important to notice that $\bar{\mathcal{R}}$ is computed from the transformation matrices T, U, V and W and is the expression of $\mathcal{R}$ in different bases (there is a base transformation for the intermediate variables and for the stored state-vector).

Remark : Thanks to proposition 4, it is possible to characterize a subset of equivalent realizations of a given transfer, but not all equivalent realizations according to (9). Greater subset may be obtained by generalizing conditions (33) in using non-square matrices (rectangular matrices $T_1 \in \mathbb{R}^{n\times\bar{n}}$ and $T_2 \in \mathbb{R}^{\bar{n}\times n}$ with $T_1T_2 = I_n$ remplace transformation matrices T and T^{-1}) or extending the *Inclusion principle* [14], [15] to the case of implicit systems. This will be done later.

The optimal realization problem originated by Gevers and Li (see equation (6) and [3]) and those defined in [3], [1], [7], [13], [16], [**?**] may be generalized as follows.

Problem 1 (Optimal design of a realization) *Let's consider a transfer function $H(z)$ and a measure M of a realization (it could be the transfer function sensitivity measure $M_{L_2}^W$ of the section IV).*
The optimal design problem consits to find the best realization $\mathcal{R}^{opt}$ for the transfer function H according to the measure M

$$\mathcal{R}^{opt} = \underset{\mathcal{R}\in\mathcal{R}_H}{arg\ min}\ M(\mathcal{R}) \tag{34}$$

Remark : This problem is a very difficult one (due to the size of $\mathcal{R}_H$). That's the reason why the problem 2 is introduced.

Problem 2 (Optimal design of a structured realization) *In addition to the previous problem, a structuration $\mathcal{S}$ is also considered. The problem is then to find the optimal structured realization $\mathcal{R}^{opt}$.*

$$\mathcal{R}^{opt} = \underset{\mathcal{R}\in\mathcal{R}_H^{\mathcal{S}}}{arg\ min}\ M(\mathcal{R}) \tag{35}$$

Indeed, the solution of problem 2 is a suboptimal solution of problem 1. The solution of problem 1 may be approached by considering simultanneously several structurations (e.g. of different order), when no particular structurations are specified *a priori*.

Since the measure M could be non-smooth and/or non-convex (like the pole-sensitivity proposed in [1], [7]), the optimizing algorithm should be an efficient global optimization one. The Adaptive Simulated Annealing (ASA) ([17], [18]) has been adopted here to search for the best realization.

VI. Examples

To apply the transfer function sensitivity measure $M_{L_2}^W$, the example choosen is the following :

$$H(z) = \frac{0.01594(z+1)^3}{z^3 - 1.9749z^2 + 1.5562z - 0.4538} \tag{36}$$

It is a low pass filter (see [3], [19]), and has a triple zero at $z = -1$, so the zero positions are very sensitive to the coefficients when realized directly.
In the following results, all the computations are done with double floating-point precision, and the results are quoted only to 4 significant digits for the fractional part. Bold font is used to exhibit parameters that risk to be approximated during the quantization process toward implementation. The others, zeros or ones, will not be deteriorated. The weighting matrices are build accordingly.
Let us now consider different realizations for $H(z)$ (null initial conditions assumed)

1) Realizations with q-operator : it corresponds to the structuration $(I, I, 0, 0, 0, 0, A_q, B_q, C_q, D_q)$
   - Canonical form :

$$\begin{array}{l} A_q^0 = \begin{pmatrix} \mathbf{1.9749} & \mathbf{-1.5562} & \mathbf{0.4538} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \\ B_q^0 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^\top \\ C_q^0 = \begin{pmatrix} \mathbf{0.0793} & \mathbf{0.0230} & \mathbf{0.0232} \end{pmatrix} \\ D_q^0 = \mathbf{0.0159} \end{array} \tag{37}$$

     According to (21), the parametric sensitivity obtain is high : $M_{L_2}^W = 93.7200$
   - an *internally balanced* realization :

$$\begin{array}{l} A_q^1 = \begin{pmatrix} \mathbf{0.8327} & \mathbf{-0.3999} & \mathbf{0.0164} \\ \mathbf{0.3999} & \mathbf{0.5935} & \mathbf{0.3425} \\ \mathbf{0.0164} & \mathbf{-0.3425} & \mathbf{0.5578} \end{pmatrix} \\ B_q^1 = \begin{pmatrix} \mathbf{-0.4424} & \mathbf{0.3799} & \mathbf{0.1671} \end{pmatrix}^\top \\ C_q^1 = \begin{pmatrix} \mathbf{-0.4424} & \mathbf{-0.3799} & \mathbf{0.1671} \end{pmatrix} \\ D_q^1 = \mathbf{0.0159} \end{array} \tag{38}$$

     In this case, and despite more coefficients are potentially deteriorated during quantization, the parametric sensitivity measure is much lower : $M_{L_2}^W = 7.9431$
2) δ-structuration : it corresponds to the structuration $(I, I, \Delta I, 0, A_\delta, B_\delta, I, 0, C_\delta, D_\delta)$. Δ is choosen as in [3] : $\Delta = 2^{-1}$.
   q-structuration and δ-structuration are equivalent ([9], [2]) in infinite precision for

$$A_\delta = \frac{A_q - I}{\Delta},\ B_\delta = \frac{B_q}{\Delta},\ C_\delta = C_q,\ D_\delta = D_q \tag{39}$$

   It is important to notice that the sensitivity measure, takes into account all the involved coefficients and include the sensitivity of the transfer function with respect to Δ, contrary to [3] (in this example, Δ is taken into account, despite it could be exactly be implemented in fixed point representation; moreover, it exists examples where the $H(z)$ is very sensitive with respect to Δ, so that a non exact implementation of Δ leads to a very large degradation).
   - Canonical form

$$\begin{array}{l} A_\delta^0 = \begin{pmatrix} \mathbf{-2.0502} & \mathbf{-2.4256} & \mathbf{-1.0200} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \\ B_\delta^0 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^\top \\ C_\delta^0 = \begin{pmatrix} \mathbf{0.1586} & \mathbf{0.7265} & \mathbf{1.0039} \end{pmatrix} \\ D_\delta^0 = \mathbf{0.0159} \end{array} \tag{40}$$

     $M_{L_2}^W = 9.0171$
   - From the balanced q-form

$$\begin{array}{l} A_\delta^1 = \begin{pmatrix} \mathbf{-0.3527} & \mathbf{-0.7998} & \mathbf{0.0329} \\ \mathbf{0.7998} & \mathbf{-0.8130} & \mathbf{0.6850} \\ \mathbf{0.0329} & \mathbf{-0.6850} & \mathbf{-0.8845} \end{pmatrix} \\ B_\delta^1 = \begin{pmatrix} \mathbf{-0.8848} & \mathbf{0.7598} & \mathbf{0.3341} \end{pmatrix}^\top \\ C_\delta^1 = \begin{pmatrix} \mathbf{-0.4424} & \mathbf{-0.3799} & \mathbf{0.1671} \end{pmatrix} \\ D_\delta^1 = \mathbf{0.0159} \end{array} \tag{41}$$

Here, $M_{L_2}^W = 6.0943$

ASA is used to solve Problem 2 and find the optimal structured realization (according to the transfer function sensitivity measure) for the q and δ-structuration of the previous example :

- Optimal realization with q-structuration

$$\begin{array}{l} A_q^{\text{opt}} = \begin{pmatrix} \mathbf{0.8173} & \mathbf{0.5439} & \mathbf{-0.0483} \\ \mathbf{-0.4718} & \mathbf{0.5373} & \mathbf{0.1256} \\ \mathbf{0.1476} & \mathbf{-0.1042} & \mathbf{0.6203} \end{pmatrix} \\ B_q^{\text{opt}} = \begin{pmatrix} \mathbf{-0.2718} & \mathbf{-0.4674} & \mathbf{-0.3215} \end{pmatrix}^\top \\ C_q^{\text{opt}} = \begin{pmatrix} \mathbf{-0.3210} & \mathbf{0.3189} & \mathbf{-0.4389} \end{pmatrix} \\ D_q^{\text{opt}} = \mathbf{0.0159} \end{array} \quad (42)$$

$M_2^{\text{opt}} = 7.8704$

- Optimal realization with δ-structuration

$$\begin{array}{l} A_\delta^{\text{opt}} = \begin{pmatrix} \mathbf{-0.7254} & \mathbf{-0.5211} & \mathbf{-0.5718} \\ \mathbf{0.2399} & \mathbf{-0.4314} & \mathbf{-0.7084} \\ \mathbf{0.3830} & \mathbf{1.0377} & \mathbf{-0.8933} \end{pmatrix} \\ B_\delta^{\text{opt}} = \begin{pmatrix} \mathbf{0.2110} & \mathbf{-0.8189} & \mathbf{0.0981} \end{pmatrix}^\top \\ C_\delta^{\text{opt}} = \begin{pmatrix} \mathbf{0.8098} & \mathbf{-0.0173} & \mathbf{-0.2696} \end{pmatrix} \\ D_\delta^{\text{opt}} = \mathbf{0.0159} \end{array} \quad (43)$$

$M_2^{\text{opt}} = 4.5671$

These results are summarized in table I, with the evaluation of the corresponding computational effort (number of additions and multiplications).

TABLE I

TRANSFER FUNCTION SENSITIVITY MEASURE OF VARIOUS REALIZATIONS

realization	$\mathbf{M_2}$	**add.**	**mul.**
canonical form q	93.7200	6	7
balanced q	7.9431	12	16
optimal q	7.8704	12	16
canonical form δ	9.0171	10	9
δ from balanced q	6.0943	15	19
δ optimal	4.5671	15	19

The results obtained are not surprising and are coherent with existing ones. The δ-structuration is less sensitive than the q-one and requires more operations for comparable realization. The q-structuration may have an admissible sensitivity when choosing an adequate realization (*e.g.* a balanced one), and the canonical δ-structuration may achieve a good compromise.

VII. Conclusion

The implicit state-space framework has been proposed in order to give a macroscopic view on the algorithm to be implemented. It allows to encapsulate different classical realizations such as q or δ-realization for digital filter and even the observer-based. These realizations are traditionally considered separately. The parametric sensitivity measure introduced in the paper applies to all of them. The characterization of equivalent realizations allows to search for optimal realization over an enlarged set of realizations than in the past. Some realizations have been compared in the case of a particular potentially sensitive transfer function. The results obtained confirm some previous one. Our future works will concern more original structuration such as the generalized delta transform, the observer-based realization, and other unexpected ones. Some interesting results have already been obtained in practical situation in the context of the automotive domain. Our present work focus on the roundoff noise gain estimation thanks to the generalized realisation propose.

VIII. Acknowledgement

The authors wish to thank PSA Peugeot Citroën for their interest and financial support.

References

[1] R. Istepanian and F. Whidborne, Eds., *Digital Controller implementation and fragility*. Springer, 2001.

[2] R. Middleton and G. Goodwin, *Digital Control and Estimation, a unified approach*. Prentice-Hall International Editions, 1990.

[3] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Probems*. Springer-Verlag, 1993.

[4] D. Williamson, *Digital Control and Implementation, Finite Wordlength Considerations*. Prentice-Hall International Editions, 1992.

[5] M. Eslami, *Theory of Sensitivity in Dynamic Systems*. Springer-Verlag, 1994.

[6] D. Williamson, "Roundoff noise minimization and pole-zero sensibivity in fixed-point digital filters using residue feedback," in *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-43, October 1986.

[7] S. Chen, J. Wu, and G. Li, "Two approaches based on pole sensitivity and stability radius measures for finite precision digital controller realizations," *Systems and Control Letters*, vol. 45, no. 4, pp. 321–329, 2002.

[8] V. Tavşanoğlu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensibility and roundoff noise," in *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. CAS-31, October 1984.

[9] M. Rostgaard, N. Poulsen, and O. Ravn, "A rapprochement between discrete-time operators," in *ECC93*, vol. 2, February 1993, pp. 426–431.

[10] R. Middleton, *Delta Toolbox : a tutorial*, March 1990.

[11] T. Hilaire, P. Chevrel, and Y. Trinquet, "Implicit state-space representation : a unifying framework for FWL implementation of LTI systems," in *IFAC05 Wolrd Congress*, July 2005.

[12] G. Li, "On the structure of digital controllers with finite word length consideration," in *IEEE Transactions on Automatic Control*, vol. 43, no. 5, May 1998, pp. 689–693.

[13] M. Rotea and D. Williamson, "Optimal realizations of finite wordlength digital filters and controllers," *IEEE Transactions on Circuits and Systems*, vol. 42, no. 2, pp. 61–72, February 1995.

[14] S. Stanković and D. Šiljak, "Contractibility of overlapping decentralized control," *System & Control Letters*, vol. 44, no. 3, October 2001.

[15] 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.

[16] J. Wu, S. Chen, G. Li, and J. Chu, "Optimal finite-precision state-estimate feedback controller realizations of discrete-time systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 8, pp. 1550–1554, August 2000.

[17] L. Ingber, "Adaptive simulated annealing (asa): Lessons learned," *Control and Cybernetics*, vol. 25, no. 1, pp. 33–54, 1996.

[18] S. Chen and B. Luk, "Adaptive simulated annealing for optimization in signal processing applications," *Signal Processing*, vol. 79, pp. 117–128, 1999.

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