Roundoff Noise Analysis of Two Efficient Digital Filter Structures

Zixue Zhao, Gang Li, and Jiong Zhou
School of EEE
Nanyang Technological University, Singapore

***Abstract*— This paper deals with the problem of efficient digital filter structures with roundoff noise consideration. Two efficient structures are analyzed for the implementation scheme of *rounding before multiplication*. The first one is the recently proposed direct-form II transposed structure in ρ-operator (ρDFIIt) [6], based on which, a revised ρDFIIt structure, denoted as ρRDFIIt, is obtained. It is shown that an ρRDFIIt structure, having the same implementation complexity as that of the corresponding ρDFIIt, yields a smaller roundoff noise gain than the latter. The roundoff noise gain for an ρRDFIIt structure with error feedback is also derived. The optimal structure problem is formulated and solved for each structure. A numerical example is presented to illustrate that the optimized structures are very competitive for that they can even over-perform the traditional optimal state-space realization in terms of the roundoff noise performance as well as the implementation complexity.**

I. Introduction

Roundoff noise has been considered as one of the most serious issues in digital filter implementation. There are two rounding schemes: *rounding after multiplication* (RAM) and *rounding before multiplication* (RBM). It is well known that roundoff noise gain can be reduced considerably by the appropriate selection of filter structures. For practical considerations, it is desired that the actually implemented filters have a nice performance as well as a simple structure that possesses many trivial parameters[1], which can be implemented exactly and produce no rounding errors.

The optimal state-space realization design [1]-[4] has been known as one of the effective methods to reduce the roundoff noise. For a digital filter of order p, such an optimal realization has $(p+1)^2$ nontrivial parameters, which is obviously not efficient for implementation. It is well known that though having poor numerical properties, the conventional shift operator based direct forms are the simplest structures and among all these forms, the direct-form II transposed (DFIIt) structure has the lowest quantization noise level at output. In [5], the DFIIt structure in delta operator (δDFIIt) was investigated for an arbitrary order IIR filter, but it was found that it has a very good performance just for the type of low-pass narrow band filters. Incited by this fact and based on the concept of polynomial operators [4], a set of special operators was derived, with which a generalized DFIIt structure, denoted as ρDFIIt, was obtained in [6]. This structure, having only $3p+1$ nontrivial parameters plus p free parameters at choice, can be used for any type of digital filters to minimize roundoff noise gain. The performance of this structure was analyzed in [6] for the RAM scheme. Error feedback [8]-[9] is another effective technique for reducing roundoff noise in the RBM scheme. This is achieved by extracting the quantization error and feeding it back through simple circuits. In [9], a number of results in computing the optimal error feedback coefficients were given for a given state-space realization.

In this paper, two efficient structures are analyzed for the RBM scheme. The first one is the ρDFIIt structure proposed in [6], for which the expressions of roundoff noise gain are derived. More importantly, a *revised* ρDFIIt structure, denoted as ρRDFIIt, is obtained. It is shown that this structure, having the same implementation complexity as that of the corresponding ρDFIIt, yields a smaller roundoff noise gain than the latter. The performance of this new structure with error feedback is also analyzed. All these structures can be optimized. An example is given to illustrate the design procedure and to compare the performance of the optimized structures with that of the traditional optimal state-space realization.

Throughout this paper, I_p and $tr(A)$ denote the identity matrix of dimension $p \times p$ and the trace of a square matrix A, respectively. The transpose of a matrix A is indicated by A^T. e_m is the m-th elementary (column) vector, whose elements are all zeros except the m-th which is one.

II. ρDFIIt Structure and Roundoff Noise Gain

Consider the following time-invariant linear digital filter $H(z)$ given by

$$H(z) = \frac{b_0 z^p + b_1 z^{p-1} + \cdots + b_{p-1} z + b_p}{a_0 z^p + a_1 z^{p-1} + \cdots + a_{p-1} z + a_p} \tag{1}$$

where $a_0 = 1$. This filter can be implemented with many different structures. In this section, we will analyze the roundoff noise performance of the ρDFIIt structure recently proposed in [6] for the RBM implementation scheme.

A. The ρDFIIt Structure

The block diagram of an ρDFIIt structure is depicted in Fig. 1 and 2, where

$$\rho_k(z) \triangleq \frac{z-\gamma_k}{\Delta_k} \triangleq \rho_k, \ k = 1, 2, \cdots, p \tag{2}$$

[1]By *trivial parameters* we mean those which are 0 and ±1. Other parameters are, therefore, referred to *nontrivial parameters*.

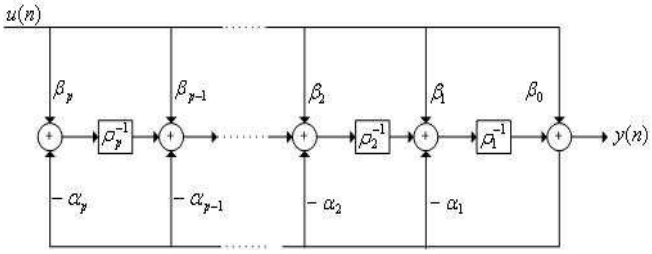

Fig. 1. ρDFIIt structure

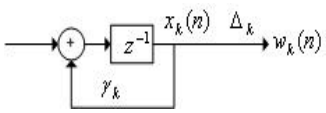

Fig. 2. A realization of operator ρ_k^{-1}

with $\{\gamma_k\}$ and $\{\Delta_k > 0\}$ two sets of free parameters and

$$\begin{cases} \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_p \end{bmatrix}^{\mathcal{T}} = \kappa^{-1}\bar{T}^{-\mathcal{T}} V_a \triangleq V_\alpha \\ \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_p \end{bmatrix}^{\mathcal{T}} = \kappa^{-1}\bar{T}^{-\mathcal{T}} V_b \triangleq V_\beta \end{cases} \tag{3}$$

where

$$V_a \triangleq \begin{bmatrix} a_0 & a_1 & \cdots & a_p \end{bmatrix}^{\mathcal{T}}, \quad V_b \triangleq \begin{bmatrix} b_0 & b_1 & \cdots & b_p \end{bmatrix}^{\mathcal{T}}$$

and $\kappa = \prod_{k=1}^{p} \Delta_k$ certifies that $\alpha_0 = 1$, $\bar{T} \in R^{(p+1)\times(p+1)}$ is an upper triangular matrix whose mth row is determined by the coefficients of the polynomial $\prod_{k=m}^{p} \rho_k$ for $m = 1, 2, \cdots, p$ and $\bar{T}(p+1, p+1) = 1$.

Choosing $\{x_k(n)\}$ indicated in Fig. 2 as the state variables, one can obtain the equivalent state-space realization $(A_\rho, B_\rho, C_\rho, \beta_0)$ as

$$\begin{cases} x(n+1) &= A_\rho x(n) + B_\rho u(n) \\ y(n) &= C_\rho x(n) + \beta_0 u(n) \end{cases} \tag{4}$$

with

$$A_\rho = D_\gamma + A_o D_\Delta, \quad B_\rho = \bar{\beta} - \beta_0 \bar{\alpha}, \quad C_\rho = e_1^{\mathcal{T}} D_\Delta \tag{5}$$

where $D_\gamma \triangleq diag(\gamma_1, \gamma_2, \cdots, \gamma_p)$ and $D_\Delta \triangleq diag(\Delta_1, \Delta_2, \cdots, \Delta_p)$ are two diagonal matrices, $A_o \in R^{p\times p}$ is a zero matrix except $A_o(k,1) = -\alpha_k$, $\forall k$ and $A_o(k, k+1) = 1$ for $k = 1, 2, \cdots, p-1$, and

$$\bar{\alpha} \triangleq \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_p \end{bmatrix}^{\mathcal{T}}, \quad \bar{\beta} \triangleq \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_p \end{bmatrix}^{\mathcal{T}}.$$

The transfer function is then

$$H(z) = \beta_0 + C_\rho (zI_p - A_\rho)^{-1} B_\rho. \tag{6}$$

It is well known [1], [2] that any structure has to be l_2-scaled in order to sustain all the state variables within a specified dynamical range. This can be achieved by the following choice of parameters $\{\Delta_k\}$ for a given set $\{\gamma_k\}$

$$\Delta_1 = \sqrt{\bar{W}_c^\rho(1,1)}, \quad \Delta_k = \sqrt{\frac{\bar{W}_c^\rho(k,k)}{\bar{W}_c^\rho(k-1,k-1)}}, \ k = 2, 3, \cdots, p \tag{7}$$

where $\bar{W}_c^\rho$ is the controllability gramian of $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$ corresponding to $\Delta_k = 1, \forall k$. For the detailed discussion on the ρDFIIt structure, please refer to [6].

One can note that under the l_2-scaling, the ρDFIIt structure is uniquely determined by the parameters $\{\gamma_k\}$, which, generally speaking, can be chosen arbitrarily. For a fixed-point implementation of B_c bits, these free parameters $\{\gamma_k\}$ can be chosen from a discrete space:

$$S_\gamma \triangleq \{-1, \ 1\} \cup \{\pm \sum_{m=1}^{B_\gamma} b_m 2^{-m}, \ b_m = 0, 1, \ \forall m\} \tag{8}$$

where B_γ is an integer, satisfying $B_\gamma << B_c$. Under such a constraint, the free parameters $\{\gamma_k\}$ produce either no roundoff noise at all or a much smaller one than that caused by the non-free parameters $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\Delta_k\}$. Consequently, the roundoff noise due to the multiplication with γ_k can be neglected, see [9], [10].

B. Roundoff Noise Gain for RBM Scheme

Looking at Fig. 1 - 2, the signals to be rounded under the RBM scheme are $y(n)$ and $\{x_k(n)\}$, which have to be multiplied with $\{\alpha_k\}$ and $\{\Delta_k\}$, respectively. With referring to the roundoff noise analysis in [6], it can be shown that the roundoff noise gain due to the rounding of $y(n)$ is given by

$$G_y = \bar{\alpha}^{\mathcal{T}} W_o^\rho \bar{\alpha} \tag{9}$$

where W_o^ρ is the observability gramian of the realization $(A_\rho, B_\rho, C_\rho, \beta_0)$. Similarly, the roundoff noise gain due to the rounding of state variables $\{x_k(n)\}$ is given by

$$G_{x_k} = \begin{cases} \Delta_1^2 (1 + \bar{\alpha}^{\mathcal{T}} W_o^\rho \bar{\alpha}), & k = 1 \\ \Delta_k^2 e_{k-1}^{\mathcal{T}} W_o^\rho e_{k-1}, & k = 2, 3, \cdots, p \end{cases} \tag{10}$$

Hence, the total roundoff noise gain of the ρDFIIt structure in terms of RBM scheme, defined as $G_\rho \triangleq G_y + \sum_{k=1}^{p} G_{x_k}$, is

$$G_\rho = \bar{\alpha}^T W_o^\rho \bar{\alpha} + \Delta_1^2 (1 + \bar{\alpha}^T W_o^\rho \bar{\alpha}) + tr(W_o^\rho Q_\Delta) \tag{11}$$

where $Q_\Delta = diag(\Delta_2^2, \Delta_3^2, \cdots, \Delta_p^2, 0)$.

III. Revised ρDFIIt Structures with/without Error Feedback

It has been found that the contribution of G_{x_1} to G_ρ in (11) can be very significant. Based on this observation, we propose a new structure depicted in Fig. 3, where

$$\beta_{r0} = \Delta_1^{-1} \beta_0, \quad \alpha_{rk} = \Delta_1 \alpha_k, \ \forall k$$

and $\rho_k^{-1}, \ k = 2, 3, \cdots, p$, is still implemented with Fig. 2. It's found that this structure, denoted as ρRDFIIt for convenience, can be viewed as a cascade implementation of the transfer function $H(z) = \Delta_1 \bar{H}(z)$ with $\bar{H}(z)$ realized by an ρDFIIt in which the operator ρ_1^{-1} has its Δ-parameter equal to one. The corresponding state-space realization of the transfer function

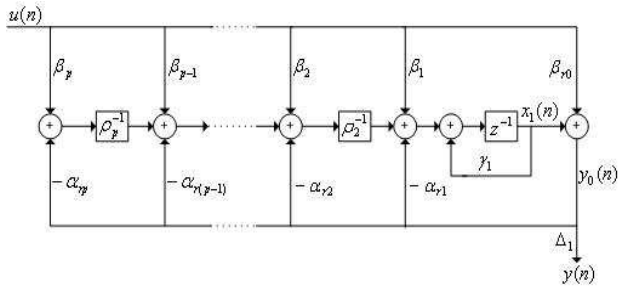

Fig. 3. ρRDFIIt structure

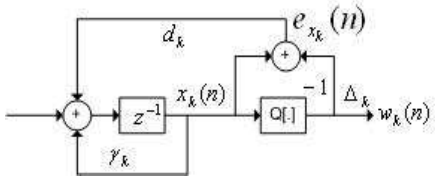

Fig. 4. A realization of operator ρ_k^{-1} with error feedback

$\bar{H}(z)$ implemented with Fig. 3, denoted as $(A_r, B_r, C_r, \beta_{r0})$, is clearly given by

$$A_r = A_\rho, \quad B_r = B_\rho, \quad C_r = \Delta_1^{-1} C_\rho. \tag{12}$$

Denote $y_0(n)$ as the output of $\bar{H}(z)$ excited with $u(n)$, hence $y(n) = \Delta_1 y_0(n)$. It can be observed that the signals to be rounded in ρRDFIIt are $y_0(n)$ and $\{x_k(n),\ k = 2, 3, \cdots, p\}$. It follows from the roundoff noise analysis in Section II-B and noting the fact that there is no roundoff noise due to the state variable $x_1(n)$ in Fig. 3 that the total roundoff noise gain of the ρRDFIIt structure is

$$G_r = \Delta_1^2 (1 + \bar{\alpha}^{\mathcal{T}} W_o^\rho \bar{\alpha}) + tr(W_o^\rho Q_\Delta) \tag{13}$$

where Q_Δ is defined before.

Comparing (11) with (13), it is clear that $G_\rho = G_r + \bar{\alpha}^{\mathcal{T}} W_o^\rho \bar{\alpha} > G_r$. The difference, $\bar{\alpha}^{\mathcal{T}} W_o^\rho \bar{\alpha}$, depending on the filters, can be very large. This is the motivation to propose the ρRDFIIt structure indeed.

The performance of this ρRDFIIt structure can be further improved with using error feedback. In what follows, we apply error feedback technique to the ρRDFIIt structure, and the resultant realization, denoted as ρREF, is depicted in Fig. 3 - 4, where the state residue $e_{x_k}(n)$ is fed back through a constant coefficient d_k into the filter. The error feedback coefficients $\{d_k\}$ are chosen such that the roundoff noise gain is minimized. In order to avoid new roundoff noise sources caused by these coefficients, $\{d_k\}$ are usually constrained to a B_d-bit format number with $B_d << B_c$ [9], that is $d_k \in S_d$, where

$$S_d \stackrel{\triangle}{=} \{-1,\ 1\} \cup \{\pm \sum_{m=1}^{B_d} b_m 2^{-m},\ b_m = 0, 1,\ \forall m\} \tag{14}$$

so that the roundoff noise due to the multiplication with d_k can be neglected.

With some manipulations, it can be shown that the total roundoff noise gain for ρREF is

$$G_e = \Delta_1^2 (1 + \bar{\alpha}^{\mathcal{T}} W_o^\rho \bar{\alpha}) + \sum_{k=2}^{p} V_k^{\mathcal{T}} W_o^\rho V_k \tag{15}$$

where $V_k \stackrel{\triangle}{=} d_k e_k - \Delta_k e_{k-1}$. Clearly, G_e is equal to G_r when $d_k = 0,\ \forall k$.

IV. Structure Optimization

In the previous sections, we have derived the expressions of roundoff noise gain for the ρDFIIt structure and its revised version ρRDFIIt with/without error feedback in terms of RBM scheme.

Denote

$$\bar{\gamma} \stackrel{\triangle}{=} \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_p \end{bmatrix}^{\mathcal{T}}, \quad \bar{d} \stackrel{\triangle}{=} \begin{bmatrix} d_2 & d_2 & \cdots & d_p \end{bmatrix}^{\mathcal{T}}.$$

Let $\bar{S}_\gamma \in R^{1\times p}$ and $\bar{S}_d \in R^{1\times(p-1)}$ as the spaces from which $\bar{\gamma}$ and $\bar{d}$ take values, respectively. It follows from (8) and (14) that

$$\bar{S}_\gamma = \{\bar{\gamma} : \gamma_k \in S_\gamma,\ \forall k\}, \quad \bar{S}_d = \{\bar{d} : d_k \in S_d,\ \forall k\}. \tag{16}$$

A. Optimized ρDFIIt and ρRDFIIt (without error feedback) structures

Since the roundoff noise performance depends on the choice of the free parameters $\{\gamma_k\}$, one can form the optimal structure problems below:

$$\min_{\bar{\gamma}\in\bar{S}_\gamma} G_\rho \Longrightarrow \bar{\gamma}(G_\rho^{opt}), \quad \min_{\bar{\gamma}\in\bar{S}_\gamma} G_r \Longrightarrow \bar{\gamma}(G_r^{opt}). \tag{17}$$

This minimization problem, can be solved using exhaustive searching because $\bar{S}_\gamma$ given by (16) contains $(2^{B_\gamma+1}+1)^p$ elements. This may need a long time to run the program when p is large, though for the off-line design, it is not a big problem. We have developed a *genetic algorithm* to find the optimal structures. A numerical example shows that this program yields a structure which has almost the same roundoff noise gain as that of the structure obtained with exhaustive searching but much more efficient than the latter.

B. Optimized ρREF structure

Since roundoff noise gain of the ρREF is a function of both $\bar{\gamma}$ and $\bar{d}$, the corresponding optimal structure problem is formulated as

$$\min_{\bar{\gamma}\in\bar{S}_\gamma,\ \bar{d}\in\bar{S}_d} G_e. \tag{18}$$

In [9], the roundoff noise gain was minimized with considering the error feedback coefficients as continuous variables and the optimal error feedback coefficients of B_d-bit format were then obtained by truncating the resultant optimal continuous

coefficients into B_d-bit representations. Following the same approach, (18) can be solved as follows.

First of all, we note that for a given $\bar{\gamma}$, the roundoff noise gain G_e can be minimized with respect to $\bar{d}$ analytically. In fact, taking the first derivative of (15) with respect to d_k and letting it be zero, the optimal continuous d_k, denoted as d_k^{opt}, can be obtained as

$$d_k^{opt} = \Delta_k \frac{e_k^{\mathcal{T}} W_o^{\rho} e_{k-1}}{e_k^{\mathcal{T}} W_o^{\rho} e_k}, \ k = 2, 3, \cdots, p. \tag{19}$$

With such a choice of $\{d_k\}$, the corresponding G_e is equal to

$$G_e = G_r - \sum_{k=2}^{p} \Delta_k^2 \frac{(e_k^{\mathcal{T}} W_o^{\rho} e_{k-1})^2}{e_k^{\mathcal{T}} W_o^{\rho} e_k} \stackrel{\triangle}{=} G_{eo} \tag{20}$$

which is obviously smaller than G_r by (13). The price paid is $p-1$ more nontrivial parameters.

The optimal $\bar{\gamma}$ can then be obtained by solving

$$\min_{\bar{\gamma} \in \bar{S}_\gamma} G_{eo} \Longrightarrow \bar{\gamma}(G_e^{opt}) \tag{21}$$

which, by nature, is exactly the same problem as (17) and can be attacked using the same genetic algorithm. With the optimal $\bar{\gamma}$, one can compute the corresponding W_o^{ρ} and hence d_k^{opt} by (19). The optimal error feedback coefficients of B_d-bit format are obtained by truncating d_k^{opt} into B_d bits.

V. Numerical Example

In this section, we will present a design example to illustrate the performance of the optimized ρDFIIt and ρRDFIIt with/without error feedback structures. In what follows, $B_\gamma = 4$ is assumed. The optimal sets of $\bar{\gamma}$ for (17) and (21) are found with using our genetic algorithm. For convenience, denote the optimal fully parameterized state-space realization [1], [2] as R_{opt}.

Example: This is a sixth order low-pass Butterworth digital filter, generated with MATLAB command $[V_b, V_a] = butter(6, 0.1)$. For this example, computation shows that

$$\begin{aligned}
\bar{\gamma}(G_{\rho}^{opt}) &= [0.8125\ 0.8125\ 0.8125\ 0.8125\ 0.75\ 0.75] \\
\bar{\gamma}(G_r^{opt}) &= [0.8125\ \ 0.8125\ \ 0.8125\ \ 0.75\ \ 0.75\ \ 0.75] \\
\bar{\gamma}(G_e^{opt}) &= [0.875\ \ 0.8125\ \ 0.75\ \ 0.75\ \ 0.75\ \ 0.75]
\end{aligned}$$

and the optimal continuous error feedback coefficients $\bar{d}$ obtained through $\bar{\gamma}(G_e^{opt})$ is

$$\bar{d}^{opt} = [0.132295\ \ 0.210031\ \ 0.218724\ \ 0.225550\ \ 0.231134].$$

Table I presents the statistics of the roundoff noise gain G and the number of nontrivial parameters N_p for the R_{opt}, ρDFIIt, ρRDFIIt and ρREF structures. One can see that R_{opt} yields a roundoff noise gain of 0.7121 and needs 49 nontrivial parameters. While the ρDFIIt structure yields a roundoff noise gain of 0.5456 with only 25 nontrivial parameters, and the ρRDFIIt structure yields a roundoff noise gain of 0.2290 which is about 42% of the ρDFIIt with the same implementation complexity. With truncating $\bar{d}^{opt}$ into 4-bits, the ρREF yields a roundoff noise gain of 0.2008, even smaller than that of the ρRDFIIt, but needs 5 more nontrivial parameters as the compensation.

TABLE I
Example I

Realization	R_{opt}	ρDFIIt	ρRDFIIt	ρREF($B_d = 4$)
G	0.7121	0.5456	0.2290	0.2008
N_p	49	25	25	30

VI. Conclusion

This paper has investigated two efficient structures with minimizing roundoff noise gain in terms of RBM implementation scheme. The first one is the ρDFIIt structure, and based on which, the second one is a revised version of the ρDFIIt structure with and without error feedback. It is shown that these structures can be optimized in terms of minimizing roundoff noise gain. An example has been given, which shows that the optimized structures can over-perform the traditional optimal state-space realization in terms of roundoff noise gain as well as the structure complexity.

References

[1] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551-562, Sept. 1976.

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

[3] D. V. B. Rao, "Analysis of coefficient quantization errors in state-space digital filters," *IEEE on Trans. on Acoust., Speech, and Signal Processing*, vol. ASSP-34, pp. 131-139, Feb. 1986.

[4] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects*, Springer Verlag London, Communication and Control Engineering Series, 1993.

[5] N. Wong and T. S. Ng, "A generalized direct-form delta operator-based IIR filter with minimum noise gain and sensitivity," *IEEE Trans. on Circuits Syst. II*, vol. 48, pp. 425-431, Apr. 2001.

[6] G. Li and Z. X. Zhao, "On the generalized DFIIt structure and its state-space realization in digital filter implementation," *IEEE Trans. Circuits Sysm. I*, vol. 51, pp. 769-778, Apr. 2004.

[7] T. L. Chang and S. A. White, "An error cancellation digital filter structure and its distributed-arithmetic implementation," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp.339-342, Apr. 1981.

[8] T. Hinamoto, S. Karino, N. Kuroda, and T. Kuma, "Error spectrum shaping in two-dimentional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. 46, pp. 1203-1215, Oct. 1999.

[9] T. Hinamoto, H. Ohnishi and W. S. Lu, "Roundoff noise minimization of state-space digital filters using separate and joint error feedback/coordinate transformation optimization," *IEEE Trans. Circuits Sysm. I*, vol.50, pp. 23-33, Jan. 2003.

[10] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 41, pp. 629-637, Feb. 1993.