Proceedings of the 32nd Chinese Control Conference
July 26-28, 2013, Xi'an, China

# All-Pass Based Efficient and Robust Structures for Finite Precision Implementation of Digital Filters

Tao Hong, Si Tang, Gang Li, Xiongxiong He, and Liping Chang
Zhejiang Provincial Key Laboratory for Signal Processing
College of Information Engineering
Zhejiang University of Technology, Hangzhou, Zhejiang 310023, China
E-mail: ieligang@zjut.edu.cn

**Abstract:** In this paper, a novel structure is derived for digital filters implemented with finite wordlength (FWL). Based on a realization of 1st order complex all-pass systems, we develop a structure for a real 2nd-order filter with complex conjugate poles, in which the transfer functions between the two states and the input signal are in a power-complementary pair. This property ensures that the state dynamics can be controlled easily such that no overflow occurs. In addition, a 2nd-order filter implemented using such a structure needs 7 multiplications and 7 additions for computing each output samples, which is more efficient than the existing optimal FWL structures. A procedure is given to extend such a structure to higher order filters and the corresponding expression for roundoff noise gain is derived. Simulations show that our proposed structure yields the excellent finite word length performance and outperforms the classical optimal structures in terms of reducing quantization errors.

**Key Words:** Digital filter structures, structure robustness, roundoff noise, overflow, all-pass systems

## 1 Introduction

Investigations on the effects of quantization errors which occur in discrete-time systems such as digital filters and controllers have been considered as one of the important research topics in real-time applications. [1] - [8].

Consider an $N$th order infinite impulse response (IIR) digital filter with transfer function $H(z)$:

$$H(z) = \frac{\sum_{k=0}^{N} b_k z^{-k}}{1 + \sum_{k=1}^{N} a_k z^{-k}} \triangleq \frac{B(z)}{A(z)} \tag{1}$$

Such a filter can be realized in many different structures[1] such as the direct-form-based structures. These structures are equivalent in infinite precision but they do have different numerical properties. In fact, although it is canonic in number of the multipliers, the direct-form-based structures are usually very sensitive to the parameter perturbation and yield a large roundoff noise propagation gain. Design of low complexity filters with high robustness against FWL errors has become one of the hot spots. [4] - [6].

The output of a delay elements $z^{-1}$ in a filter structure, usually referred as a *state* signal, has to be stored as it is required by the computations in the next clock period. Since the dynamics of the states are structure dependent, the input signal magnitude must be scaled to be as large as possible so as to maximize word length utilization at all the delay elements and, at the same time, it should not be so large as to cause overflow at any of these elements. It is desired to implement a filter using such a structure in which the dynamical ranges of all the state signals be equal in a certain sense; otherwise, the most significant bits of the delay elements with small signal magnitude will not be effectively utilized. Unequal signal magnitudes at the delay elements will result in a degradation in signal-to-noise ratio performance and this was observed by Lim in [9], where the concept of *peakedness* is proposed to measure this degradation.

[1]Here, a structure means a way in which the output of digital filter can be computed with an input given.

In this paper, we propose a novel filter structure which is based on a parallel realization of the filters. It should be pointed out that the use of parallel structure, in which each sub-system is either a 1st-order sector or 2nd-order sector that is usually implemented in a direct-form-based structure, is not new in reducing the FWL effect of parameter quantization errors. What is new in our proposed structure is that the sub-systems are implemented using a totally new structure derived based on all-pass filters. The nice properties of such a structure include overflow free to any normalized inputs (measured in any norms) and very large structure robustness.

This paper is organized as follows. Some backgrounds on digital filter implementation related issues such as quantization errors, structure scaling and overflow are provided and the problem to be considered is formulated in Section 2. In Section 3, based on all-pass systems a novel structure for 2nd-order filters and then extended to higher order filters. Section 4 is devoted to analyzing the roundoff noise performance of the proposed structure, where the expression of the roundoff gain is derived. A design example is given in 5 to demonstrate the performance of our proposed structure and to compare it with some of existing structures. To end this paper, some concluding remarks are given in Section 6.

## 2 Preliminaries and Problem Formulation

The state-space equations below provide a class of filter structures for implementation:

$$\begin{cases} x[n+1] &= Ax[n] + Bu[n] \\ y[n] &= Cx[n] + du[n] \end{cases} \tag{2}$$

where $u[n], y[n]$ are the input and output of the filter, respectively, $x[n] \in \mathcal{R}^{N\times 1}$ is the state vector, while $A \in \mathcal{R}^{N\times N}, B \in \mathcal{R}^{N\times 1}, C \in \mathcal{R}^{1\times N}, d \in \mathcal{R}$ are all constant, forming a *state-space realization* of the filter and satisfying

$$H(z) = d + C(zI - A)^{-1}B \tag{3}$$

with $I$ denoting the identity matrix of dimension $N$.

It is well known that the realizations are not unique and that the realization set is characterized with

$$A = T^{-1}A_0T,\ B = T^{-1}B_0,\ \ C = C_0T \tag{4}$$

where $(A_0, B_0, C_0, d)$ satisfies (3) and $T$ is any non-singular (real) matrix of dimension $N$.

Suppose that each state is stored in $B_s$-bit format and that the realization is fully parametrized with non-trivial coefficients[2]. Therefore, each multiplication in (2) should be rounded into a $B_s$-bit format in its fractional part. Under the assumption that each roundoff error is a statically independent white noise of variance $\sigma_0^2$, the roundoff noise gain of the structure is defined as

$$G \triangleq \frac{E[|\Delta y[n]|^2]}{\sigma_0^2} \tag{5}$$

where $E[.]$ is the statistical average and $\Delta y[n]$ is the deviation of the output.

The classical minimum roundoff noise realizations [2] under the $l_2$ scaling to be specified below later have a minimum noise gain given by

$$G = [1 + \frac{1}{N}(\sum_{k=1}^{N}\sigma_k)^2](N+1) \tag{6}$$

with $\sigma_k \triangleq \lambda_k(W_cW_o),\ \forall k$ the singular values of the filter, where $\lambda_k(M)$ denotes the $k$th eigenvalue of a matrix $M$, while $W_c, W_o$ are the controllability and observability Gramians of the realization $(A, B, C, d)$, satisfying

$$\left\{\begin{array}{lcl} W_c &=& AW_cA^{\mathcal{T}} + BB^{\mathcal{T}} \\ W_o &=& A^{\mathcal{T}}W_oA + C^{\mathcal{T}}C \end{array}\right. \tag{7}$$

These structures can reduce the roundoff noise significantly but generally require $(N+1)^2$ multiplications and $N(N+1)$ additions for computing one output sample, leading to a very complicated implementation.

### 2.1 Structures for 2nd order filter implementation

One way to reduce the implementation complexity is to realize $H(z)$ in a cascade- or a parallel-form of a number of 1st-order and 2nd-order sub-systems with each implemented with a robust structure against the FWL such as the normal structures, denoted as $S_{NF}$. See [2].

Recently, a low roundoff noise structure, denoted as $S_{WDF}$, was proposed in [6] for 2nd-order filters. See Fig. 1

For $H(z) = \frac{b_0+b_1z^{-1}+b_2z^{-2}}{1+a_1z^{-1}+a_2z^{-2}}$, the structure parameters are given

$$\left\{\begin{array}{lcl} \mu_1 &=& \sqrt{1-a_2}, \mu_2 = \sqrt{\frac{1-a_1+a_2}{2}}, \mu_3 = \sqrt{\frac{1+a_1+a_2}{2}} \\ \mu_6 &=& \frac{b_1+b_2-(a_1+a_2)b_0}{\sqrt{2+2a_1+2a_2}}, \mu_7 = \frac{b_2-b_1+(a_1-a_2)b_0}{\sqrt{2-2a_1+2a_2}} \\ \mu_4 &=& -\mu_2, \mu_5 = -\mu_3, \mu_8 = \frac{1}{\sqrt{1-a_2}}, \mu_9 = b_0 \end{array}\right.$$

### 2.2 Overflow and scaling schemes

Let $x_m[n]$ be the output of the $m$th delay element $z^{-1}$ of a given structure and $T_m(z)$ be the $z$-transform transfer function from the input signal $u[n]$ to this state variable.

[2]Here, by nontrivial parameters we mean those that do not belong to the $\{-1, 0, 1\}$, while the parameters in this set are referred to trivial, which cause no FWL errors at all.

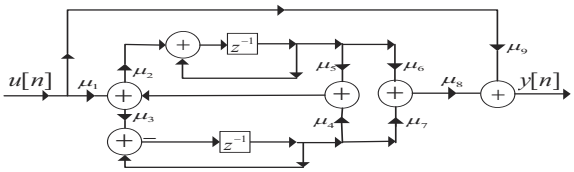


Fig. 1: Block-diagram of the structure $S_{WDF}$ proposed in [6].

Denote $h_m[n]$ be as the unit impulse response of $T_m(z)$, then

$$x_m[n] = \sum_{k=-\infty}^{+\infty} h_m[k]u[n-k]$$

Alternatively, note that

$$x_m[n] = \int_{-1/2}^{1/2} T_m(e^{j2\pi f})U(e^{j2\pi f})e^{j2\pi fn}df$$

with $U(e^{j2\pi f})$ the discrete-time Fourier transform of $u[n]$ with $f$ the normalized frequency. According to the Hölder inequality, we can show that

$$|x_m[n]| \le ||T_m(z)||_p||U(z)||_q,\ \forall\ n \tag{8}$$

which holds for *any* pair $(p, q)$ satisfying

$$\frac{1}{p} + \frac{1}{q} = 1,\ \ 1 < p,\ q < \infty \tag{9}$$

where $||S(z)||_p$ is the $L_p$-norm of a (scalar) function $S(z)$:

$$||S(z)||_p \triangleq [\int_{-1/2}^{1/2} |S(e^{j2\pi f})|^p df]^{1/p} \tag{10}$$

It is assumed that there is no overflow as long as $x_m[n]$ is absolutely not bigger than one. (8) implies that for a given structure, the input signal $u[n]$ causes no overflow at the $m$th state node at all if there exists a pair $(\tilde{p}, \tilde{q})$ satisfying (9) such that

$$||T_m(z)||_{\tilde{p}}||U(z)||_{\tilde{q}} \le 1 \tag{11}$$

The classical $l_2$-scaled structures (see [2], [3]) are defined with $||T_m(z)||_2 = 1,\ \forall m$ and hence the structures can ensure no overflow for the class of input signals $||U(z)||_2 = 1$.

Note $||T_m(z)||_\infty$ yields the maximum of the magnitude response. Without loss of generality, it is assumed that $||T_m(z)||_\infty = 1$. Consider $u[n] = \delta[n]$ - the unit sample signal. As $U(z) = 1$ and $||U(z)||_q = 1,\ \forall\ q$, one has $||T_m(z)||_\infty\ ||U(z)||_1 = 1$ and hence no overflow occurs. As the following holds:[3]

$$||T_m(z)||_q \le ||T_m(z)||_{\tilde{q}},\ q \le \tilde{q}$$

one has $|x_m[n]| \le ||T_m(z)||_1||U(z)||_\infty$ and hence

$$|x_m[n]| \le ||T_m(z)||_1 < ||T_m(z)||_\infty = 1$$

[3]It should be noted that the following does NOT hold: $||w[n]||_q \le ||w[n]||_{\tilde{q}},\ q \le \tilde{q}$.

Since the gap between $||.||_1$ and $||.||_\infty$ can be very big, the number of bit assigned to $x_m[n]$ may not be fully used or more seriously, $x_m[n]$ may be quantized to zero constantly if $||T_m(z)||_1$ is too small. Therefore, it is desired for a structure to have a flat magnitude frequency response for each $T_m(z)$. This is one of the reasons that a performance measure for a given structure, called *peakedness*, was proposed in [9]:

$$P_{eak}(m) \triangleq \frac{||T_m(z)||_\infty}{||T_m(z)||_1}, \ \forall\, m \tag{12}$$

The smaller this value is, the flatter the magnitude response $|T_m(z)|$ is.

The peakedness $P_{eak}(m)$ reflects a property of a given signal node. The overall peakedness of a structure can be measured with

$$P_{eak} \triangleq \max_m P_{eak}(m)$$

Based on the discussions above, it is desired for a structure, besides simplicity, to have $P_{eak}$ close to one. This argument motivates us to make use of all-pass filters as basic blocks for digital filter implementation.

## 3 All-pass Based Digital Filter Structures

A digital filter $H_{ap}(z)$ is said to be all-pass if

$$|H_{ap}(e^{j2\pi f})| = 1, \ \forall\, f \in [-1/2,\ 1/2] \tag{13}$$

As well known, the $N$-th order (real) all-pass filters are of the following form

$$H_{ap}(z) = \frac{z^{-N}A(z^{-1})}{A(z)} \tag{14}$$

where $A(z) = 1 - a_1 z^{-1} + \cdots - a_k z^{-k} + \cdots - a_N z^{-N}$, as defined before, with all $a_k$ are real-valued. For convenience, the 1st order all-pass filters are denoted as $A_p(\alpha, z)$:

$$A_p(\alpha, z) \triangleq \frac{\alpha^* - z^{-1}}{1 - \alpha z^{-1}} \tag{15}$$

If the transfer function $T_m(z)$ between the input $u[n]$ and the $m$-th state $x_m[n]$ is all-pass, then $P_{eak}(m) = 1$. Furthermore, if this is true for all the states, then the scaling factor $2^{-B_q}$ is equal to 1, and hence the scaling is avoided.

### 3.1 Structures for 1st and 2nd order filters

Consider a 1st-order (real) filter of transfer function $H(z)$ given by $H(z) = \frac{b_0 + b_1 z^{-1}}{1 - a_1 z^{-1}}$, which can be rewritten as

$$H(z) = c_0 + c\,\frac{a_1 - z^{-1}}{1 - a_1 z^{-1}} = c_0 + cA_p(a_1, z) \tag{16}$$

where $b_0 = c_0 + ca_1$ and $b_1 = -c_0 a_1 - c$. Therefore, with $u[n]$ and $y[n]$ being the input and output, respectively, the filter can be implemented with

$$\left\{ \begin{array}{rcl} x[n] &=& a_1\{x[n-1] + u[n]\} - u[n-1] \\ y[n] &=& cx[n] + c_0 u[n] \end{array} \right. \tag{17}$$

where the state $x[n]$ is the output of the 1st order all-pass filter $T(z) = A_p(a_1, z)$ excited by $u[n]$.

Now, let us consider a 2nd-order filter given by

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 - a_1 z^{-1} - a_2 z^{-2}} \tag{18}$$

where $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} = (1 - \alpha z^{-1})(1 - \alpha^* z^{-1})$, where $\alpha = \alpha_1 + j\alpha_2$ with $\alpha_1, \alpha_2$ are all real numbers. Such a transfer function can be decomposed into

$$H(z) = c_0 + h_1 A_p(\alpha, z) + h_2 A_p(\alpha^*, z) \tag{19}$$

where the constant coefficients $c_0, h_1$, and $h_2$ can be determined from

$$\begin{array}{rcl} B(z) &=& c_0 A(z) + h_1(\alpha^* - z^{-1})(1 - \alpha^* z^{-1}) \\ && + h_2(\alpha - z^{-1})(1 - \alpha z^{-1}) \end{array}$$

which yields

$$\left\{ \begin{array}{rcl} b_0 &=& c_0 + \alpha^* h_1 + \alpha h_2 \\ b_1 &=& -c_0 a_1 - (1 + \alpha^*)h_1 - (1 + \alpha)h_2 \\ b_2 &=& -c_0 a_2 + \alpha^* h_1 + \alpha h_2 \end{array} \right. \tag{20}$$

It is noted that with $\{b_k\}$ real-valued, $c_0$ is real-valued and that $h_1 = h_2^*$, based on the latter we can can show that with a real-valued input $u[n]$, the output $y_1[n]$ of the system $h_1 A_p(\alpha, z)$ is equal to the complex conjugate of that of the system $h_2 A_p(\alpha^*, z)$. It is then easy to understand that the output of the 2nd-order filter is given by

$$y[n] = c_0 u[n] + 2w[n]$$

where $w[n]$ is the real part of $y_1[n]$. Denote $x[n]$ as the output of $A_p(\alpha, z)$ excited by $u[n]$, one has mathematically

$$\left\{ \begin{array}{rcl} x[n] &=& \alpha x[n-1] + \alpha^* u[n] - u[n-1] \\ y_1[n] &=& h_1 x[n] \end{array} \right. \tag{21}$$

Let $x[n] = x_1[n] + jx_2[n],\ 2h_1 = c_1 - jc_2$. It can be shown that (21), a complex 1st order filter, can be implemented with the following equations in which the numbers and signals involved are all real-valued:

$$\left\{ \begin{array}{rcl} x_1[n] &=& \alpha_1(x_1[n-1] + u[n]) \\ && -\alpha_2 x_2[n-1] - u[n-1] \\ x_2[n] &=& \alpha_2(x_1[n-1] - u[n]) + \alpha_1 x_2[n-1] \end{array} \right.$$

and the output of the (real) 2nd order filter $y[n]$ is given by

$$y[n] = c_1 x_1[n] + c_2 x_2[n] + c_0 u[n]$$

The transfer function $T_k(z)$ form the input $u[n]$ to the state $x_k[n]$ can be obtained easily from the above for $k = 1, 2$:

$$\left\{ \begin{array}{rcl} T_1(z) &=& \frac{\alpha_1 + (\alpha_2^2 - \alpha_1^2 - 1)z^{-1} + \alpha_1 z^{-2}}{1 - 2\alpha_1 z^{-1} + (\alpha_1^2 + \alpha_2^2)z^{-2}} \\ T_2(z) &=& \frac{-\alpha_2 + 2\alpha_1\alpha_2 z^{-1} - \alpha_2 z^{-2}}{1 - 2\alpha_1 z^{-1} + (\alpha_1^2 + \alpha_2^2)z^{-2}} \end{array} \right. \tag{22}$$

Since $x[n] = x_1[n] + jx_2[n]$ is the output of the all-pass filter $A_p(\alpha, z)$, one has $A_p(\alpha, z) = T_1(z) + j\,T_2(z)$. It then follows from $A_p(\alpha, e^{j2\pi f})A_p^*(\alpha, e^{j2\pi f}) = 1$ that

$$|T_1(e^{j2\pi f})|^2 + |T_2(e^{j2\pi f})|^2 = 1, \ \forall\, f \tag{23}$$

Note that $T_2(e^{j2\pi f_0}) = 0$, where $\alpha_1 = cos(2\pi f_0)$, one concludes that

$$||T_1(z)||_\infty = 1, \;\; ||T_2(z)||_\infty \leq 1$$

It follows from the fact that $A_p(\alpha, z)$ is all-pass that

$$|x[n]| \leq ||U(z)||_q \tag{24}$$

Though $T_k(z)$ is not all-pass, (24) and $x[n] = x_1[n] + jx_2[n]$ imply that

$$|x_k[n]| \leq ||U(z)||_q, \;\; \forall\, q, \;\; k = 1, 2 \tag{25}$$

This means that for any input signals normalized in any $L_q$-norm, that is $||U(z)||_q = 1$, the states of our proposed structure are automatically bounded by one and hence no overflow occurs in the structure. Furthermore, since $x_2[n]$ is the output of $T_2(z)$ in response of $u[n]$ we have

$$|x_2[n]| \leq ||T_2(z)||_p ||U(z)||_q$$

This means that the state $x_2[n]$ can be scaled by $||T_2(z)||_p$ for a given $p$. With $x_2[n]$ replaced by $\epsilon_p x_2[n]$, where $\epsilon_p \triangleq ||T_2(z)||_p^{-1}$, our proposed structure is then specified by

$$\left\{ \begin{array}{rcl} x_1[n] &=& \gamma_1(x_1[n-1] + u[n]) \\ & & -\gamma_2 x_2[n-1] - u[n-1] \\ x_2[n] &=& \gamma_3(x_1[n-1] - u[n]) + \gamma_4 x_2[n-1] \\ y[n] &=& \gamma_5 x_1[n] + \gamma_6 x_2[n] + \gamma_7 u[n] \end{array} \right. \tag{26}$$

where $\gamma_1 = \alpha_1$, $\gamma_2 = \epsilon_p^{-1}\alpha_2$, $\gamma_3 = \epsilon_p\, \alpha_2$, $\gamma_4 = \alpha_1, \gamma_5 = c_1$, $\gamma_6 = \epsilon_p^{-1} c_2$, $\gamma_7 = c_0$ and the transfer function $\tilde{T}_k(z)$ form the input $u[n]$ to the state $x_k[n]$ can be obtained easily for $k = 1, 2$:

$$\left\{ \begin{array}{rcl} \tilde{T}_1(z) &=& \frac{\gamma_1 + (\gamma_2\gamma_3 - \gamma_1\gamma_4 - 1)z^{-1} + \gamma_4 z^{-2}}{1 - (\gamma_1 + \gamma_4)z^{-1} + (\gamma_1\gamma_4 + \gamma_2\gamma_3)z^{-2}} \\ \tilde{T}_2(z) &=& \frac{-\gamma_3 + 2\gamma_1\gamma_3 z^{-1} - \gamma_3 z^{-2}}{1 - (\gamma_1 + \gamma_4)z^{-1} + (\gamma_1\gamma_4 + \gamma_2\gamma_3)z^{-2}} \end{array} \right. \tag{27}$$

and hence $H(z) = \gamma_7 + \gamma_5\tilde{T}_1(z) + \gamma_6\tilde{T}_2(z)$.

Fig. 2 gives the block-diagram of the structure specified by (26).

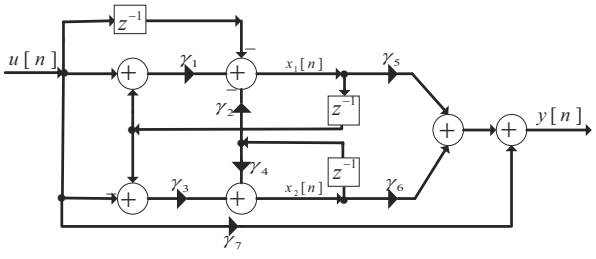


Fig. 2: Block-diagram of the proposed all-pass based structure for 2nd-order filters.

**Remark 3.1**:

- It is observed that for a 2nd order filter implemented using our proposed structure, 7 multiplications and 7 additions are needed for computing each sample of the output, a complexity slightly higher than those canonical structures such as the direction-form based structures but more efficient than the $S_{NF}$ and $S_{WDF}$.
- More importantly, our proposed structure is more robust to the input signals and hence yields a better performance against overflow and the FWL effects. All these will be demonstrated in Section 5.

### 3.2 Nth order digital filter implementation

Let $H(z)$ be the transfer function of an $N$-th order filter that has $2N_c$ complex conjugate poles $\{\alpha_k\}$ and $N_r$ real poles $\{r_k\}$ with $N = 2N_c + N_r$. Furthermore, to simplify the presentation it is assumed that there are no repeated poles.[4] Consequently, $H(z)$ can be decomposed as

$$\begin{array}{rcl} H(z) &=& \dfrac{b_0 + b_1 z^{-1} + \cdots + b_N z^{-N}}{1 + a_1 z^{-1} + \cdots + a_N z^{-N}} \triangleq \dfrac{B(z)}{A(z)} \\ &=& d + \displaystyle\sum_{k=1}^{N_r} c_k A_p(r_k, z) + \sum_{k=1}^{2N_c} h_k A_p(\alpha_k, z) \end{array}$$

where $\alpha_{2m} = \alpha_{2m-1}^*, \; \forall\, m$ is assumed.

Denote $\mathbf{b} \triangleq \left[ \begin{array}{ccc} b_0 & \cdots & b_N \end{array} \right]^T$, $\theta \triangleq \left[ \begin{array}{ccccccc} d & c_1 & \cdots & c_{N_r} & h_1 & \cdots & h_{2N_c} \end{array} \right]^T$, it can be shown that $\mathbf{b} = \mathbf{R}\theta$, where the $(N+1) \times (N+1)$ matrix $\mathbf{R} = \left[ \begin{array}{ccccccc} r_0 & q_1 & \cdots & q_{N_r} & p_1 & \cdots & p_{2N_c} \end{array} \right]$ with $r_0$ formed with the coefficients of $A(z)$, that is

$$r_0 = \left[ \begin{array}{ccccc} 1 & a_1 & b_2 & \cdots & a_N \end{array} \right]^T$$

while $q_k$ and $p_k$ constructed with the coefficients of $(r_k - z^{-1}) \prod_{\forall\, m \neq k}(1 - r_m z^{-1}) \prod_{m=1}^{2N_c}(1 - \alpha_m z^{-1})$ and $(\alpha_k^* - z^{-1}) \prod_{\forall\, m \neq k}(1 - \alpha_m z^{-1}) \prod_{m=1}^{N_r}(1 - r_m z^{-1})$, respectively. Therefore,

$$\theta = \mathbf{R}^{-1}\mathbf{b} \tag{28}$$

As $A(z), B(z)$ are polynomials (in $z^{-1}$) of real-valued coefficients, it can be shown that $d, c_k$ are all real, while the coefficients $h_k$ are complex and $h_{2k} = h_{2k-1}^*, \; \forall\, k$.

The output of $H(z)$ realized using our proposed structure is given by

$$y[n] = du[n] + \sum_{k=1}^{N_r} c_k x_k[n] + \sum_{k=1}^{N_c} y_{ck}[n] \tag{29}$$

where $x_k[n]$ is the output of the 1st order all-pass filter $A_p(r_k, z^{-1})$, computed using the first equation of (17) with $a_1$ replaced by $r_k$, while $y_{ck}[n]$ is the output of the 2nd order (real) filter

$$H_k(z) \triangleq h_{2k-1} A_p(\alpha_k, z^{-1}) + h_{2k-1}^* A_p(\alpha_k^*, z^{-1})$$

which can be implemented using (26), where $\alpha_1$ and $\alpha_2$ are the real and imaginary parts of $\alpha_k = \alpha_{1k} + j\alpha_{2k}$, $c_1, c_2$ are the real and imaginary parts of $2h_{2k-1} = c_{1k} - jc_{2k}$, and the corresponding states are now denoted as $x_{1k}[n]$ and $x_{2k}[n]$.

In the remainder of this paper, the proposed structure in this section is referred to as $S_{APF}$ for convenience.

[4] If $H(z)$ has a pole $\alpha$ with multiplicity $m$, then $H(z) = H_0(z) + \sum_{k=1}^{m} h_k A_p^k(\alpha, z^{-1})$ with $H_0(z)$ has no poles at $z = \alpha$. The part $\sum_{k=1}^{m} h_k A_p^k(\alpha, z^{-1})$ can be implemented using the all-pass filter $A_p^m(\alpha, z^{-1})$ efficiently. The details are not presented due to the limited space given.

## 4 Roundoff Noise Analysis

In this section, we investigate the performance of the proposed structure against roundoff noises that occur in the structure.

Let $\gamma$ be a parameter in a filter structure. In an actual implementation of less-than-double precision with rounding after multiplication, the product $\gamma s[n]$ has to be rounded by a quantizer $Q[\cdot]$. Denote $\epsilon_\gamma[n] \stackrel{\triangle}{=} \psi(\gamma)\{Q[\gamma s[n]] - \gamma s[n]\}$ as the roundoff noise, where $\psi(\gamma) = 1$ if $\gamma$ is nontrival, otherwise, $\psi(\gamma) = 0$. In fact, the function $\psi$ is used for indicating the fact that $\gamma$ produces no roundoff noise when it is trivial. Roundoff noises are classically modeled as statistically independent white processes with an uniform distribution within $[-\frac{2^{-B_s}}{2}, \frac{2^{-B_s}}{2}]$, where $B_s$ is the number of bits assigned for representing the fractional part of the signals.

Denote $\Delta y[n]$ as the output deviation of the filter due to $\epsilon_\gamma[n]$ and $F(z)$ as the transfer function between $\epsilon_\gamma[n]$ and $\Delta y[n]$. It is well known that $\Delta y[n]$ is a stationary process and the variance $\sigma^2_{\Delta y} = E[|\Delta y[n]|^2]$, which can be shown [3] that

$$\sigma^2_{\Delta y} = \psi(\gamma)||F(z)||_2^2 \sigma_0^2$$

where $\sigma_0^2 = \frac{2^{-2B_s}}{12}$. The roundoff noise gain for $\gamma$ is defined as $G_\gamma \stackrel{\triangle}{=} \frac{\sigma^2_{\Delta y}}{\sigma_0^2}$ and clearly,

$$G_\gamma = \psi(\gamma)||F(z)||_2^2 \tag{30}$$

where $||F(z)||_2$ is the $L_2$-norm defined before.

With some manipulations it can be shown that when $F(z) = D + J(I - z^{-1}\Phi)^{-1}L$,

$$||F(z)||_2^2 \;=\; tr[D(D + 2JL) + JW_cJ^T] \tag{31}$$

where $W_c$ is the controllability Gramian of the realization $(\Phi, L, J, D)$,[5] satisfying

$$W_c = \Phi W_c \Phi^{\mathcal{T}} + LL^{\mathcal{T}}$$

Now, let us analyze the roundoff noise effects in the proposed structure for a 2nd order filter, specified by (26) in which there are 7 multiplications involved in general.

Let $\epsilon_{\gamma_k}[n]$ denote the roundoff noise due to the multiplier $\gamma_k$ for all $k$. Taking $\gamma_1$ as example, we now shown how to derive $G_{\gamma_1}$.

Due to the introduction of $\epsilon_{\gamma_1}[n]$, (26) becomes

$$\left\{\begin{array}{rcl} \tilde{x}_1[n] &=& \gamma_1(\tilde{x}_1[n-1] + u[n]) \\ && -\gamma_2\tilde{x}_2[n-1] - u[n-1] + \epsilon_{\gamma_1}[n] \\ \tilde{x}_2[n] &=& \gamma_3(\tilde{x}_1[n-1] - u[n]) + \gamma_4\tilde{x}_2[n-1] \\ \tilde{y}[n] &=& \gamma_5\tilde{x}_1[n] + \gamma_6\tilde{x}_2[n] + \gamma_7 u[n] \end{array}\right.$$

Denote $e_k[n] \stackrel{\triangle}{=} \tilde{x}_k[n] - x_k[n],\ k = 1, 2$ and $e_y[n] \stackrel{\triangle}{=} \tilde{y}[n] - y[n]$. It then turns out that

$$\left\{\begin{array}{rcl} e_1[n] &=& \gamma_1 e_1[n-1] - \gamma_2 e_2[n-1] + \epsilon_{\gamma_1}[n] \\ e_2[n] &=& \gamma_3 e_1[n-1] + \gamma_4 e_2[n-1] \\ e_y[n] &=& \gamma_5 e_1[n] + \gamma_6 e_2[n] \end{array}\right. \tag{32}$$

[5] In a strict sense, $(\Phi, L, J, D)$ is not a realization of $F(z)$ but of $G(z) \stackrel{\triangle}{=} D + J(zI - \Phi)^{-1}L = D(1 - z^{-1}) + z^{-1}F(z)$.

and clearly, the transfer function, denoted as $F_1(z)$, between $\epsilon_{\gamma_1}[n]$ to $e_y[n]$ is given by

$$F_1(z) = D_1 + J_1(I - z^{-1}\Phi_1)^{-1}L_1$$

where the realization $(\Phi_1, L_1, J_1, D_1)$ given by

$$\left\{\begin{array}{rcl} \Phi_1 &=& \left[\begin{array}{cc} \gamma_1 & -\gamma_2 \\ \gamma_3 & \gamma_4 \end{array}\right],\ L_1 = \left[\begin{array}{c} 1 \\ 0 \end{array}\right] \\ J_1 &=& \left[\begin{array}{cc} \gamma_5 & \gamma_6 \end{array}\right],\ D_1 = 0 \end{array}\right. \tag{33}$$

and the roundoff noise gain $G_{\gamma_1} = \psi(\gamma_1)||F_1(z)||_2^2$ can then be computed using (31).

In the same manner, we can analysis the effect of $\epsilon_{\gamma_2}[n]$ on the filter output. It can be shown that $F_2(z) = F_1(z)$ and hence $G_{\gamma_2} = \psi(\gamma_2)||F_1(z)||_2^2$.

Using the same procedure, one can show that

$$F_3(z) = F_4(z) = D_1 + J_1(I - z^{-1}\Phi_1)^{-1}L_3$$

where $L_3 = \left[\begin{array}{cc} 0 & 1 \end{array}\right]^T$. Furthermore, $F_k(z) = 1,\ k = 5, 6, 7$ and hence $G_{\gamma_k} = \psi(\gamma_k),\ k = 5, 6, 7$.

The total roundoff noise gain of the structure is defined as

$$G_{APF} = \sum_{k=1}^{7} \psi(\gamma_k)||F_k(z)||_2^2 \tag{34}$$

The roundoff noise behavior of the structures $S_{NF}$ and $S_{WDF}$, proposed in [2] and [6], respectively, can be analyzed in a similar way. Due to the limited space, the results are not presented in this paper.

The overall roundoff noise gain of the proposed structure for a given $N$th-order filter with $N > 2$, as realized using a parallel form, is the sum of the total roundoff noise gains of the sub-systems of 1st-order and 2nd-order.

## 5 A Numerical Example And Simulations

In this section, we present an example to illustrate the performance of the proposed structure $S_{APF}$ and to compare it with $S_{NF}$ and $S_{WDF}$ proposed in [2] and in [6], respectively.

**Example:** This example is a sixth-order low-pass Butterworth filter of a normalized bandwidth $0.125$, generated with MATLAB command $[bb,\ aa] = butter(6,\ 0.25)$.

As it has three complex conjugate pairs of poles, the filter is implemented in a parallel form of three 2nd-order sectors, each of which is implemented using the same structure $S_x$, where $S_x = S_{WDF}, S_{NF}$ and $S_{APF}$, respectively.

Table 1: Statistics for the three structures

| | $G$ | $N_M$ | $N_A$ |
|---|---|---|---|
| $S_{APF}$ | $4.16 \times 10^1$ | 19 | 19 |
| $S_{NF}$ | $1.19 \times 10^2$ | 22 | 13 |
| $S_{WDF}$ | $1.58 \times 10^2$ | 25 | 16 |

Table 1 yields the total roundoff noise gain[6] and structure complexity measured by the number of multiplications

[6] Note both $S_{NF}$ and $S_{WDF}$ require the input signals scaled by $2^{-1}$ in order to avoid overflow and hence the output is then scaled by and $2^1$.

$N_m$ and additions $N_A$ required for computing each output sample.

For a given structure, the peakedness of each state node is computed and presented in Table 2.

Table 2: Peakedness for the three structures

| $P_{eak}(m)$ | $S_{APF}$ | $S_{NF}$ | $S_{WDF}$ |
|---|---|---|---|
| 1 | 1.4435 | 2.7790 | 2.7997 |
| 2 | 1.5664 | 3.1540 | 2.8102 |
| 3 | 1.1524 | 1.5424 | 1.4623 |
| 4 | 2.1785 | 2.0735 | 1.4691 |
| 5 | 1.0145 | 1.5889 | 1.2612 |
| 6 | 2.5512 | 2.3706 | 1.2270 |

For each structure, the actual rounding off (with states are represented in a $B_s$-format) and overflow (bounded by one) are carried out for three different input signals $u[n]$ under the constraint $\max_n |u[n]| = 1$. Table 3 yields the energy of the output error sequence of the filter realized using each structure for different input signals (of 1,000 samples), where the three signals used for testing are: $u_1[n] = 0.4\{sin(2\pi 0.125n) + sin(2\pi 0.2n) + sin(2\pi 0.35n)\}$, while $u_2[n]$ and $u_3[n]$ are random signals with Gaussian and uniform distribution, respectively.

Table 3: Energy of the output error sequence when the state variables implemented with $B_s = 12bits$.

| | $u_1[n]$ | $u_2[n]$ | $u_3[n]$ |
|---|---|---|---|
| $S_{APF}$ | 0.0144 | $9.1892\times10^{-4}$ | 0.8769 |
| $S_{NF}$ | 6.2356 | 0.0769 | 45.4650 |
| $S_{WDF}$ | 4.5314 | 0.1284 | 61.5699 |

The output signals in response to the inputs $u_1[n]$ and $u_3[n]$ are shown in Fig.s 3 and 4, where the solid line yields the ideal output, while the dotted, dashed and dash-dotted lines are those for $S_{APF}$, $S_{NF}$ and $S_{WDF}$, respectively.

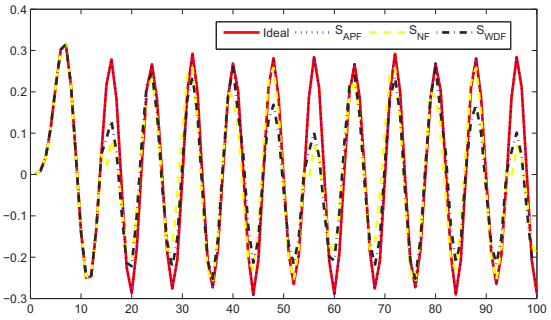


Fig. 3: The filter responses to $u_1[n]$ for different structures.

**Remark 5.1:** The results are self-explanatory. Theoretical results show that our proposed structure has smaller structure peakedness and hence should yield a better performance than the two other. This is confirmed by simulations.

## 6 Conclusions

In this paper, based on all-pass systems a novel structure for digital filter implementation has been derived. It has been shown that such a structure possesses nice properties against FWL effects such as overflow and roundoff noise. Simulations have been carried out, which demonstrate the superior performance of our proposed structure over that of two existing ones.

This piece of work inspires the use of power-complementary sub-systems for developing robust and efficient filter structures. Further researches in this direction are on-going.

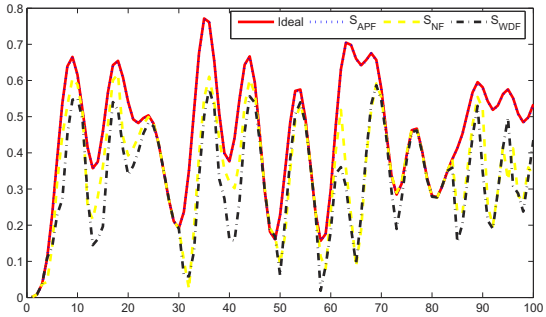


Fig. 4: The filter responses to $u_3[n]$ for different structures.

## Acknowledgment

This work was supported by the NSFC-Grant 61273195, CPSF-Grant 2012M511386 and ZSFC-Grant Y13F010050.

## References

[1] R.E. Kalman, "Nonlinear aspects of sampled-data control systems," *Proc. of Symposium on nonlinear circuit theory*, Brooklyn, NY, 1965.

[2] R.A. Roberts and C.T. Mullis, *Digital Signal Processing*, Reading, MA: Addison Wesley, 1987.

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

[4] Y. Wang and K. Roy, "CSDC: A new complexity reduction technique for multiplierless implementation of FIR filters," *IEEE Trans. Circuits Sysm. I*, vol. 52, no. 9, pp. 1845 - 1853, Sept., 2005.

[5] J. Skaf and S.P. Boyd, "Filter design with low complexity coefficients," *IEEE Trans. on Signal Processing*, *vol. 56*, no. 7, pp. 3162 - 3169, Jul., 2008.

[6] J.H.F. Ritzerfeld, "Noise gain expressions for low noise second-order digital filter structures," *IEEE Trans. on Circuits and Systems - II*, vol. 52, no.4, pp. 223-227, Apr., 2005.

[7] G. Li, Y.C. Lim, and C.G. Huang, "Very robust low complexity lattice filters," *IEEE Trans. on Signal Processing*, vol. 58, no. 12, pp. 6093-6104, Dec., 2010.

[8] G. Li, Y.C. Lim, C.G. Huang, and H. Xu, "A novel digital IIR filter design strategy - structure-based discrete coefficient filters," in *Proc. of Int. Symp. on Circuits and Systems*, Seoul, Korea, pp. 37-40, May 18-23, 2012.

[9] Y.C. Lim, "On the synthesis of IIR digital filters derived from single channel AR lattice network," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 4, pp. 741 - 749, Aug., 1984.