A Practical Strategy of an Efficient and Sparse FWL Implementation of LTI Filters

Yu Feng, Philippe Chevrel, and Thibault Hilaire

***Abstract*— The problem of finite word length implementation is discussed in this paper. Alternatively to the ρDFIIt recently proposed by G. Li et al., and leaning on the specialized implicit form for a unified analysis, a new effective and sparse structure, named ρ-modal realization, is developed. This realization meets simultaneously accuracy (low sensitivity, round-off noise gain and overflow risk), few and flexible computational efforts with a good readability (owing to sparsity), and simplicity (no tricky optimization is involved) as well. Two numerical examples are presented to confirm the theoretical results and illustrate the ρ-modal realization interest.**

I. INTRODUCTION

It is well-known that there exists an infinite set of realizations to represent a given filter. These realizations are equivalent in infinite precision since they yield the same input-output relationship. However, when digital filters are implemented, they have to be represented with finite word length (FWL) in a computing device. The FWL effects lead to a deterioration of realizations' numerical properties. Hence, the equivalent realizations are no longer equivalent in finite precision. One realization may be better suited for implementation than another.

The optimal filter implementation problem consists in minimizing the digital deterioration imposed by FWL effects. Diverse structures (like the cascade framework, the balanced state-space realization, etc.) and different digital operators (δ-operator) have been proposed with this aim since the late 1970s. In [1], rational operators suitable for discretization of both LTI and LPV systems are introduced with taking potentially into account the frequency bandwidth of each sub-system. The ρDFIIt realization proposed in [2], [3] is interesting as it is sparse and is designed in such way to minimize the transfer function sensitivity or the round-off noise gain. Other methods to establish the optimal realizations may be found e.g. in [4], [5].

In this paper, based on a multivariable ρ-operator and a modal representation, a new structure, denoted as ρ-modal realization is constructed within the specialized implicit framework (SIF) [6]. For a filter of order n, this sparse and scaled realization contains few inexactly-implemented[1] parameters, and holds n free parameters to minimize the FWL effects. It will be also be shown that the realization proposed is resilient to numerical errors and can be obtained without using optimization tools.

This paper is briefly outlined as follows. After recalling the specialized implicit form and the related analysis criteria in Section II, the dynamic-range scaling and the standard modal representation are presented in Section III. Then, the particular ρ-modal realization is proposed in Section IV and its properties are studied in Section V. Numerical illustrations are given in section VI before concluding.

Manuscript received June 15, 2009. T. Hilaire's work was supported by the National Research Network "Signal and Information Processing in Science and Engineering" (NFN SISE) project.

Y. Feng and P. Chevrel are both with the Institut de Recherche en Cybernétique et Communication de Nantes (UMR CNRS 6597) and École des Mines de Nantes, France. Corresponding author e-mail: yu.feng@emn.fr

T. Hilaire is with the Institute of Communication and Radio-Frequency Engineering, Vienna University of Technology, Austria.

[1]Exactly-implemented parameters mentioned here are those that are not modified by the process of quantization.

II. SHARED CRITERIONS IN A UNIFYING FRAMEWORK

A. Specialized Implicit Form

There exists plenty of useful and well-known realizations, such as the direct form I or II, the cascade/parallel decomposition etc., and many of these require (although implicitly in the literature) intermediate computational variables. The specialized implicit form proposed in [6] provides an explicit description of the parameters and variables involved during the implementation. The SIF representation is:

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

where the following statements are true:

1) $U(k)$ is the vector of the m current inputs, $Y(k)$ the p current outputs, $X(k)$ and $T(k)$ the vectors of n and l generalized variables; $T(k+1)$ is the vector for the intermediate variables used in the calculations of step k while $X(k+1)$ is the vector new state variables stored till the next sampling time;
2) J is a lower triangular matrix with 1's in the diagonal;
3) The computations associated with the realization (1) are executed in row order:

$$\begin{aligned} &[\text{i}] && JT(k+1) \leftarrow MX(k) + NU(k) \\ &[\text{ii}] && X(k+1) \leftarrow KT(k+1) + PX(k) + QU(k) \\ &[\text{iii}] && Y(k) \leftarrow LT(k+1) + RX(k) + SU(k) \end{aligned} \quad (2)$$

Equation (1) is equivalent in infinite precision to the classical state-space form:

$$\left(\begin{array}{c} T(k+1) \\ X(k+1) \\ \hline Y(k) \end{array}\right) = \left(\begin{array}{cc|c} 0 & J^{-1}M & J^{-1}N \\ 0 & A_Z & B_Z \\ \hline 0 & C_Z & D_Z \end{array}\right) \left(\begin{array}{c} T(k) \\ X(k) \\ \hline U(k) \end{array}\right) \quad (3)$$

with

$$A_Z \triangleq KJ^{-1}M + P, \qquad B_Z \triangleq KJ^{-1}N + Q, \quad (4)$$

$$C_Z \triangleq LJ^{-1}M + R, \qquad D_Z \triangleq LJ^{-1}N + S. \quad (5)$$

It is of importance to notice that these two realizations, though equivalent in infinite precision, are different concerning the parameters involved.

Let us recall an additional definition.

Definition 1 *([6]) A* realization *$\mathcal{R}$ is defined by the specific set of matrices J, K, L, M, N, P, Q, R and S used in (1).*

$$\mathcal{R} \triangleq (J, K, L, M, N, P, Q, R, S) \quad (6)$$

The coefficients can also be regrouped into one matrix Z:

$$Z \triangleq \begin{pmatrix} -J & M & N \\ K & P & Q \\ L & R & S \end{pmatrix} \quad (7)$$

and $\mathcal{R}$ can be defined by $\mathcal{R} := (Z, l, m, n, p)$ where l, m, n and p are the matrix dimensions given above.

Moreover, equivalent structured realizations can be defined through block diagonal similarity transform [6]:

$$Z_1 = \begin{pmatrix} \mathcal{Y} & & \\ & \mathcal{U}^{-1} & \\ & & I_p \end{pmatrix} Z_0 \begin{pmatrix} \mathcal{W} & & \\ & \mathcal{U} & \\ & & I_m \end{pmatrix} \quad (8)$$

where $\mathcal{Y}$, $\mathcal{U}$ and $\mathcal{W}$ are invertible matrices.

B. Criterion Analysis

In order to evaluate how much the digital implementations modify filters' characteristics, the I/O sensitivity measure is introduced. Let us consider the state-space system (A, B, C, D). Measure of the transfer function sensitivity through its L_2-norm is one way pointed in [7]:

$$M_{L_2} \triangleq \left\| \frac{\partial H}{\partial A} \right\|_2^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C} \right\|_2^2 + \left\| \frac{\partial H}{\partial D} \right\|_2^2 . \quad (9)$$

This measure can be generalized as follows, to the context of the SIF. Considering that the coefficients quantized without error make no contribution to the overall I/O sensitivity measure, a weighting matrix W_Z associated with Z is introduced:

$$(W_Z)_{i,j} \triangleq \begin{cases} 0, & \text{with } Z_{i,j} \in \{0, \pm 1\}; \\ 1, & \text{otherwise.} \end{cases} \quad (10)$$

Consider a realization $\mathcal{R} := (Z, l, m, n, p)$ with an associated weighting matrix W_Z. The I/O transfer function sensitivity is then defined (in the SISO case) by:

$$M_{L_2}^W \triangleq \left\| \frac{\partial H}{\partial Z} \times W_Z \right\|_2^2 \quad (11)$$

where $\times$ is the Schur product.

The following lemma describes the sensitivity with regard to each matrix of the implicit form.

Lemma 1 *The sensitivity with regard to each matrix involved in the SIF can be expressed as:*

$$\frac{\partial H}{\partial Z} = H_1^\top H_2^\top \quad (12)$$

with

$$\begin{cases} H_1 : z \mapsto C_Z(zI_n - A_Z)^{-1}M_1 + M_2 \\ H_2 : z \mapsto N_1(zI_n - A_Z)^{-1}B_Z + N_2 \\ M_1 = \begin{pmatrix} KJ^{-1} & I_n & 0 \end{pmatrix}, M_2 \triangleq \begin{pmatrix} LJ^{-1} & 0 & 1 \end{pmatrix} \\ N_1 \triangleq \begin{pmatrix} M^\top J^{-\top} & I_n & 0 \end{pmatrix}^\top, N_2 \triangleq \begin{pmatrix} N^\top J^{-\top} & 0 & 1 \end{pmatrix}^\top \end{cases} \quad (13)$$

Another measure based on pole sensitivity is also commonly used. During the quantization process, the Z matrix is perturbed to $Z + \varepsilon \times W_Z$ where ε represents digital perturbations. Hence, the poles of the implemented realization may be shifted outside the unit circle even if the initial realization is stable. Based on this consideration, a stability measure is proposed as [8]:

$$\mu_0(Z) = \inf_{\varepsilon} \left\{ \|\varepsilon\|_{\max} / Z + \varepsilon \times W_Z \text{ instable} \right\} \quad (14)$$

As this measure is difficult to evaluate, the following measure is most often used [6]:

$$\mu(Z) \triangleq \min_{1 \leqslant k \leqslant n} \frac{1 - |\lambda_k|}{\|W_Z\|_F \left\| \frac{\partial |\lambda_k|}{\partial Z} \times W_Z \right\|_F} \quad (15)$$

where λ_k denote the poles of the system, and $\|.\|_F$ represents the Frobenius norm.

The pole sensitivity describes how close the poles are to the unit circle and how sensitive they are to the parameter perturbation.

Another criterion allowing to analyze the relevance of a realization for implementation is the round-off noise gain (RNG).

It is possible to aggregate all noises, denoted as $\xi(k)$, (usually modeled as independent white sequences) corrupting T, X and Y as an additive noise $\xi'(k)$ on the output. This (colored) noise results from the filtering of $\xi(k)$ added respectively on the intermediate variables, the state and the output, through the transfer function H_1 defined in (13) (cf. Fig.1).

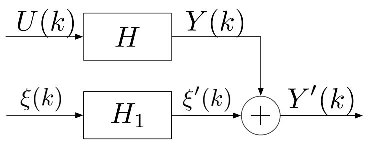

Fig. 1. Equivalent noised model

The round-off noise gain is then defined by:

$$G = \text{trace}\left(d_Z (M_1^\top W_o M_1 + M_2^\top M_2) \right) \quad (16)$$

where d_Z is a diagonal matrix with $(d_Z)_{i,i}$ defined as the number of non-trivial parameters in the i^{th} row of Z (except 0, ± 1 and powers of 2) and W_o is the observability Gramian of the state-space system (A_Z, B_Z, C_Z, D_Z). See [9], [10] for proof.

III. $L2$-SCALING & THE MODAL FORM

A. Relaxed L_2-scaling

The L_2-dynamic-range scaling constraints have been introduced by Jackson in [11] and Hwang in [12]. It consists in scaling the state variables in a way to prevent overflows or underflows. Furthermore, L_2-scaling also contributes to normalize the format of different state variables and the sensitivity criteria mentioned above.

Moreover, a SISO state-space system (A, B, C, D) is said to be L_2-scaled if the transfer functions from input to each state have a unitary L_2-norm ($1 \leqslant i \leqslant n$):

$$\left\| e_i^\top (zI - A)^{-1} B \right\|_2 = 1 \tag{17}$$

with e_i denoting the i^{th} elementary vector, whose elements are all zeros except the i^{th} which is one.

Recently in [13], new dynamic-range-scaling constraints have been proposed. It appears that, in fixed-point format, the classical L_2-scaling constraints may be uselessly too strict. It is enough, for preventing overflows, to force all state and intermediate variables to possess the same binary-point position, say the same as the input.

A fixed-point position is represented according to Fig. 2. β is the total word-length in bits of the representation, whereas γ is the fractional part word-length (it gives the binary-point position of the representation). They are fixed for each variable and coefficient, and implicit, unlike the floating-point representation. In this paper, β and γ will be suffixed by the variable/state/coefficient they refer to.

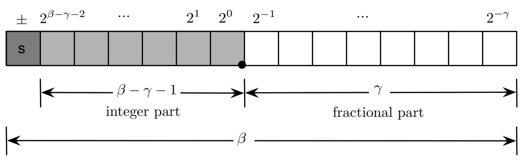

Fig. 2. Fixed-point representation

To represent a value x without overflow, a fixed-point representation (β_x, γ_x) may satisfy:

$$\beta_x - \gamma_x - 1 \geqslant \lfloor \log_2 |x| \rfloor + 1 \tag{18}$$

where the $\lfloor a \rfloor$ operator rounds a to the nearest integer lower or equal to a.

It has been proved that the overflows are avoided if the binary-point position of each state X_i is carefully chosen such that

$$\gamma_{X_i} = \beta_{X_i} - 2 - \left\lfloor \log_2 \overset{\max}{X_i} \right\rfloor, \tag{19}$$

where the upper bound $\overset{\max}{X_i}$ can be obtained by a L_2-norm estimation

$$\overset{\max}{X_i} \simeq \kappa \left\| e_i^\top (zI - A)^{-1} B \right\|_2 \overset{\max}{U} \tag{20}$$

$\overset{\max}{U}$ is the maximum amplitude of the input and κ can be interpreted as a representation of the number of standard deviation of E_i, if the input is unit-variance white centered noise ($\kappa \geqslant 1$). Since the L_2-norm estimation in (20) does not give a strict bound, κ can be seen as a safety parameter. A L_1-norm can also be used, but it is often too much conservative and less tractable.

The idea of the scaling is to choose a binary-point position for each state (usually the same as for the input), and to apply a scaling on them so as to adapt the peak values of each state to the chosen binary-point positions.

Whereas the classical L_2-scaling imposes $\overset{\max}{U} = \overset{\max}{X_i}$ (that leads to eq. (17)), we can deduce that it is sufficient to choose all the fixed-point positions to be equal to the one of the input, (i.e. $\gamma_{X_i} = \gamma_U$) and to scale accordingly, in order to prevent overflow [13].

In the case where the word length of all variables is equal (*i.e.* $\beta_U = \beta_{X_i}$), $\overset{\max}{U}$ a power of 2 and κ set equal to 1, the classical L_2-scaling is replaced by a relaxed-L_2-scaling

$$1 \leqslant (W_c)_{i,i} < 4 \tag{21}$$

This is here extended to the SIF framework, with ($1 \leqslant i \leqslant n, 1 \leqslant j \leqslant l$):

$$\overset{\max}{X_i} = \kappa \left\| e_i^\top (zI - A_Z)^{-1} B_Z \right\|_2 \overset{\max}{U} \tag{22}$$

$$\overset{\max}{T_j} = \kappa \left\| e_j^\top \left(J^{-1} M (zI - A_Z)^{-1} B_Z + J^{-1} N \right) \right\|_2 \overset{\max}{U} \tag{23}$$

These L_2-norm can be computed by the controllability Gramians associated with the state and intermediate variables respectively:

$$\left\{ \begin{array}{l} W_{cX} = A_Z W_{cX} A_Z^\top + B_Z B_Z^\top \\ W_{cT} = J^{-1} M W_{cX} M^\top J^{-\top} + J^{-1} N N^\top J^{-\top} \end{array} \right. \tag{24}$$

Proposition 1 (Relaxed L_2-scaling constraints) *In order to implement input and the descriptor variables with the same binary-point position, it makes sense to relax the L_2-scaling constraints as ($1 \leqslant i \leqslant n, 1 \leqslant j \leqslant l$):*

$$\left\{ \begin{array}{l} \frac{2^{2\alpha_{X_i}}}{\kappa^2} \leqslant (W_{cX})_{i,i} < 4 \frac{2^{2\alpha_{X_i}}}{\kappa^2} \\ \frac{2^{2\alpha_{T_j}}}{\kappa^2} \leqslant (W_{cT})_{j,j} < 4 \frac{2^{2\alpha_{T_j}}}{\kappa^2} \end{array} \right. \tag{25}$$

where

$$\left\{ \begin{array}{l} \alpha_{X_i} = \beta_{X_i} - \beta_U - \mathscr{F}_2\left(\overset{\max}{U} \right) \\ \alpha_{T_j} = \beta_{T_j} - \beta_U - \mathscr{F}_2\left(\overset{\max}{U} \right) \end{array} \right. \tag{26}$$

and $\mathscr{F}_2(x)$ is defined as the fractional value of $\log_2(x)$*:*

$$\mathscr{F}_2(x) \triangleq \log_2(x) - \lfloor \log_2(x) \rfloor \tag{27}$$

Proof: γ_U is given by $\gamma_U = \beta_U - 2 - \left\lfloor \log_2 \overset{\max}{U} \right\rfloor$, so $\gamma_U = \gamma_{X_i}$ leads to

$$\beta_U - \left\lfloor \log_2 \overset{\max}{U} \right\rfloor = \beta_{X_i} - \left\lfloor \log_2 \left(\kappa \left\| e_i^\top (zI_n - A)^{-1} B \right\|_2 \overset{\max}{U} \right) \right\rfloor \tag{28}$$

and

$$\left\lfloor \log_2\left(\kappa\sqrt{(W_{cX})_{i,i}}\right) + \mathscr{F}_2\left(\overset{\max}{U}\right)\right\rfloor = \beta_{X_i} - \beta_U \quad (29)$$

So

$$2^{\alpha_{X_i}} \leqslant \kappa\sqrt{(W_{cX})_{i,i}} < 2^{\alpha_{X_i}+1} \quad (30)$$

The same applies to the intermediate variables, with

$$\overset{\max}{T_j} = \kappa\sqrt{(W_{cT})_{j,j}}\,\overset{\max}{U}. \quad (31)$$

■

Corollary 1 *For micro-controller or DSP implementation, the word length of all variables are equal, i.e.* $\beta_{X_i} = \beta_{T_j} = \beta_U$*. Also* $\overset{\max}{U}$ *could be set to a power of* 2*. Then, if* $\kappa = 1$ *(as for classical* L_2*-scaling constraints), the relaxed* L_2*-scaling constraints become:*

$$1 \leqslant (W_{cX})_{i,i} < 4, \quad 1 \leqslant (W_{cT})_{j,j} < 4, \quad (32)$$

It is noteworthy that relaxing constraints give additional degrees of freedom to get optimal realizations, which is used in Section IV to propose a practical way for implementation.

Proposition 2 (*a posteriori* relaxed L_2-scaling) *Considering a given SIF realization, it is possible to get an equivalent one meeting constraints (25) (i.e. relaxed* L_2*-scaled one), just by applying the similarity transform (see (8)) with* $\mathcal{U}$ *and* $\mathcal{W}$ *diagonal matrices such that:*

$$\mathcal{U}_{i,i} = \kappa\sqrt{(W_{cX})_{i,i}}\,2^{-\mathscr{F}_2\left(\sqrt{(W_{cX})_{i,i}}\right)-\alpha_{X_i}} \quad (33)$$

$$\mathcal{W}_{j,j} = \kappa\sqrt{(W_{cT})_{j,j}}\,2^{-\mathscr{F}_2\left(\sqrt{(W_{cT})_{j,j}}\right)-\alpha_{T_j}} \quad (34)$$

Not specified, yet $\mathcal{Y}$ *should be chosen to preserve the structure of the realization, for example* $\mathcal{Y} = I_n$ *or* $\mathcal{Y} = \mathcal{W}^{-1}$*.*

Proof: See [13]. For $x \in \mathbb{R}$, $\bar{x} \triangleq 2^{\mathscr{F}_2(x)+a}$ is such that $2^a \leqslant \bar{x} < 2^{a+1}$, and this could be applied to eq. (25). ■

B. Modal Representation

Let us consider the transfer function $H(z)$ and its related modal realization (Λ, B, C, D):

$$\begin{aligned} H(z) &= D + C(zI - \Lambda)^{-1}B \\ &= D + \sum_{i=1}^{n} \frac{c_i b_i}{1 - \lambda_i z^{-1}} \end{aligned} \quad (35)$$

with $\lambda_i \neq \lambda_j$ for all $i \neq j$ so that Λ may be chosen as a diagonal matrix.

Rather to diagonalize the A-matrix, it is preferred in the sequel to combine the complex-conjugate pole-pairs to form a real "block-diagonal" section in which Λ has two-by-two real matrices along its diagonal as follows:

$$\Lambda = \begin{pmatrix} \alpha_1 & \beta_1 & & & & & \\ \beta_2 & \alpha_2 & & & & & \\ & & \alpha_3 & \beta_3 & & & \\ & & \beta_4 & \alpha_4 & & & \\ & & & & \ddots & & \\ & & & & & \alpha_{n-1} & \beta_{n-1} \\ & & & & & \beta_n & \alpha_n \end{pmatrix} \quad (36)$$

where α_i and β_i are linked to the real part and the imaginary part of the i^{th} pole, respectively.

Such a modal form has a low pole sensitivity: perturbation on coefficients of (36) affects directly the poles so that a tiny deviation will not lead to a dramatic change on poles. Moreover it also performs quite well in terms of quantization noises because the whole system is decomposed and realized by several parallel low-order sections which are less sensitive to quantization noises [14].

IV. A PARTICULAR ρ-MODAL REALIZATION

The sparse ρ-modal realization is introduced in this section. The proposed transformation preserves the dynamic scaling while improving the implementation.

Let us first define the following sequence of 1^{st}-order polynomial operators, named ρ-operators:

$$\rho_i = \frac{q - \gamma_i}{\Delta_i}, \quad 1 \leqslant i \leqslant n \quad (37)$$

$\{\gamma_i\}$ and $\{\Delta_i > 0\}$ are two sets of constants to determine.

The specialized implicit form related to the ρ-operator has the following particular structure:

$$\begin{pmatrix} I & 0 & 0 \\ -\Delta & I & 0 \\ 0 & 0 & I \end{pmatrix}\begin{pmatrix} T(k+1) \\ X(k+1) \\ Y(k) \end{pmatrix} = \begin{pmatrix} 0 & A_\rho & B_\rho \\ 0 & \gamma & 0 \\ 0 & C_\rho & D_\rho \end{pmatrix}\begin{pmatrix} T(k) \\ X(k) \\ U(k) \end{pmatrix} \quad (38)$$

with

$$A_\rho = \Delta^{-1}(\Lambda - \gamma), B_\rho = \Delta^{-1}B, C_\rho = C \text{ and } D_\rho = D \quad (39)$$

$$\Delta = \text{diag}(\Delta_1, \cdots, \Delta_n), \quad \gamma = \text{diag}(\gamma_1, \cdots, \gamma_n) \quad (40)$$

Proposition 3 (ρ-modal realization) *The algorithm to establish the* ρ*-modal realization is depicted as follows:*

1) *State with a* q*-based modal realization;*
2) L_2*-scale it according to Proposition 2;*
3) *Build the equivalent* ρ*-realization described in (38) and (39), then choose appropriate* $\{\gamma_i\}$ *to minimize the FWL effects;*
4) *Deduce* $\{\Delta_i\}$ *to* L_2*-scale the intermediate variables.*

Steps 3) and 4) will be detailed later, and the results are given in Proposition 4 and the equation (43).

Otherwise, according to the number of real and complex poles, the number of inexactly-implemented coefficients is between $3n + 1$ and $4n + 1$, plus $2n$ free parameters $\{\Delta_i\}$ and $\{\gamma_i\}$.

To make more precise the way to perform the relaxed L_2-scaling on both the state and the intermediate variables, let us reconsider the Gramian's equation (24)

$$\begin{cases} W_{cX} = \Lambda W_{cX} \Lambda^\top + BB^\top \\ W_{cT} = \Delta^{-2}[(\Lambda - \gamma) W_{cX} (\Lambda - \gamma)^\top + BB^\top] \end{cases} \quad (41)$$

The relaxed L_2-scaling of the intermediate variables $T(k)$ may be fulfilled by choosing appropriate $\{\gamma_i\}$ and $\{\Delta_i\}$ sets. It is clear from (41) that this choice does not modify W_{cX}, since $X(k)$ is nothing else than the state of the q-L_2-scaled modal realization obtained in step 2) of Proposition 3.

In order to explain the way to get a scaled $T(k)$, let us denote $\widetilde{W_{cT}}$ as the solution of W_{cT} in (41) when $\Delta = I_n$. Consequently the condition $1 \leqslant (W_{cT})_{i,i} < 4$ is equal to:

$$1 \leqslant \Delta_i^{-2}(\widetilde{W_{cT}})_{i,i} < 4 \tag{42}$$

Obviously all $\Delta_i \in \left]\frac{1}{2}\sqrt{(\widetilde{W_{cT}})_{i,i}}, \sqrt{(\widetilde{W_{cT}})_{i,i}}\right]$ achieves the relaxed L_2-scaling. So given the set of $\{\gamma_i\}$, it is always possible to choose $\{\Delta_i\}$ as a power of 2 according to the following expression for assuring the relaxed L_2-scaling.

$$\Delta_i = 2^{\left\lfloor \sqrt{(\widetilde{W_{cT}})_{i,i}} \right\rfloor} \tag{43}$$

The choice of the $\{\gamma_i\}$ will be further discussed in the next section.

V. PROPERTIES ANALYSIS

The properties of the proposed ρ-modal realization are investigated in this section. Furthermore an analytical solution to the optimal ρ-modal realization is exhibited. The proofs of propositions and corollaries in this section are omitted by lake of place.

A. Transfer Function Sensitivity Minimization

According to (43), Δ_i can be implemented exactly (e.g. as a power of 2), and from (39) the modification of γ_i only affects the sensitivity of matrices A_ρ, B_ρ in (38). With (12), the sensitivity with respect to these two matrices are developed as follows:

$$\frac{\partial H}{\partial A_\rho} = \left(C(zI-\Lambda)^{-1}\Delta\right)^\top \left((zI-\Lambda)^{-1}B\right)^\top \tag{44}$$

$$\frac{\partial H}{\partial B_\rho} = \left(C(zI-\Lambda)^{-1}\Delta\right)^\top \tag{45}$$

These two transfer matrices can be respectively represented by the following state-space systems:

$$\left(\begin{pmatrix} \Lambda & BC \\ 0 & \Lambda \end{pmatrix}, \begin{pmatrix} 0 \\ \Delta \end{pmatrix}, \begin{pmatrix} I_n & 0 \end{pmatrix}, 0 \right) \quad , \tag{46}$$

$$\left(\Lambda, \Delta, C, 0\right). \tag{47}$$

Considering these two systems, it is quite clear that minimizing the L_2-norm of $\frac{\partial H}{\partial A_\rho}$ and $\frac{\partial H}{\partial B_\rho}$ requires to take Δ_i as small as possible. Due to (42), it appears that the minimization of Δ_i is linked to the minimization of $(\widetilde{W_{cT}})_{i,i}$.

Proposition 4 (Optimal $\{\gamma_i\}$) *Consider the ρ-modal realization. In this context, the best choice for the γ_i in order to minimize the diagonal terms of $\widetilde{W_{cT}}$ is given by:*

$$\gamma_i = \begin{cases} \alpha_i + \dfrac{\beta_i \left(W_{cX}\right)_{i+1,i}}{\left(W_{cX}\right)_{i,i}}, & i \text{ is odd;} \\ \alpha_i + \dfrac{\beta_i \left(W_{cX}\right)_{i,i-1}}{\left(W_{cX}\right)_{i,i}}, & i \text{ is even.} \end{cases} \tag{48}$$

Corollary 2 *The choice of γ_i in (48) leads to the lowest I/O sensitivity.*

Remark 1 When the sampling frequency is high relatively to the cut-off frequency of system considered (analog filter or process), the poles of the resulting digital transfer are very close to 1 and their imaginary parts are tiny. Consequently the value of γ_i shown as (48) will be close to one. Hence, δ-operator leads to similar results as ρ-operator when a narrow low-pass filter is implemented (see *Example I* in Section VI).

B. Pole Sensitivity Minimization

The pole sensitivity is an interesting indicator to analyze the FWL effects during the digital implementation. Its definition under the specialized implicit form is given in (15).

Proposition 5 *With the ρ-modal realization, the optimization of the pole sensitivity requires to choose Δ_i as small as possible.*

C. Round-off Noise Gain Minimization

Owing to its parallel structure and second-order/first-order sub-sections, the modal realization is less sensitive to the quantization noises. According to (16) the round-off noise gain of the ρ-modal realization is

$$G = \text{trace}\Big(d_Z \begin{pmatrix}\Delta & I_n & 0\end{pmatrix}^\top W_o \begin{pmatrix}\Delta & I_n & 0\end{pmatrix} + \begin{pmatrix}0 & 0 & 1\end{pmatrix}^\top \begin{pmatrix}0 & 0 & 1\end{pmatrix}\Big) \tag{49}$$

As W_o is the observability Gramian of (Λ, B, C, D), it is constant with respect to Δ_i. The minimization of G is hence equal to choose Δ_i as small as possible.

Accordingly, the proposed ρ-modal realization has the nice feature to minimize the tree criteria simultaneously by using a unique condition.

VI. NUMERICAL EXAMPLES

In this section, two numerical examples (see [2]) are presented to illustrate the performance of the proposed realization, confirm the theoretical results of Section V, and also compare it with some existing methods.

In these examples, L_2-dynamic-range scaling is applied, and the optimal $\{\gamma_i\}$ (given by (48)) for the ρ-modal realization is rounded to the nearest exactly-implemented number. Here, only 5 bits are used to represent $\{\gamma_i\}$.

Four different realizations are engaged:

- Z_1: Cascade form with q-based second-order companion canonical sections
- Z_2: Optimized ρ-based direct-form II transposed[2] [2]
- Z_3: δ-modal realization, with $\{\gamma_i = 1\}$
- Z_4: ρ-modal realization

Numerical simulations are launched by using the FWR Toolbox[3] developed with MATLAB, and the I/O transfer function sensitivity is chosen as the criterion to optimize Z_2.

Example I: This is a fourth-order low-pass Butterworth filter with narrow bandwidth, generated by the MATLAB command *butter*$(4, 0.05)$. Its normalized bandwidth is 0.025,

[2] ρDFIIt is evaluated by the methods proposed in [2]. It is under the strict L_2-scaling constraints, without L_2-scaling on intermediate variables.

[3] Source available at http://fwrtoolbox.gforge.inria.fr

TABLE I
PERFORMANCE COMPARISON OF 4 REALIZATIONS OF EXAMPLE I

Realization	$M_{L_2}^W(Z)$	$\mu(Z)$	$G(Z)$	$N.+$	$N.\times$
Z_1	469.34	1.2326	11.277	8	12
Z_2	7.1600	0.5848	5.0231	12	16
Z_3	12.051	0.5952	5.2839	16	20
Z_4	7.1048	0.2221	6.8033	16	25

TABLE II
PERFORMANCE COMPARISON OF 4 REALIZATIONS OF EXAMPLE II

Realization	$M_{L_2}^W(Z)$	$\mu(Z)$	$G(Z)$	$N.+$	$N.\times$
Z_1	48.899	2.6938	5.4092	12	18
Z_2	26.369	9.1963	16.405	18	25
Z_3	79.921	10.500	21.609	24	30
Z_4	17.299	1.5880	11.523	24	34

and the corresponding poles are $\lambda_{1,2} = 0.9319 \pm j0.1364$, and $\lambda_{3,4} = 0.8630 \pm j0.0523$ that are clustered around $z = 1$.

For this example, the ρ-modal realization yields a transfer function sensitivity of 7.1048. The corresponding pole sensitivity and round-off noise gain are 0.2221 and 6.8033, respectively. The set of γ is computed as:

$$\gamma_{opt} = (15, 15, 15, 13) \times 2^{-4}$$

Example II: The second example is a sixth-order passband Butterworth filter which is obtained by the MATLAB command $butter(3, [0.75\ 0.90])$. Its poles are $\lambda_{1,2} = 0.6237 \pm j0.5747$, $\lambda_{3,4} = 0.8809 \pm j0.2937$ and $\lambda_{5,6} = 0.7071 \pm j0.3359$. These poles, in contrast to the first filter, are not longer clustered around $z = 1$.

The optimal I/O sensitivity measure obtained with the proposed realization is 17.299. The corresponding pole sensitivity and round-off noise gain are 1.588 and 11.523, respectively. The set of ρ is computed as:

$$\gamma_{opt} = (-9, -11, -13, -15, -19, -14) \times 2^{-4}$$

The δ-modal realization yields quite similar results as the ρ-modal one in Example I. This is coherent with Remark 1. This is however not any longer true for Example II. The expression of γ_i in (48) highlights this point. When using a high sampling frequency compared to the filters dynamics (i.e. a small discrete cut-off frequency in the command $butter$), the real part of the poles tends to one while the imaginary part tends to zero, hence bringing γ_i tending towards 1 and operator ρ tending towards δ.

Table I and II reveal the globally improved numerical properties of the ρ-modal realization, with a comparable number of computation. It should be noticed however that the ρ-modal realization is not the optimal one (but what is "optimal" in such a case with multi-objective), but it always achieves a good trade-off among the different criteria. Even in the case of oversampling, the results are good, and for example much better than those achieved with the ρDFIIt realization.

VII. CONCLUSION

This paper deals with the FWL implementation problem of digital LTI filters/controllers. Its principal contribution consists in the proposition of a systematic way to get a realization managing well the compromises between the different aspects for making a desirable FWL implementation. The proposed structure is quite easy to get. It is deduced from a modal realization and the use of a ρ-operator adapted to each mode. Among its main features, it has a sparse parameterization, leading to a low computational effort (the number of coefficients is less than $4n + 1$ for an n^{th}-order filter/controller), small pole and I/O parametric sensitivity (to the FWL degradation), as well as a small round-off noise gain. All of these are obtained under the relaxed L_2-scaling constraints allowing to normalize the intermediate computational variables (including the state), and to limit the risk of overflow as well. Last but not least, contrary to other works asking some tricky non-linear optimization, the method to develop the proposed realization requires the optimization of only parameters whose optimum can be obtained analytically.

REFERENCES

[1] P. Chevrel, "Discrétisation et codage des régulateurs en vue de leur maintenance et de leur réglage," in *Réunion du GT Commande et Robuste des Systèmes Multivariables* , Bordeaux, France, 2002.

[2] G. Li and Z. Zhao, "On the generalized dfiit structure and its state-space realization in digital filter implementation," *IEEE transaction on circuits and systems I*, vol. 51, no. 4, pp. 769–778, April 2004.

[3] J. Hao and G. Li, "An efficient controller structure with minimum round-off noise gain," *Automatica*, vol. 43, no. 5, pp. 921–927, 2007.

[4] A. Madievski, B. Anderson, and M. Gevers, "Optimum realizations of sampled-data controllers for fwl sensitivity minimization," *Automatic*, vol. 31, no. 3, pp. 367–379, 1995.

[5] W. Yan and K. Teo, "Optimal finite-precision approximation of *FIR* filers," *Signal Processing*, vol. 82, no. 11, pp. 1695–1705, 2002.

[6] T. Hilaire, P. Chevrel, and J. Whidborne, "A unifying framework for finite wordlength realizations," *IEEE transaction on Circuits and Systems*, vol. 54, no. 8, pp. 1765–1774, August 2007.

[7] M. Gevers and G. Li, *Parameterizations in Control, Estimation and Filtering Problems*. Springer-Verlag, 1993.

[8] I. Fialho and T. Georgiou, "On stability and performance of sampled-data systems subject to wordlength constraint," *IEEE Trans. Automatic Control*, vol. 39, no. 12, pp. 2476–2481, December 1994.

[9] C. Mullis and R. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. on Circuits and Systems*, vol. CAS-23, pp. 551–562, 1976.

[10] T. Hilaire, D. Mnard, and O. Sentieys, "Roundoff noise analysis of finite wordlength realizations with the implicit state-space framework," in *15th European Signal Processing Conference (EUSIPCO07)*, September 2007.

[11] L. Jackson, "Dynamic-range constraint in state-space digital filtering," *IEEE Trans. On Acoustics, Speech and Signal Processing*, pp. 591–593, December 1975.

[12] S. Hwang, "Roundoff noise analysis for fixed-point digital filters realized in cascade or parralel form," *IEEE Trans. Audioand Elec.*, vol. 18, no. 2, August 1970.

[13] T. Hilaire, "Low parametric sensitivity realizations with relaxed-L_2-dynamic-range-scaling constraints," *IEEE transaction on circuits and systems II, to appear*, 2009.

[14] A. Oppenheim, R. Schafer, and J. Buck, *Discrete-Time Signal Processing (Second Edition)*. Prentice-Hell, 1999.