ROUNDOFF NOISE ANALYSIS OF FINITE WORDLENGTH REALIZATIONS WITH THE IMPLICIT STATE-SPACE FRAMEWORK

Thibault HILAIRE, Daniel MÉNARD, Olivier SENTIEYS

IRISA (UMR CNRS 6074) - R2D2 Team
LANNION, France
{firstname.lastname}@irisa.fr

ABSTRACT

The analytic evaluation of the system output roundoff noise is an interesting approach to analyze the effect of a Finite Word Length implementation of linear filters. Previous works have introduced the Roundoff Noise Gain *in this context and applied it to shift and δ-realizations.*
To generalize them, the paper is based on a more general representation and exhibits the output noise power in the general case.
Finally, the problem optimal realization problem, according to the Roundoff Noise Gain *measure, is considered.*

1. INTRODUCTION

The majority of signal processing systems (and also control systems) are digitally implemented with general purpose microprocessors, DSP or specific computing devices like FPGA. Since the processor cannot compute with infinite precision and use most of the time fixed-point arithmetic, the Finite Word Length (FWL) implementation of signal processing algorithms leads to deterioration in performance. This induced deterioration has two separate origins [1]:

- the quantization of embedded coefficients,
- the roundoff errors in the numerical computations.

They can respectively be formalized as parametric errors and numerical noises. If the filter and the target architecture are specified, they both depend on the realization (the parameters of the mathematical algorithm used to implement the filter), the fixed-point representation of the variables and coefficients, and the software/hardware design.
Numerous works in the control and filtering communities have been done ([1, 7, 17, 9, 16], etc.) on the parametric errors (transfer function sensitivity, etc.) and the search of *optimal* realizations (relatively to different criteria).

The roundoff errors have been studied in two different ways: works in [4, 13, 1] deal with a roundoff noise measure (the Roundoff Noise Gain) unlinked to hardware considerations in order to optimize the realization with respect to that criteria, whereas works in [11, 8] deal with the Signal Quantization Noise Ratio and are more focused on the software/hardware realizations.

Various algorithms exist to numerically realize a linear time invariant filter. As shown in section 5, they present various computational cost and FWL comportment (evaluate here with the RNG). In order to compare and find optimal realizations, our objective is to propose an analytic, efficient and general measure to evaluate the output noise power of various possible implementation schemes. It allows to answer the optimal design problem and help to solve the computation cost and roundoff noise errors trade-off. Due to a lack of place, only the RNG scheme is studied here.

This paper is organized as follows. After presenting the implicit state-space framework in section 2, a general roundoff noise analysis, applied to this form, is exhibited in section 3. The classical RNG scheme is then derived. Section 4 presents the optimal design problems and some numerical results are finally proposed in section 5.

2. THE IMPLICIT STATE-SPACE FRAMEWORK

Work in [2] highlights the interest of the implicit state-space representation in the context of FWL implementation problems and proposes to use a specialized form directly connected to the in-line computations to be performed. It can be used as a unifying framework to allow a more detailled (but macroscopic) description of FWL implementations. Various realizations, like q (shift) or δ-realizations, classical Direct Form I and II, cascade or parallel decompositions, mixed structures, etc. may be then described in a single unifying form.
Equation (1) recalls the specialized implicit form that explicitly expresses the parametrization and the intermediate variables used:

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

where

- $J \in \mathbb{R}^{l\times l}$, $K \in \mathbb{R}^{n\times l}$, $L \in \mathbb{R}^{p\times l}$, $M \in \mathbb{R}^{l\times n}$, $N \in \mathbb{R}^{l\times m}$, $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times m}$, $R \in \mathbb{R}^{p\times n}$, $S \in \mathbb{R}^{p\times m}$, $T(k) \in \mathbb{R}^l$, $X(k) \in \mathbb{R}^n$, $U(k) \in \mathbb{R}^m$ and $Y(k) \in \mathbb{R}^p$,
- $U(k)$ represents the m inputs, and $Y(k)$ the p outputs,
- $X(k+1)$ is the n stored states ($X(k)$ is effectively stored from one step to the next, in order to compute $X(k+1)$ at step k),
- $T(k+1)$ is the l intermediate variables 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 characterizes the concept of intermediate variables),
- the matrix J is lower triangular with 1 on the diagonal,

$T(k+1)$ 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 inside one time step.

It is implicitly considered throughout the paper that the computations associated to the realization (1) are ordered from top to bottom, associated in a one to one manner to the following algorithm:

(i) intermediate variable computation: J is lower triangular, so $T_0(k+1)$ is first calculated, and then $T_1(k+1)$ using

$T_0(k+1)$ and so on (the computation of J^{-1} is not necessary):
$J.T(k+1) \leftarrow M.X(k)+N.U(k)$

(ii) state-vector update:
$X(k+1) \leftarrow K.T(k+1)+P.X(k)+Q.U(k)$

(iii) outputs computation:
$Y(k) \leftarrow L.T(k+1)+R.X(k)+S.U(k)$

Steps (ii) and (iii) can be swapped: the computational delay could be reduced by evaluating $Y(k)$ first.
Equation (1) is equivalent in infinite precision to the classical state-space form:

$$\begin{pmatrix} T(k+1) \\ X(k+1) \\ \hline Y(k) \end{pmatrix} = \left(\begin{array}{cc|c} 0 & J^{-1}M & J^{-1}N \\ 0 & A_Z & B_Z \\ \hline 0 & C_Z & D_Z \end{array}\right) \begin{pmatrix} T(k) \\ X(k) \\ \hline U(k) \end{pmatrix} \quad (2)$$

with $A_Z \in \mathbb{R}^{n\times n}$, $B_Z \in \mathbb{R}^{n\times m}$, $C_Z \in \mathbb{R}^{p\times n}$ and $D_Z \in \mathbb{R}^{p\times m}$ and where

$$\begin{array}{rclrcl} A_Z &=& KJ^{-1}M+P & B_Z &=& KJ^{-1}N+Q \\ C_Z &=& LJ^{-1}M+R & D_Z &=& LJ^{-1}N+S \end{array} \quad (3)$$

However, equation (2) corresponds to a different parametrization than the one in eq. (1).
The equivalent transfer function considered is then given by

$$H : z \mapsto C_Z(zI_n - A_Z)^{-1}B_Z + D_Z \quad (4)$$

In the following, a realization $\mathscr{R}$ will be defined in the implicit form by its parameters used for the internal description

$$\mathscr{R} \triangleq (J,K,L,M,N,P,Q,R,S) \quad (5)$$

It could also be equivalently written in a compact form $\mathscr{R} = (Z,l,m,n,p)$ with

$$Z \triangleq \begin{pmatrix} -J & M & N \\ K & P & Q \\ L & R & S \end{pmatrix} \quad (6)$$

The usual realizations (Direct Forms, state-space, δ-realizations, cascade, parallel, mixed realization, etc.) can be easily expressed in the Implicit State-Space Framework. For example, a δ-state-space realization

$$\left\{ \begin{array}{rcl} \delta[X(k)] &=& A_\delta X(k)+B_\delta U(k) \\ Y(k) &=& C_\delta X(k)+D_\delta U(k) \end{array} \right. \quad (7)$$

where $\delta = \frac{q-1}{\Delta}$ (Δ is strictly positive constant, q is the shift-operator), can be expressed as:

$$\begin{pmatrix} I_n & 0 & 0 \\ -\Delta I_n & I_n & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} T(k+1) \\ X(k+1) \\ Y(k) \end{pmatrix} = \begin{pmatrix} 0 & A_\delta & B_\delta \\ 0 & I_n & 0 \\ 0 & C_\delta & D_\delta \end{pmatrix} \begin{pmatrix} T(k) \\ X(k) \\ U(k) \end{pmatrix} \quad (8)$$

This form is well known [1, 12] to be numerically superior to the usual shift-operator, because it generally results in less sensitive implementation with less roundoff noise. Other examples can be found in [2].

3. OUTPUT NOISE POWER

3.1 Preliminaries

Let G be a MIMO[1] $l \times m$ system and U a noise to be propagated in ($U(k) \in \mathbb{R}^l$).
In that paper, the first (μ) and second (σ,Ψ) order moments of a noise b (it may be a vector of noises) are defined by

$$\mu_b \triangleq E\{b(k)\} \quad (9)$$

$$\Psi_b \triangleq E\left\{b(k)b^\top(k)\right\} \quad (10)$$

$$\sigma_b^2 \triangleq E\left\{b^\top(k)b(k)\right\} = tr(\Psi_b) \quad (11)$$

where $E\{.\}$ is the *mean* operator and $tr(.)$ the *trace* operator.

Definition 1 (L_2-norm, gramians) *The L_2-norm of G is defined by*

$$\|G\|_2^2 \triangleq \frac{1}{2\pi}\int_0^{2\pi} tr\left(G\left(e^{j\omega}\right)G^H\left(e^{j\omega}\right)\right)\, d\omega \quad (12)$$

where $\cdot^H$ the transpose conjugate operator. Let (A,B,C,D) be a state-space representation of G. So

$$G : z \mapsto C(zI_n - A)^{-1}B + D \quad (13)$$

and it can be shown that

$$\|G\|_2^2 = tr\left(DD^\top + CW_cC^\top\right) = tr\left(D^\top D + B^\top W_o B\right) \quad (14)$$

where W_c and W_o are respectively the controllability and observability gramians of the realization (A,B,C,D). They are solutions of the Lyapunov equations:

$$W_c = AW_cA^\top + BB^\top, \qquad W_o = A^\top W_oA + C^\top C \quad (15)$$

The following proposition is necessary to recall the properties of noises through a transfer function.

Proposition 1 *Let suppose the noise $U(k)$ to satisfy*

$$E\left\{U(k)U^\top(k-l)\right\} = \delta_{0,l}\Psi_U \quad (16)$$

where $\delta_{i,j}$ is the Kronecker symbol.
Let Y denote the resulting noise of U through G. If (A,B,C,D) is a state-space representation of G, the first and second order moments of Y are given by:

$$\mu_Y = G(0)\mu_U \quad (17)$$

$$\sigma_Y^2 = tr\left(\Psi_U(D^\top D + B^\top W_oB)\right) \quad (18)$$

where W_o is the observability Gramian of G.
Proof:
The proof for the mean can be found in [14].
The power spectrum densities Φ_U and Φ_Y satisfy ([14])

$$\Phi_Y(z) = G(z)\Phi_U(z)G^H(z) \qquad \forall z \in \mathbb{C} \quad (19)$$

[1] Mulitple Input Multiple Output

The Fourier Transform (FT) gives $FT(\delta_{0,l}) = 1$ and $FT(\Phi) = \Psi$, so

$$\begin{aligned}\sigma_Y^2 &= tr\left(\frac{1}{2\pi}\int_0^{2\pi}\Phi_Y\left(e^{j\omega}\right)\,d\omega\right)\\ &= \frac{1}{2\pi}\int_0^{2\pi}tr\left(\left(G\varphi_U\right)\left(e^{j\omega}\right)\left(G\varphi_U\right)^H\left(e^{j\omega}\right)\right)\,d\omega\\ &= \|G\varphi\|_2^2 \end{aligned} \quad (20)$$

where φ_U is such that $\Psi_U = \varphi_U\varphi_U^\top$ (Ψ_U is a symetric positive definite matrix, so a Cholesky decomposition is available). As $G\varphi_U$ and G share the same observability gramian W_o, it comes, with eq. (14):

$$\begin{aligned}\sigma_Y^2 &= tr\left(\varphi_U^\top D^\top D\varphi_U + \varphi_U^\top B^\top W_o B\varphi_U\right)\\ &= tr\left(\varphi_U\varphi_U^\top(D^\top D + B^\top W_o B)\right)\end{aligned} \quad (21)$$

■

Remark 1 $G(0)$ is the static gain of G. In the SISO[2] case ($l = m = 1$), one can find the classical results [14]:

$$\mu_Y = G(0)\mu_U, \qquad \sigma_Y^2 = \|G\|_2^2\,\sigma_U^2 \quad (22)$$

3.2 Roundoff Noise Analysis

Let us consider a realization $\mathscr{R}$ described with the implicit form (1), with transfer function H. When implemented, the steps (i) to (iii) are modified by the add of noises B_T, B_X and B_Y:

$$\begin{aligned} J.T^*(k+1) &\leftarrow M.X^*(k) + N.U(k) + B_T(k)\\ X^*(k+1) &\leftarrow K.T^*(k+1) + P.X^*(k) + Q.U(k) + B_X(k)\\ Y^*(k) &\leftarrow L.T^*(k+1) + R.X^*(k) + S.U(k) + B_Y(k)\end{aligned} \quad (23)$$

(the noise $J^{-1}B_T$ is added on T).
These noises added depend on:

- the way the computations are organized (the order of the sums) and done,
- the fixed-point representation of the inputs,
- the fixed-point representation of the outputs,
- the fixed-point representation chosen for the states and the intermediate variables.

These noises are independent white noises.

It is possible to express the implemented system as the initial system (H) with a noise B added on the ouputs: the implemented system (23) is then equivalent to the system described in figure 1, where the transfer functions H_T, H_X and H_Y are respectively the transfer functions from the noises added on the intermediate variables (on $J.T$), the states and the outputs to the outputs:

$$H_T : z \mapsto C_Z(zI_n - A_Z)^{-1}KJ^{-1} + LJ^{-1} \quad (24)$$

$$H_X : z \mapsto C_Z(zI_n - A_Z)^{-1} \quad (25)$$

$$H_Y : z \mapsto I_p \quad (26)$$

These transfer function have values in $\mathbb{C}^{l\times p}$, $\mathbb{C}^{n\times p}$ and $\mathbb{C}^{p\times p}$ respectively.

[2] Single Input Single Output

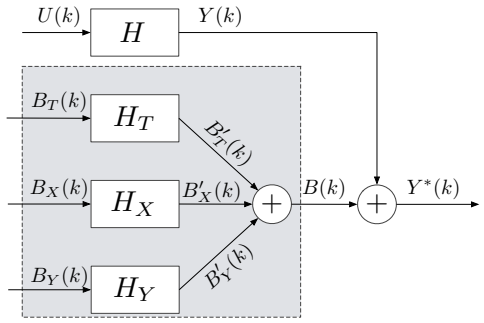

Figure 1: Equivalent system, with noises extracted

Then, the expression of the output noise power is given by

$$P \triangleq \sigma_B^2 = E\left\{B^\top(k)B(k)\right\} \quad (27)$$

with B the sum of the filtered noises B'_T, B'_X and B'_Y (see figure 1):

$$B(k) = B'_T(k) + B'_X(k) + B'_Y(k) \quad (28)$$

From eq. (27), it comes

$$\begin{aligned}P &= \sigma_{B'_T}^2 + \sigma_{B'_X}^2 + \sigma_{B'_Y}^2 + \mu_{B'_T}^\top\left(\mu_{B'_X} + \mu_{B'_Y}\right)\\ &\quad + \mu_{B'_X}^\top\left(\mu_{B'_T} + \mu_{B'_Y}\right) + \mu_{B'_Y}^\top\left(\mu_{B'_T} + \mu_{B'_X}\right)\end{aligned} \quad (29)$$

Proposition 2 *Finally, the output noise power is given by*

$$\begin{aligned}P &= tr\left(\Psi_{B_T}J^{-\top}\left(L^\top L + K^\top W_o K\right)J^{-1}\right)\\ &\quad + tr(\Psi_{B_X}W_o) + tr\left(\Psi_{B_Y}\right)\\ &\quad + \mu_{B_T}^\top H_T^\top(0)\left(H_X(0)\mu_{B_X} + \mu_{B_Y}\right)\\ &\quad + \mu_{B_X}^\top H_X^\top(0)\left(H_T(0)\mu_{B_T} + \mu_{B_Y}\right)\\ &\quad + \mu_{B_Y}^\top\left(H_T(0)\mu_{B_T} + H_X(0)\mu_{B_X}\right)\end{aligned} \quad (30)$$

Proof: The transfer functions H, H_T and H_X share the same observability gramian W_o, and the proposition 1 links the moments $\sigma_{B'_T}$, $\sigma_{B'_X}$, $\sigma_{B'_Y}$, $\mu_{B'_T}$, $\mu_{B'_X}$ and $\mu_{B'_Y}$ to the noises' moments Ψ_{B_T}, Ψ_{B_X}, Ψ_{B_Y}, μ_{B_T}, μ_{B_X} and μ_{B_Y}. ■

Remark 2 Equation (30) is a good illustration of the relationship between the works done in the *hardware/software* community and the one done in the *control* community: the first ones are based on the accurate evaluation of the noises for particular H/S fixed-point implementation on various targets (DSP, FPGA) whereas the second ones are based on the search of *good* realizations with particular well-conditioned structures. In the first case, only the classical direct form is studied, whereas the real HW/SW impact is neglected in the second case.
The moments Ψ_{B_T}, Ψ_{B_X}, Ψ_{B_Y}, μ_{B_T}, μ_{B_X} and μ_{B_Y} only depend on the H/W implementation, whereas the other terms (W_o, $J^{-\top}\left(L^\top L + K^\top W_o K\right)J^{-1}$, $H_T(0)$ and $H_X(0)$) only depend on the algorithm used.

3.3 Roundoff Noise Gain

The *Roundoff Noise Gain* is the output noise power in a specific computational scheme: the noises are supposed to appear only after each multiplication (Roundoff After Multiplication scheme) and are modeled by centered white noise statically independent. Each noise hase the same power σ_0^2 (determined by the wordlength chosen for all the variables and coefficients).

Definition 2 *The* Roundoff Noise Gain *is defined by*

$$G \triangleq \frac{P}{\sigma_0^2} \tag{31}$$

This measure was studied by [13, 4, 1] and has been established for state-space realizations and some other particular realizations (the ρDFIIt, see [10, 18]). This particular computational scheme fixes the moments of B_T, B_X and B_Y: they only depend here on the number of non-trivial parameters in the realization.
Let introduce the matrices d_J to d_S. They are diagonal matrices defined by

$$(d_X)_{i,i} \triangleq \text{number of non-trivial parameters in the i}^{\text{th}} \text{ row of } X \tag{32}$$

The trivial parameters considered are 0, 1 and -1 because they do not imply a multiplication.

The first step of the algorithm (1) is

$$J.T(k+1) \leftarrow M.X(k) + N.U(k) \tag{33}$$

and is realized as follows ($1 \leqslant i \leqslant l$):

$$T_i(k+1) \leftarrow \sum_{j=1}^{n} M_{ij}X_j(k) + \sum_{j=1}^{m} N_{ij}U_j(k) - \sum_{j<i} J_{ij}T_j(k+1) \tag{34}$$

Each multiplication by a non-trivial parameter implies a quantization noise. Since they are independent centered white noise, Ψ_{B_T} is given by:

$$\begin{aligned} \Psi_{B_T} &= E\left\{B_T(k)B_T^\top(k)\right\} && (35)\\ &= (d_M + d_N + d_J)\,\sigma_0^2 && (36)\end{aligned}$$

(J is a diagonal matrix with 1 on the diagonal, so the number of non-trivial parameters of each row in the sub-diagonal part of J is equal to the number of non-trivial parameter of each of its row).
For the same reasons,

$$\begin{aligned} \Psi_{B_Y} &= (d_L + d_R + d_S)\,\sigma_0^2 && (37)\\ \Psi_{B_X} &= (d_K + d_P + d_Q)\,\sigma_0^2 && (38)\end{aligned}$$

Proposition 3 *Then, the RNG is given by*

$$\begin{aligned} G &= tr\left((d_M + d_N + d_J)J^{-\top}\left(L^\top L + K^\top W_o K\right)J^{-1}\right)\\ &\quad + tr\left((d_K + d_P + d_Q)W_o\right) + tr\,(d_L + d_R + d_S) \end{aligned} \tag{39}$$

Remark 3 In the state-space case, eq. (39) leads fortunately to the classical result enonced by Mullis & Roberts [13, 1]

$$G = tr\left((d_A + d_B)W_o\right) + tr\,(d_C + d_D) \tag{40}$$

Remark 4 Recently, a new sparse structure have been proposed, the ρDFIIt [18, 10] and the RNG developped for it. It is obvious to describe this structure in the Implicit State-Space Framework and find again, with the general equation (39), the RNG in that case.

4. OPTIMAL DESIGN

Since the Roundoff Noise power depends on the realization chosen to numerically realize the filter, it is of interest to find, among the equivalent realizations set, those with lower roundoff noise.

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. The *Inclusion Principle* introduced by Šiljak and Ikeda [5] in the context of decentralized control, could be extended to the Specialized Implicit Form in order to characterize equivalent classes of realizations [2]. Although this extension gives the formal description of equivalent classes, it is of practical interest to consider only realizations with the same dimensions, where transformation from one realization to another is only a similarity transformation.

Proposition 4 *Consider a realization $\mathscr{R}_0 = (Z_0, l, m, n, p)$. All realizations $\mathscr{R}_1 = (Z_1, l, m, n, p)$ such that*

$$Z_1 = \begin{pmatrix} \mathscr{Y} & & \\ & \mathscr{U}^{-1} & \\ & & I_p \end{pmatrix} Z_0 \begin{pmatrix} \mathscr{W} & & \\ & \mathscr{U} & \\ & & I_m \end{pmatrix} \tag{41}$$

are equivalent (with $\mathscr{U} \in \mathbb{R}^{n\times n}$, $\mathscr{Y} \in \mathbb{R}^{l\times l}$ and $\mathscr{W} \in \mathbb{R}^{l\times l}$ non-singular matrices).

For particular structured realizations, the transformation matrices $\mathscr{U}$, $\mathscr{Y}$ and $\mathscr{W}$ may be linked (for δ-state-space, $\mathscr{Y} = \mathscr{U}^{-1}$ and $\mathscr{W} = \mathscr{U}$, see (8) ; and for classical state-space, $\mathscr{Y} = \mathscr{W} = I_l$).

The optimal design problem consists in finding the best realization, among the equivalent realizations set, according to a FWL criteria, here the Roundoff Noise Gain.
Due to the size of equivalent realizations set, this problem cannot be solved practically: the search is done among equivalent realizations with particular structure (δ-state-space, cascade decomposition, etc.).

Let consider a realization $\mathscr{R}_0 = (J_0, K_0, L_0, M_0, N_0, P_0, Q_0, R_0, S_0)$ and a realization $\mathscr{R}_1$ deduced from eq. (41).
Assuming that the transformation doesn't change the trivial parameters (this is the case when the search is done among equivalent realizations with particular structure), then the moments Ψ_{B_T}, Ψ_{B_X}, Ψ_{B_Y}, μ_{B_T}, μ_{B_X} and μ_{B_Y} are independent of the transformation. It is obvious to remark that the roundoff noise P is now a function of $\mathscr{U}$ and $\mathscr{Y}$, because:

$$\begin{aligned} W_o\big|_{Z_1} &= \mathscr{U}^\top W_o\big|_{Z_0}\,\mathscr{U}\\ \left(J^{-\top}\left(L^\top L + K^\top W_o K\right)J^{-1}\right)\Big|_{Z_1} &= \mathscr{Y}^{-\top}\left(J^{-\top}\left(L^\top L + K^\top W_o K\right)J^{-1}\right)\Big|_{Z_0}\mathscr{Y}^{-1}\\ H_T\big|_{Z_1}^\top(0) &= H_T\big|_{Z_0}^\top(0)\mathscr{Y}^{-1}\\ H_X\big|_{Z_1}(0) &= H_X\big|_{Z_0}(0)\mathscr{U} \end{aligned} \tag{42}$$

5. EXAMPLES

To illustrate the RNG measure and the optimal design problem, we consider the following filter

$$H(z) = \frac{0.01594(z+1)^3}{z^3 - 1.9749z^2 + 1.5562z - 0.4538} \tag{43}$$

It is a low pass filter (see [1, 4]) and has a triple zero at $z = -1$, so the zero positions are very sensitive to the roundoff noise when realize directly.
The following realizations are considered:

Z_1: direct form I with shift-operator,

Z_2: RNG-optimal state-space realization,

Z_3: RNG-optimal implicit state-space realization: we consider all the equivalent realizations described by

$$\begin{cases} EX(k+1) &= AX(k) + BU(k), \\ Y(k) &= CX(k) + DU(k). \end{cases} \tag{44}$$

where E is a lower triangular matrix. This can be described in the implicit state-space framework by

$$Z_0 = \left(\begin{array}{c|c|c} -E & A & B \\ \hline I_n & 0 & 0 \\ \hline 0 & C & D \end{array}\right) \tag{45}$$

and equivalent realizations can be searched with proposition 4, with $\mathscr{W} = \mathscr{U}$,

Z_4: RNG-optimal δ-realization (described by eq. (7)), with $\Delta = 2^{-5}$.

The RNG-optimal realizations Z_2, Z_3 and Z_4 are obtained by solving the optimal design problem for the RNG measure (the RNG depends on $\mathscr{U}$ and $\mathscr{Y}$, with eq. (42)). The Adaptive Simulated Annealing (ASA) method [6] has been chosen.
To satisfy the RNG computational scheme (same fixed-point representation for every states and intermediate variables), a norm dynamic-range scaling constraint (L_2-scaling, see [3, 13, 15]) is added on the optimal problem.

Table 1: RNG measure and computational cost

realization	RNG	Nb. operations
Z_1	$27.53dB$	$6+\ 7\times$
Z_2	$16.40dB$	$12+\ 16\times$
Z_3	$12.05dB$	$15+\ 19\times$
Z_4	$13.35dB$	$15+\ 19\times$

The results given in table 1 are coherent with existing results on RNG. First, it is possible to find the optimal realization among equivalent ones with same structure. Secondly, even if it is not the main goal of this paper, it is possible to compare realizations: those realization with lower RNG requires more operations. Last point: in that particular example, the realization RNG-optimal implicit state-space realization Z_3 presents better RNG than RNG-optimal δ-realization Z_4 for the same number of operations, but the result is case-depedent: it is possible to find examples where Z_4 could be better than Z_3.

6. CONCLUSION

The Implicit State-Space Form provides a general framework for the analysis and design of digital filter implementation with minimum output noise power. The general output noise power measure and the optimal design problem have been exhibited. The Roundoff Noise Gain corresponds to this measure for a particular and quite simple computational scheme. Due to its limitations (the L_2-scaling changes the parameters and can lead to a higher transfer function sensitivity), our present work focuses on output noise power for general computational scheme (DSP, FPGA, etc.) and a global methodology to search the *optimal* realizations with criteria like power consumption, area (FPGA), output noise power, transfer function sensitivity, pole sensitivity, etc.

REFERENCES

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

[2] T. Hilaire, P. Chevrel, and J.F. Whidborne. A unifying framework for finite wordlength realizations. *to be published in IEEE Trans. on Circuits and Systems*, 2007.

[3] T. Hinamoto, O. Omoifo, and W.-S. Lu. L2-sensitivity minimization for mimo linear discrete-time systems subject to l2-scaling constraints. In *Proc. ISCCSP 2006*, 2006.

[4] S.Y. Hwang. Minimum uncorrelated unit noise in state-space digital filtering. *IEEE Trans. on Acoust., Speech, and Signal Processing*, 25(4):273–281, August 1977.

[5] M. Ikeda, D. Šiljak, and D. White. An inclusion principle for dynamic systems. *IEEE Trans. Automatic Control*, 29(3):244–249, March 1984.

[6] L. Ingber. Adaptive simulated annealing (ASA): Lessons learned. *Control and Cybernetics*, 25(1):33–54, 1996.

[7] R. Istepanian and J.F. Whidborne, editors. *Digital Controller implementation and fragility*. Springer, 2001.

[8] S. Kim, KI. Kum, and W. Sung. Fixed-point optimization utility for C and C++ based digital signal processing programs. *IEEE Transactions on Circuits and Systems*, 45(11):1455–1464, November 1998.

[9] G. Li. On the structure of digital controllers with finite word length consideration. In *IEEE Trans. on Autom. Control*, volume 43, pages 689–693, May 1998.

[10] G. Li and Z. Zhao. On the generalized DFIIt structure and its state-space realization in digital filter implementation. *IEEE Trans. on Circuits and Systems*, 51(4):769–778, April 2004.

[11] D. Ménard and O. Sentieys. Automatic evaluation of the accuracy of fixed-point algorithms. In *Proceedings of DATE02 (Design Automation and Test in Europe)*, march 2002.

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

[13] C. Mullis and R. Roberts. Synthesis of minimum roundoff noise fixed point digital filters. In *IEEE Transactions on Circuits and Systems*, volume CAS-23, September 1976.

[14] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. Mc Graw Hill, 1991.

[15] K.K. Parhi. *VLSI Digital Signal Processing Systems: Design and Implementation of Digital Controllers*. Number ISBN 0-471-24186-5. John Wiley & Sons, 1999.

[16] 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*, volume CAS-31, October 1984.

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

[18] Z. Zhao and G. Li. Roundoff noise analysis of two efficient digital filter structures. *IEEE Transactions on Signal Processing*, 54(2):790–795, February 2006.