INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2025, VOL. 71, NO. 3, PP. 1–9
Manuscript received July 3, 2025; revised July 2025. doi: 10.24425/ijet.2025.153611

Error analysis of digital filters using fixed point arithmetic

Robert Wirski, and Paweł Poczekajło

***Abstract*—Based on an electrocardiogram filter, measurement methods of magnitude and phase responses, quantization and overflow errors, as well as limit circles in digital filters for fixed number representation are presented. A computer library for SCILAB has been created to simplify simulations. Direct form II, cascade, and rotation structures performance has been compared. It has been shown that there is no the best structure but the rotation one is superior to classical structures except for quantization errors. However, due to its low overflow errors, quantization noise can be further minimised by relocation of integer bits to fractional part of fixed point number representation.**

***Keywords*—digital signal processing; finite word length effects; simulation; quantization; overflow**

I. Introduction

A. Motivation

Between the 1960s and the 1970s, a number of papers has been published presenting problems witnessed in digital filters, which are implemented using finite-length registers. For a survey, see references in Ch. 14 of [1]. It occurred, that processing quality of simple, direct implementations of digital signal processing (DSP) algorithms is less than expected in terms of errors and intrinsic oscillations. Several design approaches considering finite word length effects are used in hardware DSP designs. To minimize quantization errors, the word-length optimization of computational units is usually performed, measured with signal-to-noise ratio and hardware costs [2]–[5]. In [6], a technique is proposed to measure DSP system errors using a single excitation consisting of a sum of sine signals of harmonic frequencies and selected amplitudes and phases. It allows to determine overall coefficient sensitivity, quantization, and overflow errors. The drawback of that technique follows from the specific shape of the multisine input signal which may not fit well to the real ones. In [7], the authors address a problem of last-bit accurate implementations of a direct form I filter structure. In the papers above, no coefficient sensitivity and limit circles are discussed. In [8], the authors designed RLS adoptive controller for which resource utilization of FPGA chip and signal-to-noise performance have been measured. Entire path of a hardware digital filter design is discussed in [10], where the authors optimized word length of several filter structures to reach the desired signal-to-noise ratio. Overflow and limit circles are also signalled but no optimization in this area is presented. In the literature presented above, the authors deal with narrow set of parameters, usually quantization and a cost. However, performed simulations show, that improving one parameter causes deterioration of another. Moreover, the structure optimizations are based mostly on bit length manipulation. A number of papers were published presenting filter designs. However, rarely their performance is measured under real computational structures and number representations [11], [12]. We believe, that if the filter is about to do real processing, one cannot stop a design just on a system function synthesis disregarding real computational structures. It might happen, that a given filter design will not be feasible under a dedicated software or hardware the designer has in mind. Up to the authors knowledge, no approach has been developed to evaluate coefficient quantization sensitivity, quantization, overflow, and limit circles in a whole.

Those effects are of nonlinear nature that are hard to describe analytically. That is why we decided to create a simulation tool for that as a toolbox for SCILAB [13]. We present magnitude coefficient sensitivities, quantization errors, overflow errors, and limit circles for a selected electrocardiogram (ECG) filter implemented using direct form II, cascade and rotation structures. We show, that the latter structure performs better then classical ones even for lower bit length. We also propose our original algorithm to find and classify limit circles for any given structure node, as well as measurement benchmark for them. Finally, we introduce a new parameter provisionally called a decrease coefficient of overflow errors, discussed in Sec. IV-B.

R. Wirski and P. Poczekajło are with Faculty of Electronics and Computer Science, Koszalin University of Technology, Koszalin, Poland (e-mail: robert.wirski, pawel.poczekajlo@tu.koszalin.pl).

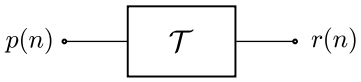


Fig. 1. Digital signal processing system.

B. Background

In this paper we deal with DSP systems as shown in Fig. 1. The system $\mathcal{T}$ is a linear time invariant DSP system which processes input $p(n)$ and produces output $r(n)$ [1]. Both $p(n)$ and $r(n)$ are considered real functions of independent integer

variable n, called discrete functions. Such systems are usually described by a discrete convolution in the following form

$$r(n) = \sum_{k=-\infty}^{\infty} p(k)h(n-k), \tag{1}$$

where $h(n)$ is called an impulse response function. It is the response to the unit step sequence defined as

$$\delta(n) = \begin{cases} 1 \text{ for } n = 0 \\ 0 \text{ for } n \neq 0 \end{cases}. \tag{2}$$

There are two main classes of DSP systems: finite impulse response (FIR) and infinite impulse response (IIR). The latter needs to be implemented using feedback loops. A useful starting point for filter structures development is a recursive system described by a difference equation

$$r(n) = -\sum_{k=1}^{N} a_k r(n-k) + \sum_{k=0}^{M} b_k p(n), \tag{3}$$

where a_k, b_k are real constants for time-invariant systems. We usually deal with a $\mathcal{Z}$-transform of a discrete function $f(n)$, given by

$$F(z) = \mathcal{Z}\{f(n)\} = \sum_{n=-\infty}^{\infty} f(n)z^{-n}, \tag{4}$$

where z is a complex number. Applying (4) to (1), one gets

$$R(z) = H(z)P(z), \tag{5}$$

where $R(z)$, $P(z)$, and $H(z)$ are $\mathcal{Z}$-transforms of $r(n)$, $p(n)$, and $h(n)$, respectively. Applying (4) to (3) and comparing to (5) one obtains

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 + \sum_{k=1}^{N} a_k z^{-k}}, \tag{6}$$

which is called a system function. We define a magnitude response $\left|H(e^{j\omega})\right|$ and a phase response $\arg H\left(e^{j\omega}\right)$, both called frequency responses, where $\omega = [0, \pi]$ covers all digital system's usable frequency range from zero to Nyquist frequency. To measure a quantization impact of c coefficient on a frequency response, one defines magnitude coefficient sensitivity $S_c^{\text{mag}}(\omega)$, and phase coefficient sensitivity $S_c^{\text{phase}}(\omega)$, given by

$$S_c^{\text{mag}}(\omega) = \frac{\partial |H(\omega, c)|}{\partial c} \tag{7a}$$

$$S_c^{\text{phase}}(\omega) = \frac{\partial \arg H(\omega, c)}{\partial c}. \tag{7b}$$

To get overall structure coefficient sensitivities for all parameters c_k $(k = 1, \cdots, K)$, we use statistical and pessimistic approach:

$$S_{\text{stat}}^{\text{mag/phase}}(\omega) = \frac{1}{K}\sum_{k=1}^{K} S_{c_k}^{\text{mag/phase}} \tag{8a}$$

$$S_{\text{pesim}}^{\text{mag/phase}}(\omega) = \frac{1}{K}\sum_{k=1}^{K} \left|S_{c_k}^{\text{mag/phase}}\right|. \tag{8b}$$

One can not build ideal linear time invariant digital systems, given in Fig. 1, but we can come close to them using DSP systems utilising finite word length numbers. They are subject to effects manifested as inaccurate both frequency and time domain responses. When some exact computational result requires more precision than it is available, it must be replaced by another number which is less accurate but fits system's representation. It is called quantization, which is typically categorized to rounding or ceiling. Overflow occurs when a computational result exceeds allowed range of numbers. There are two common overflow characteristics: wrap around and saturation.

Typically, fixed point numbers or floating point numbers are used to represent finite word length numbers. The former are usually marked by Q*i.f*, which describes a *i+f* bit length binary number consisting of i bits for the integer part and a sign, and f bits that are the fraction. The most common method of representing negative numbers is two's complement (U2) used widely by contemporary microprocessors. For example sake lest us consider a computing system based on fixed-point 8-bit number precision which in U2 covers number range [-128,127]. Multiplication of such numbers produces 16-bit results which may fit the range but do not fit system's register length. So, it is required to limit the result's precision to 8 bits which causes quantization. If the multiplication or summation result is beyond the [-128,127] range, overflow effects occurs and the result must be replaced by another number within the range.

A limit circle is another finite word length effect which manifests itself as an intrinsic steady oscillation observed in nodes of a system. It occurs only in feedback structures and does not disappear even when the input vanishes.

C. Orthogonal Filters

Due to poor performance of direct form structures, several techniques have been elaborated to mimic a behaviour of lossless analog filters in digital domain. The most common are wave filters [14] and orthogonal filters [15]. In [9], [10], [16], the authors presented synthesis and realization techniques of one-dimensional (1-D), two-dimensional and three-dimensional orthogonal filters. These are called lossless because they preserve energy of processed signals. The energy of a real vector $x(n)$ is defined as follows

$$\mathcal{E}\{x(n)\} = \sum_{n=-\infty}^{\infty} x^T(n)x(n). \tag{10}$$

For 1-D case, the filter design starts with the embedding, given by

$$H(z) = \begin{bmatrix} h(z) \\ g(z) \end{bmatrix}, \tag{11}$$

such that $h(z)h(1/z) + g(z)g(1/z) = 1$, which is required to get an orthogonal filter with usable frequency characteristics. Then, state-space equations are computed, given by

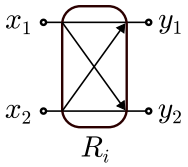
$$\begin{bmatrix} x(n+1) \\ r(n) \end{bmatrix} = \tau \begin{bmatrix} x(n) \\ p(n) \end{bmatrix}, \tag{12}$$

$$A = \begin{bmatrix} 0.9561 & -0.1980 & -0.1326 & -0.0601 & 0.0514 & -0.0095 & 0.0804 \\ 0.1687 & 0.9416 & -0.0486 & -0.0220 & 0.1968 & -0.0530 & -0.0085 \\ 0 & 0 & 0.7846 & -0.2221 & 0.1163 & -0.0147 & 0.3127 \\ 0 & 0 & 0 & 0.9105 & -0.0898 & -0.0475 & 0.1935 \\ 0 & 0 & 0 & 0.2064 & 0.7218 & 0.1683 & 0.0220 \\ 0 & 0 & 0 & 0 & 0 & 0.9503 & -0.1622 \\ 0 & 0 & 0 & 0 & 0 & 0.1791 & 0.8133 \end{bmatrix}$$

$$B = \begin{bmatrix} -0.0303 & 0.1239 \\ 0.1727 & 0.1030 \\ -0.1877 & 0.4339 \\ 0.0043 & 0.3510 \\ -0.5445 & -0.3334 \\ 0.1548 & 0.2160 \\ 0.3505 & -0.4284 \end{bmatrix} \tag{9}$$

$$C = \begin{bmatrix} -0.2396 & -0.1274 & -0.5633 & -0.2554 & 0.3435 & -0.0752 & 0.3147 \\ 0 & -0.2406 & 0.2172 & 0.0985 & 0.5460 & -0.1598 & -0.2662 \end{bmatrix}$$

$$D = \begin{bmatrix} 0.0008 & 0.5667 \\ 0.7005 & 0.0008 \end{bmatrix}$$



$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = R_i(\alpha_i) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Fig. 2. Givens rotation symbol and its equation.

where

$$\tau = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{13}$$

is an orthogonal matrix given by (9). To obtain a structurally lossless system Givens rotations are used. We will use a graphical symbol presented in Fig. 2 to denote the Givens rotation, where

$$R_i(\alpha_i) = \begin{bmatrix} \cos(\alpha_i) & \sin(\alpha_i) \\ -\sin(\alpha_i) & \cos(\alpha_i) \end{bmatrix}. \tag{14}$$

It is known, that (13) can be decomposed into

$$\tau = \left(\prod_i R_i^T(\phi_i) \right) E, \tag{15}$$

where E is a diagonal matrix containing ± 1 on its main diagonal [17, p. 252].

II. Computing System Parameters

To perform simulations, we have elaborated a result representation, called num4, presented in Tab. I. It consists of four variables: *ideal* (highest accuracy, no quantization or overflow), *quantized* (quantization applied), *overflowed* (overflow applied), *hardware* (real response, both quantization and overflow applied). Using such an approach we can calculate all assumed parameters using one model description, which simplifies programming and is less prone to coding errors. The parameters chosen to be analysed are described in the following subsections.

TABLE I
The num4 structure format

structure members	*ideal*	*quantized*	*overflowed*	*hardware*
is quantized?	✗	✓	✗	✓
is overflowed?	✗	✗	✓	✓

A. Magnitude and phase responses

To get magnitude and phase responses, we apply a simulation of the Kronecker delta (2) to the system. The value of $\delta(0)$ should be scaled to be not too small and not too big, so nonlinear phenomena will not be dominant. Typically, is it chosen to be 10% of a maximum value for a given number representation. However, it might be too conservative. So, to improve accuracy, usually it can be set to the half of the maximum value of the chosen number representation.

B. Coefficient sensitivities to quantization

Suppose we are given a system for which we have isolated c coefficient for sensitivity analysis with the precision Δ. Its sensitivity to quantization is evaluated using the following **Algorithm 1**:

a) Compute impulse responses $h_1(n)$ and $h_2(n)$ for the system with c replaced by $c_1 = c - \frac{\Delta}{2}$ and $c_2 = c + \frac{\Delta}{2}$, respectively.
b) Compute FFT for $h_1(n)$ and $h_2(n)$ obtaining magnitudes $A_1(n)$, $A_2(n)$ and phases $P_1(n)$, $P_2(n)$, respectively.
c) Evaluate coefficient sensitivities, given by

$$S_c^{\text{mag}}(\omega) \approx \frac{A_2(n) - A_1(n)}{\Delta} \tag{16}$$

$$S_c^{\text{phase}}(\omega) \approx \frac{P_2(n) - P_1(n)}{\Delta} \tag{17}$$

which are approximations of (7).

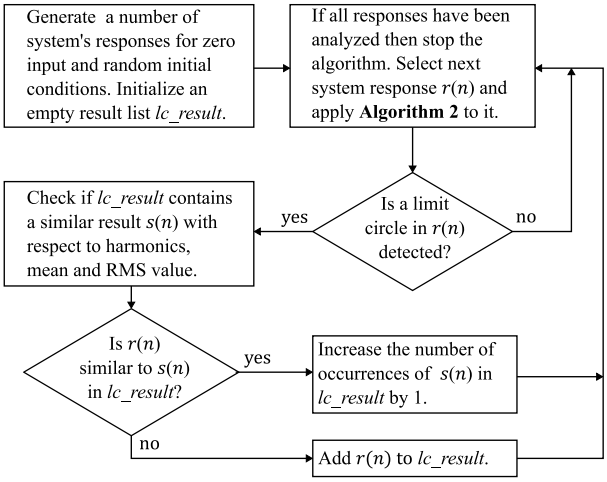


Fig. 3. Classification of limit circles

To obtain overall sensitivities, **Algorithm 1** is used with respect to all system's coefficients. Then, (8) is applied to get statistical and pessimistic sensitivities.

C. Quantization and overflow

The excitation of the system for quantization and overflow analysis is a random input for the zero initial condition. Then, we subtract obtained *num4.quantized*, *num4.overflowed*, and *num4.hardware* values from *num4.ideal*. Obtained errors are noise equivalents observed in analogue processing. To compare quantization and overflow errors for different structures using the same bit format representation, we utilize the standard deviation. For different bit formats, the standard deviation is not a suitable measure, because a signal dynamics is what matters rather then noise levels. In that case, we use the signal-to-noise ratio, defined by

$$\mathrm{SNR} = \frac{\mathrm{FS}}{\sigma_Q}, \tag{18}$$

where σ_Q is the standard deviation of the analysed error, and FS is the maximal positive value for Q*i.f*. In U2 representation, it is given by $\mathrm{FS} = 2^{i-1} - 2^{-f}$. Different types of input probability densities can be considered to mimic real signals to be processed. Additionally, scaling factor in the range $(0, 1]$ can be applied to the input to meet expected signal levels. However to perform a stress test, a uniform pseudorandom signal covering all available number range should be used.

D. Limit circles

We have elaborated fast and simple procedure for evaluation of limit circle parameters, which are: mean $\overline{r_c}$, root mean square (RMS) r_{RMS}, fundamental frequency f_c, and most significant harmonics. It is based on assumption, that the limit circle manifests as a steady-state oscillation. It is presented below as **Algorithm 2**:

a) Suppose we are given l_r samples of the system response output $r(n)$ for a random initial condition X_0 and zero input.

TABLE II
PARAMETERS OF THE ECG FILTER DESIGNED IN [11]

IIR filter parameter	Value
Type	Chebyshev type II
Order	7
Sampling frequency	2 kHz
Bandform	low pass
Lower cutoff frequency	100 Hz
Upper cutoff frequency	200 Hz
Stopband attenuation	60 dB

b) Get $r_c(n)$ as the final $l_c = 2^M$ (M natural) samples of $r(n)$ such that l_c is roughly half of l_r. It can be computed using:

$$l_c = 2^{\mathrm{floor}(\log_2(l_r))-1}. \tag{19}$$

c) Calculate a root mean square of $r_c(n)$:

$$r_{\mathrm{RMS}} = \sqrt{\frac{1}{l_c}\sum_{n=1}^{l_c} r_c^2(n)} \tag{20}$$

d) Calculate a mean value $\overline{r_c}$ for $r_c(n)$, and subtract it from $r_c(n)$.

e) If $r_c(n)$ is zero for some chosen precision, then there is no limit circle in $r(n)$. Otherwise continue the algorithm.

f) Obtain fast Fourier transform (FFT) of $r_c(n)$ and choose one side of its spectrum, indicated as a_j.

g) Find oscillation harmonics frequencies f_i and complex Fourier coefficients a_j for which $|a_j| > k \cdot \max|a_j|$, for some $k \in (0, 1)$.

h) Find limit circle fundamental frequency $f_o = \min(f_i)$.

The limit circle measurements are repeated and obtained results are compared. Two limit circles $r(n)$ and $s(n)$ are considered similar if both are described by the same set of frequencies f_r, f_s, mean values $\overline{r}$, $\overline{s}$, and root mean squares r_{RMS}, s_{RMS}, which satisfy:

$$\frac{|s_{\mathrm{RMS}} - r_{\mathrm{RMS}}|}{s_{\mathrm{RMS}}} < k_{\mathrm{RMS}} \text{ and } |\overline{r} - \overline{s}| < k_{\mathrm{mean}}, \tag{21}$$

for some precision coefficients k_{RMS} and k_{mean}. The entire technique is presented in Fig. 3. Levels of limit circles are presented in dB referenced to maximum amplitude FS as

$$r_{\mathrm{dBFS}} = 20\log_{10}\frac{r_{\mathrm{RMS}}}{\mathrm{FS}}. \tag{22}$$

III. ECG FILTER DESIGN

To illustrate developed measurement techniques we have chosen the IIR filter meant to de-noise ECG signals, proposed in [11] that satisfies a set of requirements gathered in Tab. II. We present three structures: a direct from II, a cascade structure, and a rotation one. For brevity, we will call them DFII, CAS, and ROT, respectively. To get the system function of the filter, a standard filter design command found in Scilab [13] has been used, namely **iir**. Obtained coefficients of the system function (6) are presented in Tab. III. The DFII structure has been presented in Fig. 6. Using Scilab's **factor** command, the CAS structure has been obtained such that each section has the same maximum value of the magnitude

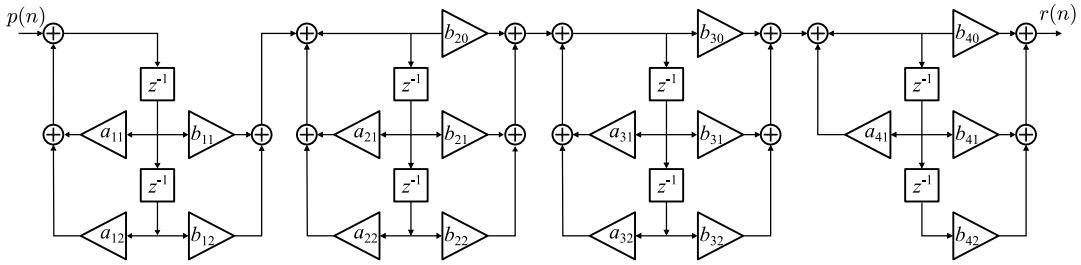


Fig. 4. The CAS structure.

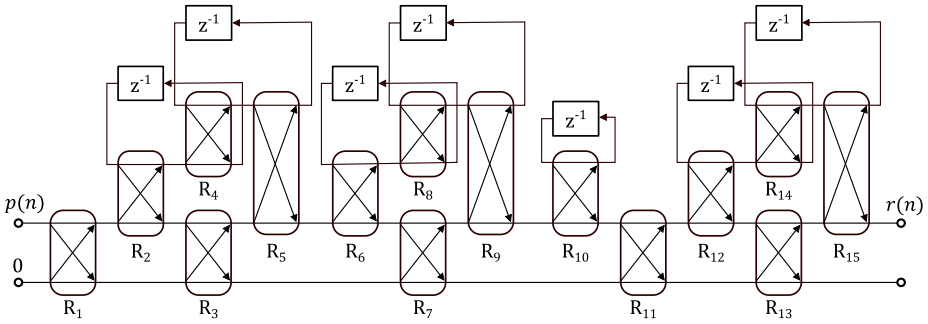


Fig. 5. The ROT structure.

TABLE III
THE DFII STRUCTURE COEFFICIENTS

Coefficent	Value
b_0	0.000845864
b_1	-0.00361109
b_2	0.00587547
b_3	-0.00310379
b_4	-0.00310379
b_5	0.00587547
b_6	-0.00361109
b_7	0.000845864
a_1	-6.07821
a_2	15.8879
a_3	-23.1455
a_4	20.2912
a_5	-10.7032
a_6	3.14476
a_7	-0.396975

TABLE IV
THE CAS STRUCTURE COEFFICIENTS

Coefficent	Value
b_{11}	0.0121714
b_{12}	0.0121714
a_{11}	-1.89773
a_{12}	0.933694
b_{20}	0.147268
b_{21}	-0.225271
b_{22}	0.147268
a_{21}	-1.76363
a_{22}	0.801965
b_{30}	0.532855
b_{31}	-0.981686
b_{32}	0.532855
a_{31}	-1.63226
a_{32}	0.67571
b_{40}	0.885611
b_{41}	-1.68013
b_{42}	0.885611
a_{41}	-0.784591

response. Its coefficients are gathered in Tab. IV and the structure is shown in Fig. 4. Then, by the technique in [10], state-space equations (12) has been evaluated where τ has been decomposed into a product of Givens rotations, whose parameters are given in Tab. V. The resulting rotation structure has been presented in Fig. 5 .

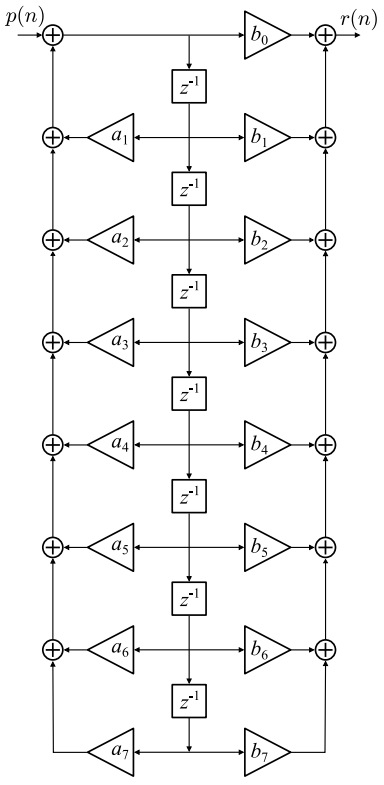


Fig. 6. The DFII structure.

IV. SIMULATIONS

A. Results

All simulations have been performed using U2 fixed-point numbers with rounding in quantization and wrap around overflow. To show magnitude responses for the structures, the number of fractional bits have been chosen to present a deterioration of a given structure. Obtained results have been presented in Fig. 7-10. Then, using (16), magnitude sensitivity has been evaluated for $\Delta = 2^{-17}$ for every coefficient of all structures. The overall statistical sensitivities have been

TABLE V
THE ROT STRUCTURE COEFFICIENTS

Rot. matrix	$sin(\alpha)$	$sin(\alpha)$
R_1	0.773973	-0.633219
R_2	0.562626	0.826712
R_3	-0.991289	-0.131702
R_4	0.179131	0.983825
R_5	-0.258877	0.96591
R_6	0.675188	0.737645
R_7	-0.977747	-0.209786
R_8	0.206435	0.97846
R_9	0.366162	0.930551
R_{10}	0.620014	0.784591
R_{11}	-0.963121	0.26907
R_{12}	0.295617	0.955307
R_{13}	-0.813926	-0.580968
R_{14}	0.168671	0.985672
R_{15}	-0.243085	0.970005

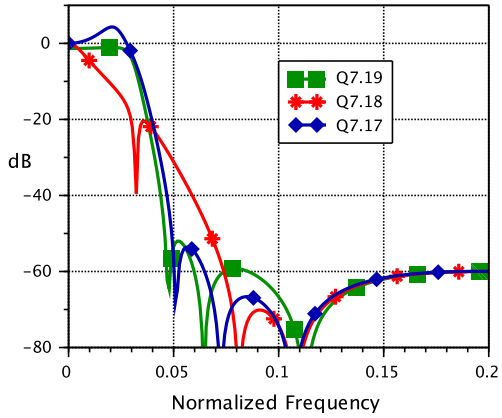


Fig. 7. Magnitude response of the DFII structure (Q7.17-19).

presented in Fig. 11-12. By the technique presented in Sec. II-C, quantization errors have been obtained for Q7.19 to Q7.8 (Tab. VII, Fig. 13), as well as SNR of overflow errors for Q7.17 to Q12.17 (Tab. VIII, Fig. 14). Additionally, overflow errors have been simulated for several input scaling factors (Tab. IX, Fig. 15). Using the technique presented in Sec. II-D, 1000 simulations of limit circles have been performed for each structure for Q18.18. Obtained result has been presented in

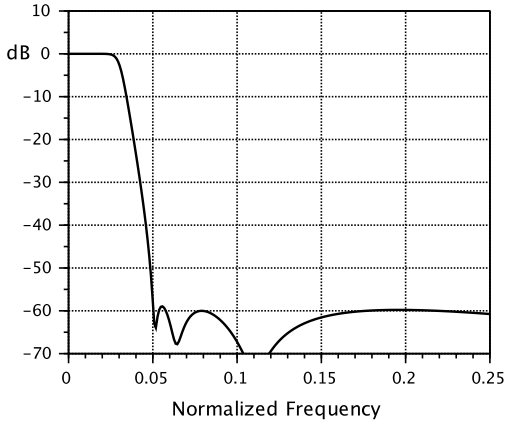


Fig. 8. Magnitude response of the ROT structure (Q7.13).

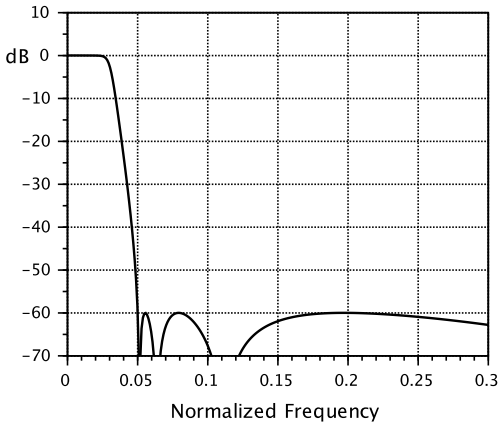


Fig. 9. Magnitude response of the CAS structure (Q7.13).

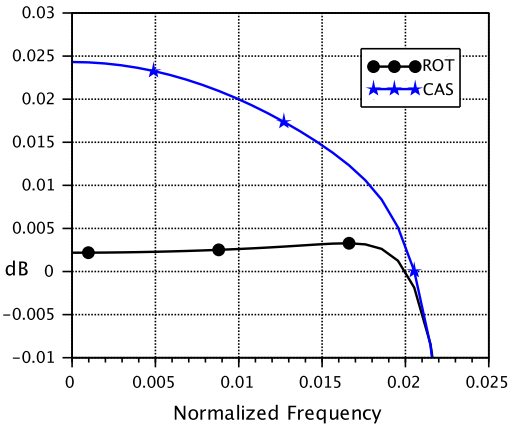


Fig. 10. Comparison of passband magnitude of the CAS and ROT structures (Q7.13).

Tab. VI. The FFT of the limit circles with the highest obtained values have been presented in Fig. 16-18.

B. Conclusions

Obtained overall statistical magnitude sensitivity for DFII (Fig. 11) has enormously bigger values then for ROT and

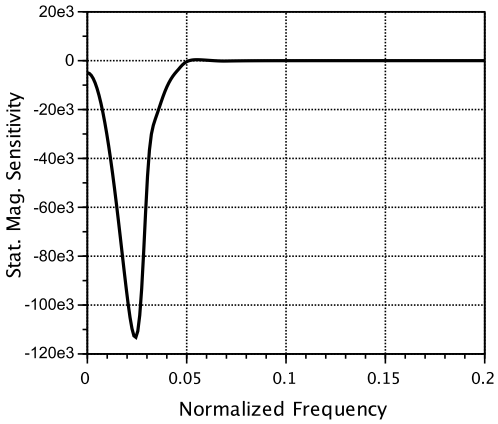


Fig. 11. Overall sensitivity of the DFII structure.

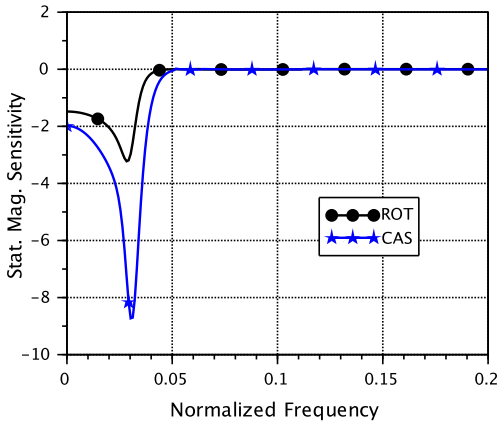


Fig. 12. Overall sensitivity of the ROT and the CAS structures.

TABLE VI
LIMIT CIRCLES OBSERVED IN DFII, CAS, AND ROT STRUCTURES

structure	number of limit circles	Levels
DFII	1000	$[-50.9\,\text{dBFS}, -43.4\,\text{dBFS}]$
ROT	560	$[-211\,\text{dBFS}, -205\,\text{dBFS}]$
CAS	779	$[-65.4\,\text{dBFS}, -1.16\,\text{dBFS}]$

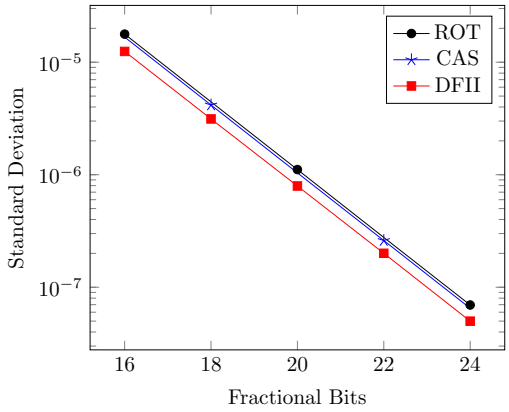


Fig. 13. Standard deviation of quantization errors for the DFII, CAS, and ROT structures

CAS structures (Fig. 12). That is why it is hard to satisfy the frequency constrains using the DFII structure. The magnitude response of the DFII structure (Fig. 7) reveals that even Q7.19 does not fullfill the required stopband attenuation which is -52 dB instead of -60dB. The ROT structure for Q7.13, has

TABLE VII
STANDARD DEVIATION OF QUANTIZATION ERRORS FOR THE DFII, CAS, AND ROT STRUCTURES

Format	DFII std. dev.	ROT std. dev.	CAS std. dev.
Q22.16	$1.2475 \cdot 10^{-5}$	$1.7775 \cdot 10^{-5}$	$1.6719 \cdot 10^{-5}$
Q22.18	$3.1326 \cdot 10^{-6}$	$4.4437 \cdot 10^{-6}$	$4.1646 \cdot 10^{-6}$
Q22.20	$7.9443 \cdot 10^{-7}$	$1.1116 \cdot 10^{-6}$	$1.0388 \cdot 10^{-6}$
Q22.22	$2.0039 \cdot 10^{-7}$	$2.771 \cdot 10^{-7}$	$2.6018 \cdot 10^{-7}$
Q22.24	$4.994 \cdot 10^{-8}$	$6.9534 \cdot 10^{-8}$	$6.5046 \cdot 10^{-8}$

TABLE VIII
SNR OF OVERFLOW ERRORS FOR THE DFII AND ROT AND CAS STRUCTURES

Format	DFII SNR(dB)	ROT SNR(dB)	CAS SNR(dB)
Q1.18	22.1368	12.0218	16.9267
Q2.18	28.0758	∞	22.9449
Q4.18	38.7359	∞	34.8233
Q6.18	43.4945	∞	328.7164
Q8.18	44.1065	∞	∞
Q10.18	44.1329	∞	∞
Q12.18	44.1293	∞	∞
Q14.18	44.137	∞	∞
Q16.18	44.1595	∞	∞
Q18.18	327.8671	∞	∞
Q20.18	339.0925	∞	∞
Q22.18	∞	∞	∞

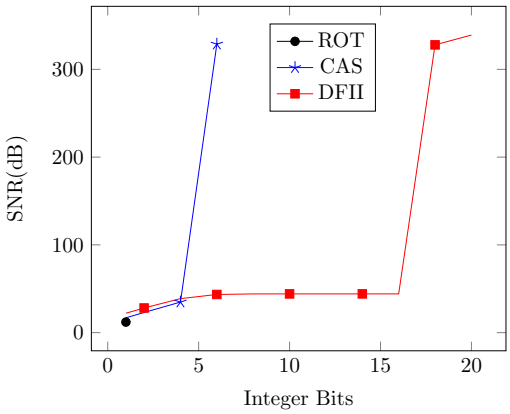


Fig. 14. SNR of overflow errors for the DFII, ROT and CAS structures

the first stopband lobe slighty above -60dB (Fig. 8). For the same Q7.13, the CAS structure is the best when it comes to satisfy the stopband part of the magnitude response (Fig. 9). However, in the passband, the ROT structure is superior to the CAS one (Fig. 10). It is due to very low sensitivity in the passband (which is zero where magnitude equals to 1) of the lossless systems.

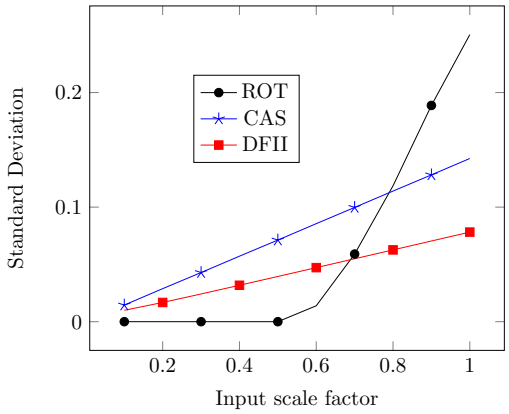


Fig. 15. Standard deviation of overflow errors for the scaled input (Q1.18).

The standard deviation of quantization errors of the ROT and CAS structures for Q22.16 to Q22.24 are, respectively, higher about 40% and 31.5% when compared to the DFII

TABLE IX
STANDARD DEVIATION OF OVERFLOW ERRORS FOR THE SCALED INPUT (Q1.18)

Input scaling factor	DFII	ROT	CAS
0.1	$9.9468 \cdot 10^{-3}$	0	0.0145
0.2	0.0168	0	0.0288
0.3	0.0242	0	0.0429
0.4	0.0318	0	0.0572
0.5	0.0395	0	0.0712
0.6	0.0472	0.0139	0.0856
0.7	0.0549	0.059	0.0997
0.8	0.0627	0.1186	0.114
0.9	0.0704	0.1888	0.1281
1	0.0782	0.2506	0.1425

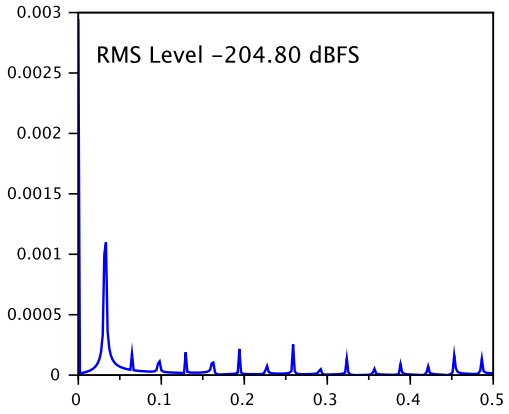


Fig. 16. FFT of the highest limit circle observed in the ROT structure (Q8.18).

structure. It is the only parameter by which the DFII beats the rest of structures. It is caused by the lower number of multiplications in the DFII structure.

The SNR of overflow errors is the best for the ROT structure, for which we did not observe overflow errors as low as for Q2.18. For the CAS and DFII structures, they where absent for Q8.18 and Q22.18, respectively.

We would like to draw reader's attention to the Fig. 15, which shows dependence of standard deviation of overflow errors to input scaling. It uncovers, that structures may differ here in their decreasing rate. Clearly, the structures posses a linear dependence of the overflow errors to input scaling where they are sufficiently greater then zero. We believe, that it is a new parameter, unknown in literature, which can be taken into account when choosing structures. It can be defined as a decrease coefficient of a linear trend where the overflow errors are present. Obtained gradients of overflow errors are 0.66 for ROT; 0.14 for CAS 0.14; 0.08 for DFII. The highest decrease rate of overflow errors has the ROT structure.

To perform limit circles simulations, we chose Q18.18 which is the smallest assuring the acceptable overflow errors amongst all structures (Tab. VIII). The results are presented in Tab. VI. The ROT structure presents the best performance having 560 limit circles with levels lower then -205 dbFS which can be neglected. Surprisingly, the highest levels of limit circles were observed for the CAS stucture. That so popular structure amongst filter designers presented 779 limit circles with levels reaching -1.16 dBFS! This example reveals that a great care must be exercised in application of cascade structures, especially in medical applications when an invoked oscillation may destroy measurement results or a controlled object in automation. The DFII structure presents limit circles for all 1000 attempts with narrow 7.5 dB range of values, below -43 dBFS. The character of oscillations varies from almost clean sine waves (Fig. 17) to a noise-like process (Fig. 18).

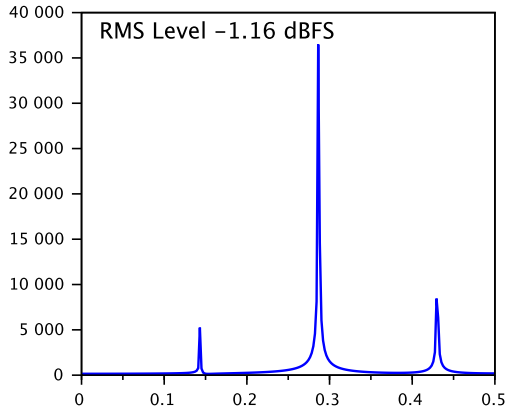


Fig. 17. FFT of the highest limit circle observed in the CAS structure (Q8.18).

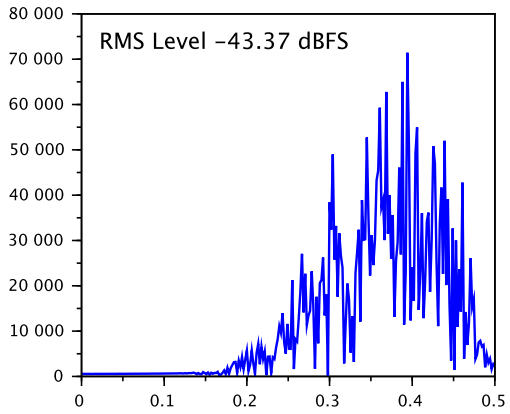


Fig. 18. FFT of the highest limit circle observed in the DFII structure (Q18.18).

Summarizing, we would like to emphasise that the ROT structure won in all presented categories except for the quantization errors. However, due to its low overflow errors, we can greatly decrease the number of integer bits and sightly increase the number of friction bits of its number representation. As a result, the ROT structure can work with lower number of bits having better parameters when compared to other classical structures.

Performed simulations concern the one particular type of the ECG filter. So, it is not clear how the obtained results extend to other types of filters. That is why, for the further work, we plan to perform simulations using wide range of filters and present more general, statistical results.

REFERENCES

[1] A. Antoniou, Digital Signal Processing, McGraw-Hill, 2006.

[2] K.-I. Kum, W. Sung, Combined word-length optimization and high-level synthesis of digital signal processing systems, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 20 (8) (2001) 921–930. https://doi.org/10.1109/43.936374.

[3] G. Constantinides, P. Cheung, W. Luk, Wordlength optimization for linear digital signal processing, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 22 (10) (2003) 1432–1442. https://doi.org/10.1109/TCAD.2003.818119.

[4] G. Li, L. Meng, Z. Xu, J. Hua, A novel digital filter structure with minimum roundoff noise, Digital Signal Processing 20 (4) (2010) 1000–1009. https://doi.org/10.1016/j.dsp.2009.10.018.

[5] V. L. R. da Costa, H. V. Schettino, Ândrei Camponogara, F. P. de Campos, M. V. Ribeiro, Digital filters for clustered-OFDM-based PLC systems: Design and implementation, Digital Signal Processing 70 (2017) 166–177. https://doi.org/10.1016/j.dsp.2017.08.004.

[6] J. Paduart, J. Schoukens, Y. Rolain, Fast measurement of quantization distortions in DSP algorithms, IEEE Transactions on Instrumentation and Measurement 56 (5) (2007) 1917–1923. https://doi.org/10.1109/TIM.2007.903644.

[7] A. Volkova, M. Istoan, F. De Dinechin, T. Hilaire, Towards hardware IIR filters computing just right: Direct form I case study, IEEE Transactions on Computers 68 (4) (2019) 597–608. https://doi.org/10.1109/TC.2018.2879432.

[8] H. H. Thannoon, I. A. Hashim, Efficient fpga implementation of recursive least square adaptive filter using non-restoring division algorithm, International Journal of Electronics and Telecommunications 69 (4) (2024) 175–182. http://dx.doi.org/10.24425/ijet.2023.147705.

[9] P. Poczekajlo, An Overview of the Methods of Synthesis, Realization and Implementation of Orthogonal 3-D Rotation Filters and Possibilities of Further Research and Development, International Journal of Electronics and Telecommunications 67 (2) (2021) 295–300. https://dx.doi.org/10.24425/ijet.2021.135979.

[10] R. Wirski, Synthesis and realization of two-dimensional separable denominator orthogonal systems via decomposition into 1-D systems, IEEE Transactions on Circuits and Systems I: Regular Papers 66 (11) (2019) 4309–4322. https://doi.org/10.1109/TCSI.2019.2927673.

[11] K. D. Shinde, D. Khanapure, N. Shetti, J. Athavani, N. Hattiholi, Denoising of ECG signal using optimized IIR filter architecture—a CSD-based design, in: S. Kalya, M. Kulkarni, S. Bhat (Eds.), Advances in VLSI, Signal Processing, Power Electronics, IoT, Communication and Embedded Systems, Springer Nature Singapore, Singapore, 2024, pp. 143–157.

[12] R. Wu, X. Tang, J. He, Y. Cao, L. Xiao, F. Xiao, The DST-O method for multiband IIR filter, Circuits, Systems, and Signal Processing 42 (2023) 431–448. https://doi.org/10.1007/s00034-022-02129-w.

[13] The SCILAB homepage, http://www.scilab.org.

[14] A. Fettweis, Digital filter structures related to classical filter networks, AEÜ 25 (2) (1971) 79–89.

[15] E. Depretere, P. Dewilde, Orthogonal cascade realization of real multiport digital filters, Int. J. Circuits Theory Appl. 8 (1980) 245–272.

[16] P. Poczekajlo, R. Wirski, Synthesis and realization of 3-D orthogonal FIR filters using pipeline structures, Circuits Syst Signal Process 37 (2018) 1669–1691. https://doi.org/10.1007/s00034-017-0618-2.

[17] G. H. Golub, C. F. Van Loan, Matrix Computations, 4th Edition, The Johns Hopkins Univ. Press, Baltimore, MD, 1996.