Appl. Sci. **2014**, *4*, 99-127; doi:10.3390/app4020099

OPEN ACCESS

applied sciences

ISSN 2076-3417

www.mdpi.com/journal/applsci

Article

Round-Off Noise of Multiplicative FIR Filters Implemented on an FPGA Platform

Jean-Jacques Vandenbussche [1,*], Peter Lee [2] and Joan Peuteman [1]

[1] Department ESAT, KU Leuven Kulab, Zeedijk 101 Ostend 8400, Belgium;
E-Mail: joan.peuteman@kuleuven.be

[2] School of Engineering and Digital Arts, University of Kent, Canterbury Kent CT2-7NT, UK;
E-Mail: P.Lee@kent.ac.uk

* Author to whom correspondence should be addressed;
E-Mail: jeanjacques.vandenbussche@kuleuven.be; Tel.: +32-59-569-017; Fax: +32-59-569-001.

Received: 28 November 2013; in revised form: 7 February 2014 / Accepted: 19 February 2014 / Published: 25 March 2014

Abstract: The paper analyzes the effects of round-off noise on Multiplicative Finite Impulse Response (MFIR) filters used to approximate the behavior of pole filters. General expressions to calculate the signal to round-off noise ratio of a cascade structure of Finite Impulse Response (FIR) filters are obtained and applied on the special case of MFIR filters. The analysis is based on fixed-point implementations, which are most common in digital signal processing algorithms implemented in Field-Programmable Gate-Array (FPGA) technology. Three well known scaling methods, *i.e.*, L_2 bound; infinity bound and absolute bound scaling are considered and compared. The paper shows that the ordering of the MFIR stages, in combination with the scaling methods, have an important impact on the round-off noise. An optimal ordering of the stages for a chosen scaling method can improve the round-off noise performance by 20 dB.

Keywords: MFIR; FIR-filters; linear phase filters; FPGA; fixed point digital signal processing DSP; round-off noise; filter cascade structure

---

1. Introduction

Multiplicative Finite Impulse Response (MFIR) filters are a class of filter structures that were originally introduced by Fam in the early 1980s [1]. It was shown that MFIR filters can be used to

replace recursive Infinite Impulse Response (IIR) filters with FIR equivalents requiring significantly less hardware than classical FIR architectures that fulfill the same specifications [2]. The replacement of a pole that is implemented in a recursive structure, by a non-recursive FIR, has the advantage that it will always be stable. This is particularly interesting when the original pole is situated close to the unit circle. The MFIR filters are able to realize low-pass, high-pass, band-pass, and notch filters. Although the MFIR structures require approximately the same number of delay elements as the classical FIR implementations, they require, logarithmically, fewer adders and multipliers [1,2].

MFIR filters are based on the identity [3]:

$$\sum_{i=0}^{2^P-1} x^i = \prod_{i=0}^{P-1}\left(1+x^{2^i}\right). \tag{1}$$

This identity can be used to approximate both real pole and conjugate pole pair filters. In case $H_r(z)$ is the transfer function of a stable IIR filter with a real pole, the cascade MFIR filter approximation using Equation (1) yields:

$$\begin{aligned} H_r(z) &= \frac{1}{1-\lambda z^{-1}} = \sum_{i=0}^{\infty}\left(\lambda z^{-1}\right)^i \qquad |\lambda|<1 \\ H_r(z) &\approx \sum_{i=0}^{2^P-1}\left(\lambda z^{-1}\right)^i = \prod_{i=0}^{P-1}\left(1+\left(\lambda z^{-1}\right)^{2^i}\right) = \prod_{i=0}^{P-1} M_i(z) = M(z). \end{aligned} \tag{2}$$

with the obvious definitions for $M_i(z)$ and $M(z)$. In this correspondence, every single $M_i(z)$ will be called a "stage" of the MFIR filter. It has been shown in [1,2] that even for the approximation of poles extremely close to the unit circle, maximum 10 stages (P = 10) are required when a deviation smaller than |0.01| dB in the magnitude response between the IIR filter and the MFIR approximation is allowed. The efficiency of the MFIR approximation is immediately clear from Equation (2). Only P multipliers and adders are required for the MFIR filter while the direct form requires 2^P multipliers and adders.

An IIR filter with a transfer function $H_c(z)$ having a conjugate pole pair $\lambda = re^{+j\theta}$ and $\lambda^* = re^{-j\theta}$ with $|r| < 1$, can be approximated with a cascade MFIR structure [1,2]:

$$\begin{aligned} H_c(z) &= \frac{1}{\left(1-\lambda z^{-1}\right)\left(1-\lambda^* z^{-1}\right)} \\ &\approx \prod_{i=0}^{P-1}\left(1+\left(\lambda z^{-1}\right)^{2^i}\right)\prod_{i=0}^{P-1}\left(1+\left(\lambda^* z^{-1}\right)^{2^i}\right) \\ &\approx \prod_{i=0}^{P-1}\left(1+2\left(rz^{-1}\right)^{2^i}\cos\left(2^i\theta\right)+\left(rz^{-1}\right)^{2^{i+1}}\right) = \prod_{i=0}^{P-1} M_i(z) = M(z) \end{aligned} \tag{3}$$

MFIR filters with a linear phase and a desired magnitude response $\left|H\left(e^{j\omega_n}\right)\right|$ (Here ω_n is the normalized angular frequency: $\omega_n = \omega T_S$, and T_S is the sampling period.) can also be realized using the following procedure [1,4].

1. Design an IIR filter that approximates $\left|H\left(e^{j\omega_n}\right)\right|^{1/2}$.
2. Approximate the poles of the IIR filter with the MFIR structure.
3. Cascade to every zero in the resulting MFIR filter its reciprocal with respect to the unit circle.

Consequently, the linear phase approximation of a stable IIR filter $H_r(z)$ with a real pole at λ, can be designed by determining $\left|H_r^{'}\left(e^{j\omega_n}\right)\right| = \left|H_r\left(e^{j\omega_n}\right)\right|^{1/2}$ and approximating the real pole λ' of $H_r'(z)$ with:

$$M(z) = \prod_{i=0}^{P-1}\left(1+\left(\lambda' z^{-1}\right)^{2^i}\right)\left(1+\left(\frac{z^{-1}}{\lambda'}\right)^{2^i}\right)$$
$$M(z) = \prod_{i=0}^{P-1}\left(1+\left((\lambda')^{2^i}+\frac{1}{(\lambda')^{2^i}}\right)z^{-2^i}+z^{-2^{i+1}}\right). \quad (4)$$

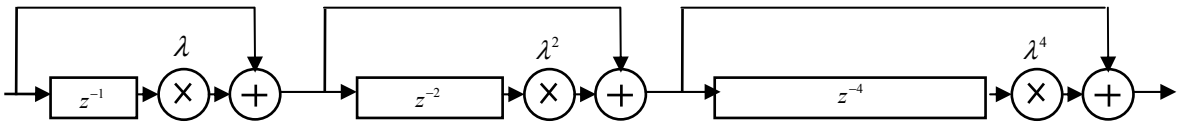
The linear phase approximation of a stable IIR filter $H_c(z)$ with a *complex*-conjugate pole pair $\lambda = re^{+j\theta}$ and $\lambda^* = re^{-j\theta}$, can be designed by determining $\left|H_c^{'}\left(e^{j\omega_n}\right)\right| = \left|H_c\left(e^{j\omega_n}\right)\right|^{1/2}$ and approximating the poles $\lambda' = r'e^{+j\theta'}$and $\lambda'^* = r'e^{-j\theta'}$ of $H_c'(z)$ with:

$$M(z) = \prod_{i=0}^{P-1}\begin{bmatrix}1+\left(2\left((r')^{2^i}+(r')^{-2^i}\right)\cos\left(2^i\theta'\right)\right)z^{-2^i}+\left((r')^{-2.2^i}+4\cos^2\left(2^i\theta'\right)+(r')^{2.2^i}\right)z^{-2.2^i}\\ +\left(2\left((r')^{2^i}+(r')^{-2^i}\right)\cos\left(2^i\theta'\right)\right)z^{-3.2^i}+z^{-4.2^i}\end{bmatrix}. \quad (5)$$

Despite the advantage of the logarithmically more efficient use of multipliers and adders, MFIR filters have not been popular. Indeed, the large number of delay elements required to approximate the behavior of an IIR filter was considered prohibitively expensive. This made them impractical for implementation on standard DSP platforms with fixed memory maps. Advances in Very Large Scale Integration VLSI technology in general, and Field-Programmable Gate-Array (FPGA) architectures in particular, make it necessary to re-evaluate MFIR filters and the technical barriers to their widespread use. Several applications and implementations of MFIR filters in modern FPGA fabrics [4–6], have shown that FPGAs are an ideal target platform for implementing efficient MFIR filters that are competitive to standard FIR and IIR equivalents implemented on the same fabric. The effects of coefficient-quantization have been studied and have shown that MFIR filter structures are less coefficient-quantization susceptible than the IIR filter they approximate [7].

It is clear from Equations (2–5) that MFIR filters are basically a cascade of simple sparse FIR filters. An example of a cascade of three MFIR stages realizing an MFIR filter approximating a real pole is given in Figure 1.

Figure 1. General architecture of an MFIR filter approximating a real pole (three stages).



Every stage in an MFIR filter has the same structure as the other stages but the number of delay elements differs. Therefore, it is obvious to design an optimized component that implements a general MFIR stage per MFIR filter approximation, *i.e.*, a real pole approximation, a complex-conjugate pole pair approximation and their respective linear phase types.

As floating-point arithmetic is only recently available in FPGAs, only fixed-point arithmetic is taken into consideration. Unfortunately, every fixed-point addition or multiplication requires a widening of the bus width, which has to be avoided; this implies that rounding must be applied to keep the data path widths manageable and implementable.

In this text, rounding is defined as the process whereby the width of the data path after a multiplication or an addition is reduced to the original width of the data path before the multiplication or addition. This is done by taking the most significant bits, conventional rounding of the result, and using saturation if necessary.

Scaling is defined as the process that changes the filter coefficients in order to increase the SNR based on the fulfillment of a specified criterion (as will be defined in Section 3) at the output of the filter.

In order to create general MFIR stage components, the data path bit-width at the input and at the output of each stage are kept constant. Inside the stage, the bit width is appropriately incremented to avoid accumulation of round-off errors. Consequently, in practice at the end of each stage, a rounding block will bring the output bit width back to the original input bit-width. It is, however, not excluded that a change in bus width between the stages would yield better results. However, there are so many possible combinations that it is almost impossible to investigate the behavior of all these possible implementations.

In this correspondence, the Signal to Noise Ratio SNR degradation effect of the consecutive roundings and scalings in the cascaded MFIR structures is analyzed and conclusions on optimal ordering of the stages are drawn. In Section 2, the round-off model and the different transfer functions are defined. Section 3 discusses the scaling methods as defined in [8]. Section 4 introduces a general theory for round-off noise determination in filter cascade structures, implying the theory developed in this section is general and is applicable on any filter having a cascade structure. In the fifth section, the developed general theory is applied on the MFIR structure. The paper ends with a conclusion and suggestions for future work.

Although it will not always be explicitly mentioned, in this correspondence, SNR refers to the Signal to round-off Noise Ratio, *i.e.*, noise due to rounding errors is considered and other noise sources are not taken into account.

2. The Round-Off Noise Model

2.1. Introduction

In order to avoid overflow of the signal data due to the successive multiplications of the signal with the filter coefficients, rounding (as defined above) will have to be performed. Rounding can be seen as a quantization action on the signal. Each rounding action is treated as a random process with uniform probability density function, producing white noise that is uncorrelated with the signal and other quantization sources in the filter. As in this correspondence the implementation of the MFIR filters is investigated for fixed-point implementation, the rounding process is of vital importance.

A. Fam developed, in [1], the “pure multiplicity property” which forms the basis of the investigation of the relationship between the noise variance at the output (relative to the round-off

noise variance) and the ordering of the different stages of the MFIR filter. This is done for an implementation without scaling (which is not very realistic for fixed point implementations) and an implementation with L_2 scaling. The analysis in [1] is performed for MFIR structures that approximate a real pole or a conjugate pole pair using the forward lattice structure. The approximation of a conjugate pole pair in cascade is not analyzed as "it does not have the pure multiplicity property" [1]. However, in the present research, the forward lattice MFIR implementation is not considered because of its (unpractical) large hardware impact [2]. Consequently, the study of the noise behavior of the complex-conjugate pole pair cascade MFIR implementation and the linear phase MFIR implementations are completely new.

Moreover, it is suggested in [1] that the analysis of the round-off noise for the MFIR structures that do not have the "pure multiplicity property" should be done in the style of [8] or [9]. In this text, the method of [8] will be used for the real pole approximation as well as for the complex-conjugate pole pair cascade approximation. All three well known scaling methods (L_2 bound, infinity bound and absolute bound) [8] will be considered. The noise performance is evaluated by first calculating the noise variance at the output due to the noise variance of the round-off error sources. However, a good evaluation of the round-off noise performance can only be made when the actual signal to round-off noise ratio (SNR) is considered [10]. With scaling, the output signal is also scaled, implying the actual signal to noise ratio at the output and not only the noise variance is considered for all possible scaling cases.

The objective of the paper is to determine how much the SNR of a signal is deteriorated by the noise (due to round-off errors) in the overall MFIR structure. Although the round-off noise performance of FIR and IIR cascade structures has been studied intensively over the past decades, [8–18] no directly applicable expression of the SNR degradation in cascade structures could be found. Therefore, the theory will first be developed in a general manner. More precisely, the stages in the cascade will not be considered as MFIR stages with transfer function $M_i(z)$, but as general filter stages with transfer function $H_k(z)$. The index k will be used as an index for the general stages in the cascade. Note that k is not the same as the index i that is used to indicate the stages of the MFIR structure *and* the power of the multiplier coefficients of the MFIR stages (as in Equations (2) and (3)).

It will be shown that the ordering has a large impact on the SNR. However, P stages of an MFIR structure can be ordered in $P!$ possible orderings, making an exhaustive search for the optimal ordering unpractical. Unlike typical cascade structures studied in [15–18], the transfer functions of the stages of the MFIR structure are fixed by the expressions given in Equations (2–5), *i.e.*, in every stage the grouping of the zeros is fixed by the MFIR approach. This implies that the optimization of the SNR performance can only focus on the ordering of the stages.

2.2. The Round-Off Noise Model

The study is based on the following assumptions. Each multiply and accumulate action in a stage is modeled as an infinite precision multiplier, followed by a summation node. After the summation node, rounding is performed and consequently round-off noise is added to the system. In the present paper, it is assumed that the rounding process uses conventional rounding and saturation. For the real pole

approximation, the conjugate pole pair cascade approximation and their respective linear phase approximations, each stage has one single noise source at the output of the stage. It is assumed that

- every sample of the noise source is uncorrelated with the previous sample,
- all noise sources are uncorrelated,
- the noise sources are uncorrelated with the input signal,
- every noise source is a time discrete stationary zero mean white random process with output variance $q^2/12$. Here, q is the smallest quantization step ($q = 2^{-b}$ where b is the number of bits (without the sign bit) used to quantize the signal).

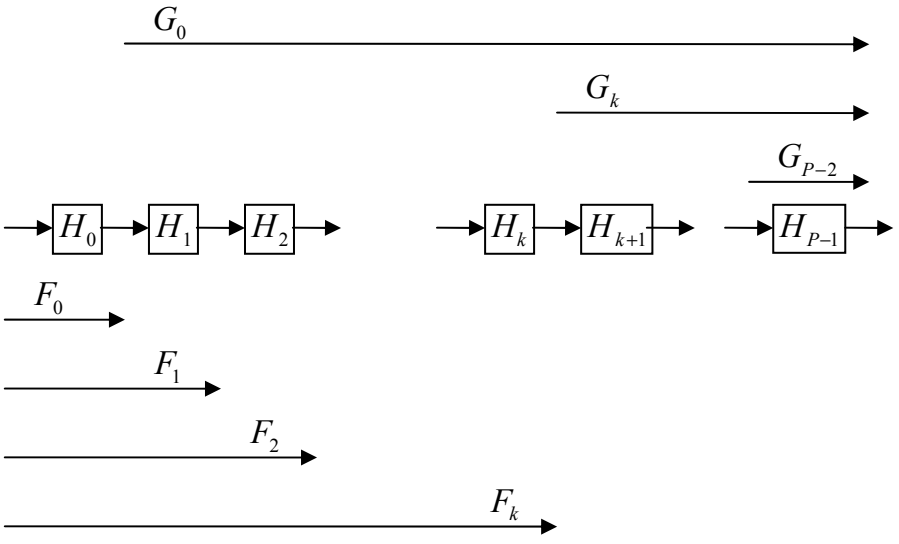
2.3. Symbol Conventions and Transfer Functions

The following definitions are used throughout the text.

Every stage without rounding or scaling is indicated by $H_k(z)$ or $H_k(e^{j\omega_n})$ or in short form H_k. The total filter transfer function is written as $H(z)$ or $H(e^{j\omega_n})$. A (general) stage with rounding and scaling is indicated by $\overline{H_k(z)}$, $\overline{H_k(e^{j\omega_n})}$ or in short form $\overline{H_k}$. The respective scaling factors per stage are indicated by S_k. The time samples of the round-off noise source are indicated by $e_k(n)$ (where n is the discrete time index) or in short form e_k. The noise variance of a noise source is given by σ_e^2.

A number of transfer functions are defined in Figure 2. $F_k(z)$ is the transfer function from the filter input to the output of the stage with transfer function $H_k(z)$ (without rounding). $G_k(z)$ is the transfer function from the output of the stage with transfer function $H_k(z)$ to the output of the filter (without rounding).

Figure 2. The filter cascade without rounding or scaling and its transfer functions.



For the round-off noise analysis, it is assumed there is a round-off noise source at the output of each stage. The transfer function from the filter input to the output of the stage with transfer function $\overline{H_k(z)}$ (noise source of stage k not included, as shown in Figures 3 and 4) is indicated by $\overline{F_k(z)}$.

The transfer function from the output of the stage with transfer function $\overline{H_k(z)}$ (noise source of stage k included, as shown in Figure 4) to the output of the total scaled filter is indicated by $\overline{G_k(z)}$.

Figure 3. The rounded and scaled filter cascade and its transfer functions.

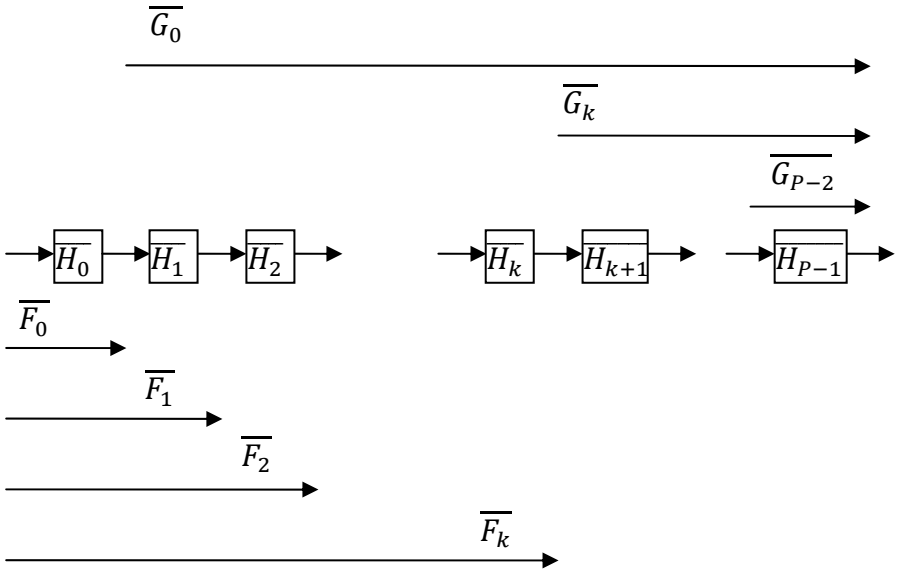
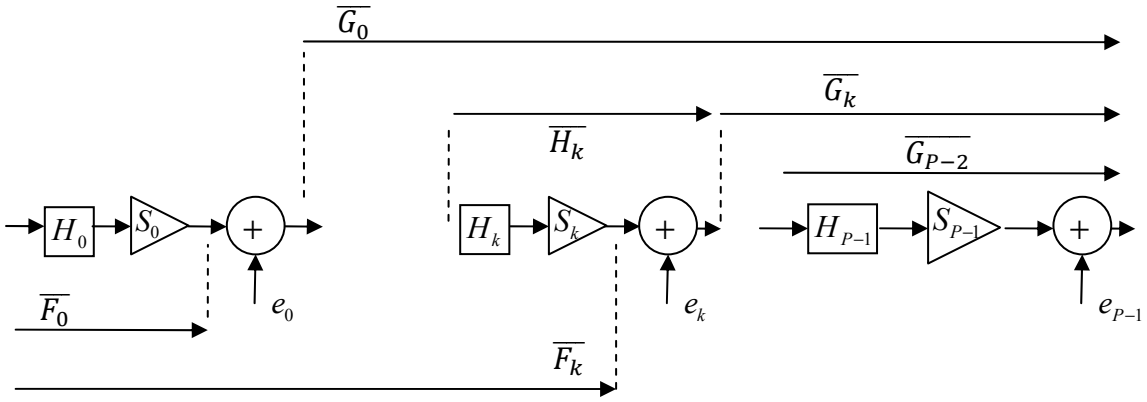


Figure 4. The "worked out" rounded and scaled filter cascade and its transfer functions.



Consequently, the following holds:

$$G_k(z) = \prod_{m=k+1}^{P-1} \left(H_m(z)\right), \tag{6}$$

$$\overline{G_k(z)} = \prod_{m=k+1}^{P-1} \left(\overline{H_m(z)}\right) = \prod_{m=k+1}^{P-1} \left(S_m H_m(z)\right) = \left(\prod_{m=k+1}^{P-1} (S_m)\right) G_k(z), \tag{7}$$

$$F_k(z) = \prod_{m=0}^{k} H_m(z), \tag{8}$$

$$\overline{F_k(z)} = \prod_{m=0}^{k} \left(\overline{H_m(z)}\right) = \prod_{m=0}^{k} \left(S_m H_m(z)\right) = \left(\prod_{m=0}^{k} (S_m)\right) F_k(z). \tag{9}$$

3. The General Scaling Methods

In analogy to [1] and [8], the considered scaling methods, are defined in this section. A scaling factor is used to multiply the filter stage coefficients in order to obtain a specific criterion at the output of the stage as defined in [8].

- For L_2 bound scaling, the scaling factors are determined by:

$$\prod_{m=0}^{k} S_m = \left[\left\| F_k\left(e^{j\omega_n}\right) \right\|_2 \right]^{-1}, \tag{10}$$

where $\left\| F_k\left(e^{j\omega_n}\right) \right\|_2$ is defined as the L_2 norm in the frequency domain, given by:

$$\left\| F_k\left(e^{j\omega_n}\right) \right\|_2 = \left[\frac{1}{2\pi} \int_{-\pi}^{+\pi} \left| F_k\left(e^{j\omega_n}\right) \right|^2 d\omega_n \right]^{\frac{1}{2}} = \left[\sum_n \left| f_k(n) \right|^2 \right]^{\frac{1}{2}}. \tag{11}$$

Here, $f_k(n)$ are the impulse response samples of the filter given by the transfer function $F_k(z)$. The recursive version of Equation (10) can be used to calculate the scale factor per stage. It is given by:

$$S_k = \begin{cases} \left(\left\| H_0\left(e^{j\omega_n}\right) \right\|_2 \right)^{-1} & k = 0 \\ \left(\prod_{m=0}^{k-1} S_m \right)^{-1} \left(\left\| \prod_{m=0}^{k} H_m\left(e^{j\omega_n}\right) \right\|_2 \right)^{-1} & k = 1 \ldots\ldots P-1. \end{cases} \tag{12}$$

The following holds when L_2 bound scaling is used: if the RMS value (over ω_n) of the input signal is bounded by unity, the RMS value (over ω_n) of the signal at each stage output will be bounded by unity.

- For infinity bound scaling, L_∞, the scaling factors are determined by:

$$\prod_{m=0}^{k} S_m = \left(\sup_{-\pi<\omega_n<+\pi} \left(\left| F_k\left(e^{j\omega_n}\right) \right| \right) \right)^{-1} \tag{13}$$

The recursive version of Equation (13) can be used to calculate the scale factor per stage. It is given by:

$$S_k = \begin{cases} \left(\sup_{-\pi<\omega_n<+\pi} \left| H_0\left(e^{j\omega_n}\right) \right| \right)^{-1} & k = 0 \\ \left(\prod_{m=0}^{k-1} S_m \right)^{-1} \left(\sup_{-\pi<\omega_n<+\pi} \left(\left| \prod_{m=0}^{k} H_m\left(e^{j\omega_n}\right) \right| \right) \right)^{-1} & k = 1 \ldots\ldots P-1. \end{cases} \tag{14}$$

The L_∞ bound scaling sets the maximum of the frequency responses of all respective $F_k(e^{j\omega n})$ at 0 dB.

- For absolute bound scaling, the scaling factors are determined by:

$$\prod_{m=0}^{k} S_m = \left(\sum_{n=0}^{p_k} \left| f_k(n) \right| \right)^{-1} \tag{15}$$

where $f_k(n)$ are the impulse response samples of the filter given by the transfer function $F_k(z)$. p_k is the length of the impulse response. Equation (15) in recursive form yields:

$$S_k = \begin{cases} \left(\sum_{n=0}^{p_0} \left| f_0(n) \right| \right)^{-1} & k = 0 \\ \left(\prod_{m=0}^{k-1} S_m \right)^{-1} \left(\sum_{n=0}^{p_k} \left| f_k(n) \right| \right)^{-1} & k = 1.....P-1. \end{cases} \tag{16}$$

Absolute bound scaling is based on the reasoning that if the peak value of the input signal is bounded by unity, the peak absolute value of the signal at each stage output will be bounded by unity when absolute bound scaling is used. The absolute bound scaling criterion is avoiding overflow in all cases.

Contrary to the absolute bound scaling approach, when using infinity bound scaling or L_2 bound scaling, overflow is still possible. Although all scaling methods prevent overflow according to a certain criterion, none of them will force all multiplier coefficients to be smaller than (or equal to) one. This implies that in a practical implementation it can happen that the bits used to represent a multiplier coefficient are not sufficient. The problem can be solved in several ways [19].

The present paper uses a method that has minimum impact on the overall accuracy. More precisely, the multiplier values of the stages where the coefficients are larger than 1 are divided by a power of 2 (using shifting) to force all coefficients of this stage to be smaller than 1. This division by the power of 2 is undone in the output signal of the stage, *i.e.*, by shifting the result.

4. The Ordering of the Stages

As the sequential ordering of the stages has a large impact on the scaling factors and the noise performance of the filter, the round-off output noise variance as a function of the ordering must be calculated and a method to determine the optimal sequential ordering must be found.

4.1. The Round-Off Output Noise Variance

The round-off output noise variance is the summation of the noise sources inside the filter cascade, with suitable weightings and filtering. This variance is affected by the ordering of the stages of the filter structure. The variance of the noise of each rounding operation equals σ_e^2. The input signal of the filter has an amplitude in the interval (−1, +1). In case b + 1 bits (b bits + a sign bit) are used to represent the signal in two's complement, the variance of the noise generated by one round-off noise source is (under the assumptions of Section 2.2) given by:

$$\sigma_e^2 = \frac{q^2}{12}. \tag{17}$$

Here, $q = 2^{-b}$. In general, the variance of the round-off noise source is also given by:

$$\sigma_e^2 = \frac{1}{2\pi} \int_{-\pi}^{+\pi} P_{ee}(\omega_n)\, d\omega_n - |m_e|^2 \tag{18}$$

where: $P_{ee}(\omega_n)$ is the Power Spectral Density (PSD) of the noise source.

m_e is the mean value of the noise source.

Under the assumptions given in Section 2.2, the mean $m_e = 0$ and the PSD of the round-off noise source is independent of the frequency, implying:

$$\sigma_e^2 = P_{ee}(\omega_n). \tag{19}$$

The PSD of the noise generated at the output of the total filter structure, by the noise source of a (scaled and rounded) stage k is given by:

$$P_{vv}(\omega_n) = \left|\overline{G_k\left(e^{j\omega_n}\right)}\right|^2 \sigma_e^2 \tag{20}$$

The variance of this noise is given by:

$$\sigma_v^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{vv}(\omega_n)\, d\omega_n - |m_v|^2. \tag{21}$$

As the noise source has a zero mean value and the stages are linear, $m_v = 0$. The PSD of the noise generated at the output of the filter structure, by all round-off noise sources is given by:

$$P_{nn}(\omega_n) = \sigma_e^2\left[1 + \sum_{k=0}^{P-2}\left|\overline{G_k\left(e^{j\omega_n}\right)}\right|^2\right]. \tag{22}$$

The output noise variance due to the round-off noise sources of all stages is thus given by (using Equation (11)):

$$\sigma_n^2 = \sigma_e^2\left[1 + \sum_{k=0}^{P-2}\left\|\overline{G_k\left(e^{j\omega_n}\right)}\right\|_2^2\right] \tag{23}$$

As can be seen in Equation (23), the contribution of a noise source of any stage k to the output noise variance only, depends on the transfer function from this noise source to the output. As this is valid for a general stage k, it is valid for all stages. From this reasoning, it is obvious to order the stages from high power amplification to low power amplification in order to keep all $\left\|\overline{G_k\left(e^{j\omega_n}\right)}\right\|_2^2$ (for any k) as small as possible. This approach minimizes the output round-off noise variance. However, in Equation (23), the transfer function from a noise source k to the output is scaled, implying that the optimal ordering must be derived from the scaled stage equations, which is rather inconvenient. Therefore, Equation (23) will be further worked out. It is clear from Figure 3 and Equation (7) that Equation (22) can be written as:

$$P_{nn}(\omega_n) = \sigma_e^2\left[1 + \sum_{k=0}^{P-2}\left|\prod_{m=k+1}^{P-1}\left(S_m H_m\left(e^{j\omega_n}\right)\right)\right|^2\right]. \tag{24}$$

Independent of the stage ordering, the following holds for infinity bound and L_2 bound scaling:

$$\prod_{m=k+1}^{P-1} S_m = \frac{\prod_{m=0}^{P-1} S_m}{\prod_{m=0}^{k} S_m} = \frac{\left\|F_k\left(e^{j\omega_n}\right)\right\|_\alpha}{\left\|H\left(e^{j\omega_n}\right)\right\|_\alpha} \tag{25}$$

where, $\alpha = \infty$ or $\alpha = 2$ respectively.

Using Equations (7) and (25) in Equation (24) yields:

$$P_{nn}(\omega_n) = \sigma_e^2 \left[1 + \frac{1}{\left\| H\left(e^{j\omega_n}\right) \right\|_\alpha^2} \sum_{k=0}^{P-2} \left(\left\| F_k\left(e^{j\omega_n}\right) \right\|_\alpha^2 \left| G_k\left(e^{j\omega_n}\right) \right|^2 \right) \right] \tag{26}$$

and by applying Equation (11):

$$\sigma_n^2 = \sigma_e^2 \left[1 + \frac{1}{\left\| H\left(e^{j\omega_n}\right) \right\|_\alpha^2} \sum_{k=0}^{P-2} \left(\left\| F_k\left(e^{j\omega_n}\right) \right\|_\alpha^2 \left\| G_k\left(e^{j\omega_n}\right) \right\|_2^2 \right) \right]. \tag{27}$$

In Equations (26) and (27), the scaling is taken into account and the un-scaled stage equations can now be used in $F_k(e^{j\omega n})$ and $G_k(e^{j\omega n})$ to determine the optimal ordering for minimal output round-off noise variance. It is clear that every $F_k(e^{j\omega n})$ and $G_k(e^{j\omega n})$ contribute to the output noise. For a given filter, the output noise variance is minimal when $\sum_{k=0}^{P-2} \left(\left\| F_k\left(e^{j\omega_n}\right) \right\|_\alpha^2 \left\| G_k\left(e^{j\omega_n}\right) \right\|_2^2 \right)$ is minimal, *i.e.*, an optimal ordering must be found to minimize this sum.

- In case of L_2 bound scaling ($\alpha = 2$), $\left\| F_k\left(e^{j\omega_n}\right) \right\|_2^2$ and $\left\| G_k\left(e^{j\omega_n}\right) \right\|_2^2$ in every sum term will have the same contribution, implying that the ordering of the stages has no importance. (*I.e.*, in case of ordering from $i = 0$ to $i = P - 1$ and using L_2 bound scaling, Equation (27) yields: $\sigma_{n_{L2}}^2 = \sigma_e^2 \left[1 + \frac{1}{\|H\|_2^2} \left(\|H_0\|_2^2 \|H_1 H_2 H_3 ...\|_2^2 + \|H_0 H_1\|_2^2 \|H_2 H_3 ...\|_2^2 + ... \right) \right]$. In case of ordering from $i = P - 1$ to $i = 0$ and using L_2 bound scaling, Equation (27) yields: $\sigma_{n_{L2}}^2 = \sigma_e^2 \left[1 + \frac{1}{\|H\|_2^2} \left(\|H_1 H_2 H_3 ...\|_2^2 \|H_0\|_2^2 + \|H_2 H_3 ...\|_2^2 \|H_0 H_1\|_2^2 + ... \right) \right]$. Two identical equations are obtained.)
- If infinity bound scaling, L_∞, is considered ($\alpha = \infty$), the optimal ordering will be determined by the ratio:

$$\frac{\left\| F_k\left(e^{j\omega_n}\right) \right\|_\infty^2}{\left\| G_k\left(e^{j\omega_n}\right) \right\|_2^2} \tag{28}$$

for every $k \in \{0,1,...,P-2\}$.

In case this ratio is significantly larger than 1 for every k value, it is best to order the stages from small peak gain to large peak gain. In case this ratio is not significantly larger than 1, the optimal ordering should be determined exhaustively.

- In case of absolute bound scaling the equivalent of Equation (25) is given by:

$$\prod_{m=k+1}^{P-1} S_m = \frac{\prod_{m=0}^{P-1} S_m}{\prod_{m=0}^{k} S_m} = \frac{\sum_{n=0}^{p_k} \left| f_k(n) \right|}{\sum_{n=0}^{p} \left| h(n) \right|}. \tag{29}$$

Here $h(n)$ are the impulse response (having a length p) samples of the filter given by the transfer function $H(z)$. Applying Equation (29) on Equation (22) and using Equation (7) yields:

$$P_{nn}(\omega_n) = \sigma_e^2 \left[1 + \frac{1}{\left(\sum_{n=0}^{p} \left| h(n) \right| \right)^2} \sum_{k=0}^{P-2} \left(\left(\sum_{n=0}^{p_k} \left| f_k(n) \right| \right)^2 \left| G_k\left(e^{j\omega_n} \right) \right|^2 \right) \right] \tag{30}$$

and by applying Equation (11):

$$\sigma_n^2 = \sigma_e^2 \left[1 + \frac{1}{\left(\sum_{n=0}^{p} \left| h(n) \right| \right)^2} \sum_{k=0}^{P-2} \left(\left(\sum_{n=0}^{p_k} \left| f_k(n) \right| \right)^2 \left\| G_k\left(e^{j\omega_n} \right) \right\|_2^2 \right) \right]. \tag{31}$$

As for infinity bound scaling, the optimal ordering to obtain a minimal output round-off noise variance is determined by the ratio:

$$\frac{\left(\sum_{n=0}^{p_k} \left| f_k(n) \right| \right)^2}{\left\| G_k\left(e^{j\omega_n} \right) \right\|_2^2}, \tag{32}$$

for every $k \in \{0,1,...,P-2\}$.

In case this ratio is significantly larger than 1 for every k value, it is best to order the stages in increasing coefficient magnitude, *i.e.*, the stage with the largest coefficient(s) at the end. In case the ratio is not significantly larger than 1, the optimal ordering should be determined exhaustively.

4.2. The Signal to Round-off Noise Ratio

The round-off noise performance of a filter may not be correctly evaluated by only analyzing the round-off noise. A more reliable result is obtained by calculating the Signal to round-off noise ratio (SNR). The signal to noise ratios (when using the previously discussed scaling methods) are investigated in this section.

The discrete input signal of the filter is indicated by $x(n)$ and the discrete output signal by $y(n)$. The variance of the discrete input signal $x(n)$ is indicated by σ_{in}^2. The variance of the discrete output signal $y(n)$ of the filter will be indicated by σ_{out}^2. All calculations presented in this section are based on the conditions that the input signal $x(n)$ is zero mean and has a constant, frequency independent, Probability Density Function (PDF).

In case the input signal is a wide sense stationary random signal with uniform PDF and variance σ_{in}^2, the output signal variance of the filter is given by:

$$\sigma_{out}^2 = \frac{\sigma_{in}^2}{2\pi}\int_{-\pi}^{+\pi}\left|\overline{H\left(e^{j\omega_n}\right)}\right|^2 d\omega_n \tag{33}$$

$$\sigma_{out}^2 = \sigma_{in}^2\left\|\overline{H\left(e^{j\omega_n}\right)}\right\|_2^2 \tag{34}$$

$$\sigma_{out}^2 = \sigma_{in}^2\sum_n\left|\overline{h(n)}\right|^2 \tag{35}$$

where $\sum_n\left|\overline{h(n)}\right|^2$ is the sum of the squared impulse response samples of the (scaled and rounded) filter.

- In case of L_2 or L_∞ bound scaling, Equation (34) can be written as (using Equation (10) or Equation (13) as appropriate):

$$\sigma_{out}^2 = \sigma_{in}^2\frac{\left\|H\left(e^{j\omega_n}\right)\right\|_2^2}{\left\|H\left(e^{j\omega_n}\right)\right\|_\alpha^2}, \tag{36}$$

where $\alpha = 2$ or $\alpha = \infty$, respectively. Combining Equations (36) and (27) yields.

$$SNR = \frac{\sigma_{out}^2}{\sigma_n^2} = \frac{\sigma_{in}^2}{\sigma_e^2}\frac{\left\|H\left(e^{j\omega_n}\right)\right\|_2^2}{\left[\left\|H\left(e^{j\omega_n}\right)\right\|_\alpha^2 + \sum_{k=0}^{P-2}\left(\left\|F_k\left(e^{j\omega_n}\right)\right\|_\alpha^2\left\|G_k\left(e^{j\omega_n}\right)\right\|_2^2\right)\right]}. \tag{37}$$

- In case of absolute bound scaling, the SNR is given by:

$$SNR = \frac{\sigma_{out}^2}{\sigma_n^2} = \frac{\sigma_{in}^2}{\sigma_e^2}\frac{\left\|H\left(e^{j\omega_n}\right)\right\|_2^2}{\left[\left(\sum_{n=0}^{p}\left|h(n)\right|\right)^2 + \sum_{k=0}^{P-2}\left(\left(\sum_{n=0}^{p_k}\left|f_k(n)\right|\right)^2\left\|G_k\left(e^{j\omega_n}\right)\right\|_2^2\right)\right]}. \tag{38}$$

In Equations (37) and (38) σ_e^2 is dependent on the type of rounding used and the number of bits that are used to quantize the signal in the filter structure. σ_{in}^2 depends on the input signal. Quantization and input signal independent factors are given by:

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{L2} = \frac{1}{\left[1+\frac{\sum_{k=0}^{P-2}\left(\left\|F_k\left(e^{j\omega_n}\right)\right\|_2^2\left\|G_k\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|H\left(e^{j\omega_n}\right)\right\|_2^2}\right]} \tag{39}$$

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{L\infty} = \frac{1}{\left[\frac{\left\|H\left(e^{j\omega_n}\right)\right\|_\infty^2}{\left\|H\left(e^{j\omega_n}\right)\right\|_2^2} + \frac{\sum_{k=0}^{P-2}\left(\left\|F_k\left(e^{j\omega_n}\right)\right\|_\infty^2 \left\|G_k\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|H\left(e^{j\omega_n}\right)\right\|_2^2}\right]}. \tag{40}$$

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{abs} = \frac{1}{\left[\frac{\left(\sum_{n=0}^{p}\left|h(n)\right|\right)^2}{\left\|H\left(e^{j\omega_n}\right)\right\|_2^2} + \frac{\sum_{k=0}^{P-2}\left(\left(\sum_{n=0}^{p_k}\left|f_k(n)\right|\right)^2 \left\|G_k\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|H\left(e^{j\omega_n}\right)\right\|_2^2}\right]}. \tag{41}$$

In general for an arbitrary $A(\omega)$ [8]:

$$\|A\|_1 \le \|A\|_2 \le \ldots \le \|A\|_\infty \le \sum_n |a(n)|, \tag{42}$$

which implies that:

$$\left\|H\left(e^{j\omega_n}\right)\right\|_2^2 \le \left\|H\left(e^{j\omega_n}\right)\right\|_\infty^2 \le \left(\sum_{n=0}^{p}\left|h(n)\right|\right)^2. \tag{43}$$

Note that Equation (43) is independent of the ordering of the stages.

As mentioned in Section 4.1, in case of L_2 bound scaling, $\sum_{k=0}^{P-2}\left(\left\|F_k\left(e^{j\omega_n}\right)\right\|_2^2 \left\|G_k\left(e^{j\omega_n}\right)\right\|_2^2\right)$ in Equation (39) is independent of the stage ordering, implying that the stage ordering has no impact on the SNR. For a given ordering of the stages, it is clear from Equations (39–41) and (43) that

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{L2} \ge \left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{L\infty} \ge \left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{abs}. \tag{44}$$

Even in case of an optimized ordering for infinity bound scaling or absolute bound scaling, L_2 bound scaling will always have the best SNR. Note that SNR in this text is signal to round-off noise ratio and not the overall SNR of the filter. However, an optimized ordering for absolute bound scaling can have a better SNR than a non-optimized ordering for infinity bound scaling and vice versa. In case of L_∞ bound scaling, the optimal SNR must be determined by finding an ordering that minimizes

$$\sum_{k=0}^{P-2}\left(\left\|F_k\left(e^{j\omega_n}\right)\right\|_\infty^2 \left\|G_k\left(e^{j\omega_n}\right)\right\|_2^2\right). \tag{45}$$

In case of absolute bound scaling the optimal SNR must be determined by finding an ordering that minimizes

$$\sum_{k=0}^{P-2}\left(\left(\sum_{n=0}^{p_k}\left|f_k(n)\right|\right)^2 \left\|G_k\left(e^{j\omega_n}\right)\right\|_2^2\right). \tag{46}$$

Comparing these requirements with those of the output noise variance (formulated in Section 4.1), it can be concluded that the optimization of the SNR, by finding the optimal ordering, uses the same criteria as the minimization of the output round-off noise as discussed in Section 4.1.

Notice that the equations derived in this section are generally valid for any filter cascade (with one noise source at the output of each filter stage) and not only for MFIR structures.

5. The SNR of MFIR Filters

In this section, the general theory developed in Section 4 will be applied to the MFIR filter structures.

5.1. The Transfer Functions

In case *MFIR* filters are considered, the stages are indicated by $M_i(z)$ or $M_i(e^{j\omega n})$ or in short form M_i. The total filter transfer function is written as $M(z)$, $M(e^{j\omega n})$ or M. A scaled stage is indicated by $\overline{M_i(z)}$, $\overline{M_i(e^{j\omega_n})}$ or $\overline{M_i}$.

The ordering where the first stage is stage $i = 0$, the second is $i = 1$ and so on, will be called the forward (sequential) ordering. Consequently, in case of MFIR structures and forward ordering:

$$G_i\left(e^{j\omega_n}\right)=\prod_{m=i+1}^{P-1}M_m\left(e^{j\omega_n}\right), \tag{47}$$

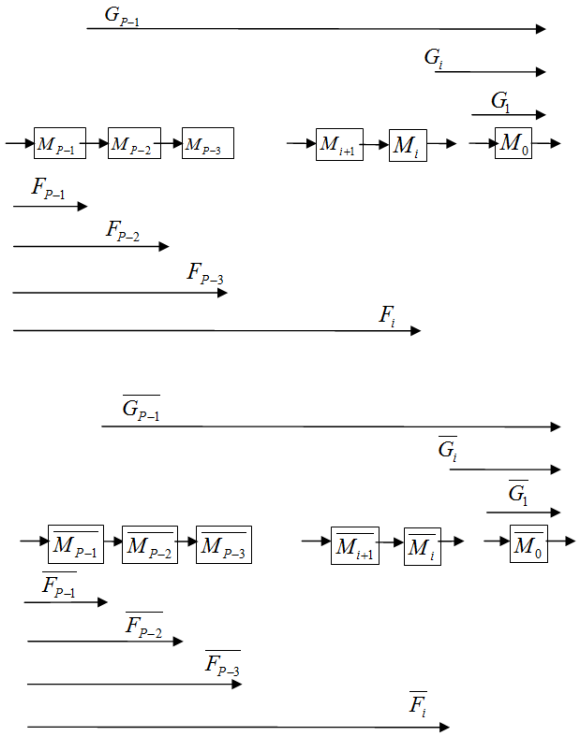
$$\overline{G_i\left(e^{j\omega_n}\right)}=\prod_{m=i+1}^{P-1}\left(\overline{M_m\left(e^{j\omega_n}\right)}\right)=\prod_{m=i+1}^{P-1}\left(S_mM_m\left(e^{j\omega_n}\right)\right)=\left(\prod_{m=i+1}^{P-1}\left(S_m\right)\right)G_i\left(e^{j\omega_n}\right), \tag{48}$$

$$F_i\left(e^{j\omega_n}\right)=\prod_{m=0}^{i}M_m\left(e^{j\omega_n}\right), \tag{49}$$

$$\overline{F_i\left(e^{j\omega_n}\right)}=\prod_{m=0}^{i}\left(\overline{M_m\left(e^{j\omega_n}\right)}\right)=\prod_{m=0}^{i}\left(S_mM_m\left(e^{j\omega_n}\right)\right)=\left(\prod_{m=0}^{i}\left(S_m\right)\right)F_i\left(e^{j\omega_n}\right). \tag{50}$$

Notice that $F_{P-1}(e^{j\omega_n}) = M(e^{j\omega_n})$ and $G_{P-1}(e^{j\omega_n}) = 1$ (see Figures 2 and 3).

Figure 5. The unscaled and scaled MFIR filter cascade transfer functions for reverse ordering.



The ordering where stage $i = P - 1$ is the first stage, $i = P - 2$ is the second stage and so on, is called the reverse (sequential) ordering. In case of reverse ordering of MFIR stages, the transfer functions are shown in Figure 5 and are defined by:

$$G_i\left(e^{j\omega_n}\right) = \prod_{m=0}^{i-1} M_m\left(e^{j\omega_n}\right), \quad i > 0 \tag{51}$$

$$\overline{G_i\left(e^{j\omega_n}\right)} = \prod_{m=0}^{i-1}\left(\overline{M_m\left(e^{j\omega_n}\right)}\right) = \prod_{m=0}^{i-1}\left(S_m M_m\left(e^{j\omega_n}\right)\right) = \left(\prod_{m=0}^{i-1}(S_m)\right) G_i\left(e^{j\omega_n}\right), \qquad i > 0 \tag{52}$$

$$F_i\left(e^{j\omega_n}\right) = \prod_{m=i}^{P-1} M_m\left(e^{j\omega_n}\right), \tag{53}$$

$$\overline{F_i\left(e^{j\omega_n}\right)} = \prod_{m=i}^{P-1}\left(\overline{M_m\left(e^{j\omega_n}\right)}\right) = \prod_{m=i}^{P-1}\left(S_m M_m\left(e^{j\omega_n}\right)\right) = \left(\prod_{m=i}^{P-1}(S_m)\right) F_i\left(e^{j\omega_n}\right). \tag{54}$$

Notice that in case of reverse ordering $F_0(e^{j\omega n}) = M(e^{j\omega n})$ and $G_0(e^{j\omega n}) = 1$.

Although there are $P!$ Possible ordering combinations of the MFIR stages, experiments have shown that choosing the best option between forward and reverse sequential ordering is usually satisfactory.

5.2. The Scaling Factors

The equations derived in Section 3 adapted to MFIR structures are given in Table 1.

Table 1. Scaling factors for Multiplicative Finite Impulse Response (MFIR) structures.

	Forward ordering	**Reverse ordering**
L_2 bound	$\prod_{m=0}^{i} S_m = \left(\left\|F_i\left(e^{j\omega_n}\right)\right\|_2\right)^{-1}$	$\prod_{m=i}^{P-1} S_m = \left(\left\|F_i\left(e^{j\omega_n}\right)\right\|_2\right)^{-1}$
Infinity bound	$\prod_{m=0}^{i} S_m = \left(\sup_{-\pi<\omega_n<+\pi}\left(\left\|F_i\left(e^{j\omega_n}\right)\right\|\right)\right)^{-1}$	$\prod_{m=i}^{P-1} S_m = \left(\sup_{-\pi<\omega_n<+\pi}\left(\left\|F_i\left(e^{j\omega_n}\right)\right\|\right)\right)^{-1}$
Absolute bound	$\prod_{m=0}^{i} S_m = \left(\sum_{n=0}^{p_i}\left\|f_i(n)\right\|\right)^{-1}$	$\prod_{m=i}^{P-1} S_m = \left(\sum_{n=0}^{p_i}\left\|f_i(n)\right\|\right)^{-1}$

Table 2. Length of the impulse responses of the partial MFIR transfer functions.

Forward ordering	$f_i(n)$ or $\overline{f_i(n)}$
Real pole	$0 \rightarrow p_i = 2^{i+1} - 1$
Real pole linear phase	$0 \rightarrow p_i = 2^{i+2} - 2$
Complex-conjugate pole pair	$0 \rightarrow p_i = 2^{i+2} - 2$
Complex-conjugate pole pair linear phase	$0 \rightarrow p_i = 2^{i+3} - 4$
Reverse ordering	$f_i(n)$ or $\overline{f_i(n)}$
Real pole	$0 \rightarrow p_i = 2^{P} - 2^{i}$
Real pole linear phase	$0 \rightarrow p_i = 2^{P+1} - 2^{i+1}$
Complex-conjugate pole pair	$0 \rightarrow p_i = 2^{P+1} - 2^{i+1}$
Complex-conjugate pole pair linear phase	$0 \rightarrow p_i = 2^{P+2} - 2^{i+2}$

In case of absolute bound scaling, the length p_i of the impulse response $f_i(n)$ must be known. As this length is filter dependent, Table 2 gives an overview of the impulse response lengths for forward and reverse ordering of MFIR stages [2,19].

It is shown in [19] that in case of a real pole approximation, the scaling factors only depend on the stage on which they are applied, *i.e.*, not on any of the previous or next stages.

5.3. The SNR and Optimal Ordering for MFIR Filters

5.3.1. General MFIR SNR Expressions

In case of forward ordering, Equations (39–41) become:

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{L2} = \frac{1}{\left[1+\frac{\sum_{i=0}^{P-2}\left(\left\|F_i\left(e^{j\omega_n}\right)\right\|_2^2\left\|G_i\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}\right]}, \tag{55}$$

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{L\infty} = \frac{1}{\left[\frac{\left\|M\left(e^{j\omega_n}\right)\right\|_\infty^2}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}+\frac{\sum_{i=0}^{P-2}\left(\left\|F_i\left(e^{j\omega_n}\right)\right\|_\infty^2\left\|G_i\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}\right]}, \tag{56}$$

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{abs} = \frac{1}{\left[\frac{\left(\sum_{n=0}^{p}\left|m\left(n\right)\right|\right)^2}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}+\frac{\sum_{i=0}^{P-2}\left(\left(\sum_{n=0}^{p_i}\left|f_i\left(n\right)\right|\right)^2\left\|G_i\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}\right]}. \tag{57}$$

Here, $G_i\left(e^{j\omega_n}\right)$ and $F_i\left(e^{j\omega_n}\right)$ are given by Equations (47) and (49) respectively. The value of p_i is given in Table 2 for forward ordering. $m(n)$ are the impulse response samples of the total MFIR filter. The value p in $\left(\sum_{n=0}^{p}\left|m\left(n\right)\right|\right)^2$ in Equation (57) is the length of the total MFIR filter impulse response and can be found by setting $i = P - 1$ in Table 2 for forward ordering.

In case of reverse ordering, Equations (39–41) become:

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{L2} = \frac{1}{\left[1+\frac{\sum_{i=1}^{P-1}\left(\left\|F_i\left(e^{j\omega_n}\right)\right\|_2^2\left\|G_i\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}\right]}, \tag{58}$$

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{L\infty} = \frac{1}{\left[\frac{\left\|M\left(e^{j\omega_n}\right)\right\|_\infty^2}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}+\frac{\sum_{i=1}^{P-1}\left(\left\|F_i\left(e^{j\omega_n}\right)\right\|_\infty^2\left\|G_i\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}\right]}, \tag{59}$$

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}.SNR\right)_{abs} = \frac{1}{\left[\frac{\left(\sum_{n=0}^{p}\left|m(n)\right|\right)^2}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2} + \frac{\sum_{i=1}^{P-1}\left(\left(\sum_{n=0}^{p_i}\left|f_i(n)\right|\right)^2\left\|G_i\left(e^{j\omega_n}\right)\right\|_2^2\right)}{\left\|M\left(e^{j\omega_n}\right)\right\|_2^2}\right]}. \quad (60)$$

Here, $G_i\left(e^{j\omega_n}\right)$ and $F_i\left(e^{j\omega_n}\right)$ are given by Equations (50) and (53) respectively. The value of p_i is given in Table 2 for reverse ordering. The value p in $\left(\sum_{n=0}^{p}\left|m(n)\right|\right)^2$ in Equation (60) is the length of the total MFIR filter impulse response and can be found by setting $i = 0$ in Table 2 for reverse ordering.

In order to keep the text more readable, the figures of the several calculation results of Equations (55–60) are grouped together at the end of the section. Table 3 gives an overview of the results.

Table 3. Overview of the figures of the SNR calculations.

Approximation of	Scaling type	Ordering	Range	Figure
Real pole	L_2	Forward and Reverse	$\|\lambda\| \in [0.1,\ 1)$	Figure 6
Real Pole	Abs and Inf.	Forward and Reverse	$\|\lambda\| \in [0.1,\ 1)$	Figure 7
Real Pole Linear Phase	L_2	Forward and Reverse	$\|\lambda\| \in [0.1,\ 1)$	Figure 8
Real Pole Linear Phase	Abs and Inf.	Forward and Reverse	$\|\lambda\| \in [0.1,\ 1)$	Figure 9
Compl. Conj. Pole pair	L_2	Forward and Reverse	$r = 0.9$ $\theta \in [0,\ \pi)$	Figure 10
Compl. Conj. Pole pair	Inf	Forward and Reverse	$r = 0.9$ $\theta \in [0,\ \pi)$	Figure 11
Compl. Conj. Pole pair	Abs	Forward and Reverse	$r = 0.9$ $\theta \in [0,\ \pi)$	Figure 12
Compl. Conj. Pole pair	Inf	Reverse	$r = 0.8; 0.85; 0.9; 0.95$ $\theta \in [0,\ \pi)$	Figure 13
Compl. Conj. Pole pair	L_2, Inf, Abs	Forward and Reverse	$r = 0.9$ $\theta \in [0,\ \pi)$	Figure 14
Compl. Conj. Pole pair, Lin Phase	L_2, Inf, Abs	Forward	$r = 0.9$ $\theta \in [0,\ \pi)$	Figure 15
Compl. Conj. Pole pair, Lin Phase	L_2, Inf, Abs	Reverse	$r = 0.9$ $\theta \in [0,\ \pi)$	Figure 16
Compl. Conj. Pole pair, Lin Phase	Inf	Reverse	$r = 0.8; 0.85; 0.9; 0.95$ $\theta \in [0,\ \pi)$	Figure 17

5.3.2. SNR Performance of an MFIR Filter Approximating a Real Pole Filter

Figures 6 and 7 show the $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ values for real poles $|\lambda|$ in the interval [0.1, 1). In case of the MFIR approximation, for each $|\lambda|$ value, the required number of stages, P, is determined to obtain a maximum difference of |0.01| dB between the MFIR magnitude response and the magnitude response of the approximated IIR filter. The edges in the curves indicate where an extra MFIR stage is added to fulfill this requirement. The IIR filter results are based on ([10], Equation 12.148):

$$\left(\frac{\sigma_e^2}{\sigma_{in}^2}\right)SNR = S^2 \tag{61}$$

Here, S is the scaling factor of the IIR filter.

Figure 6. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ for real poles $|\lambda|$ in the interval [0.1, 1) in case of L_2 bound scaling. (The MFIR forward and MFIR reverse results are superimposed).

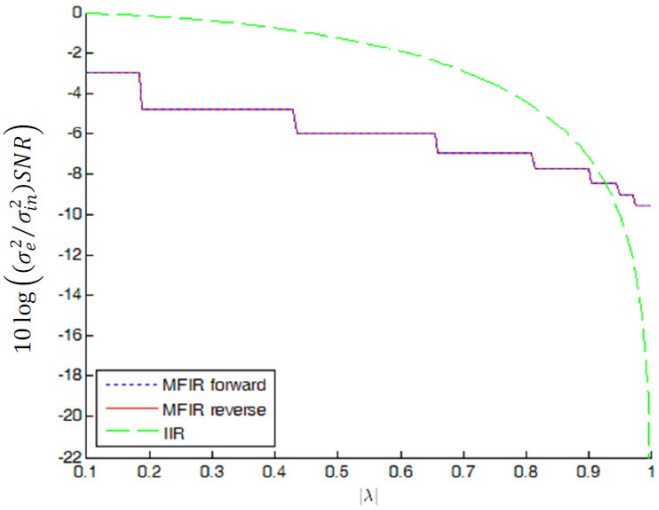


Figure 7. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ for real poles $|\lambda|$ in the interval [0.1, 1) in case of absolute or infinity bound scaling.

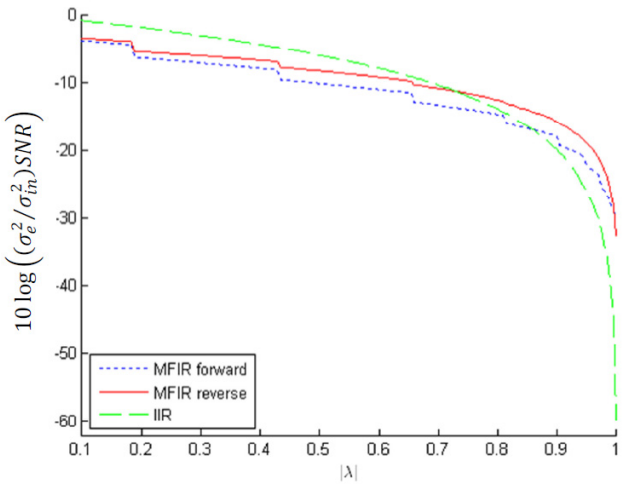


Figure 8. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ of MFIR filters approximating the squared magnitude response of real poles $|\lambda|$ in the interval [0.1, 1) in case of L_2 bound scaling. (The MFIR forward and MFIR reverse results are superimposed).

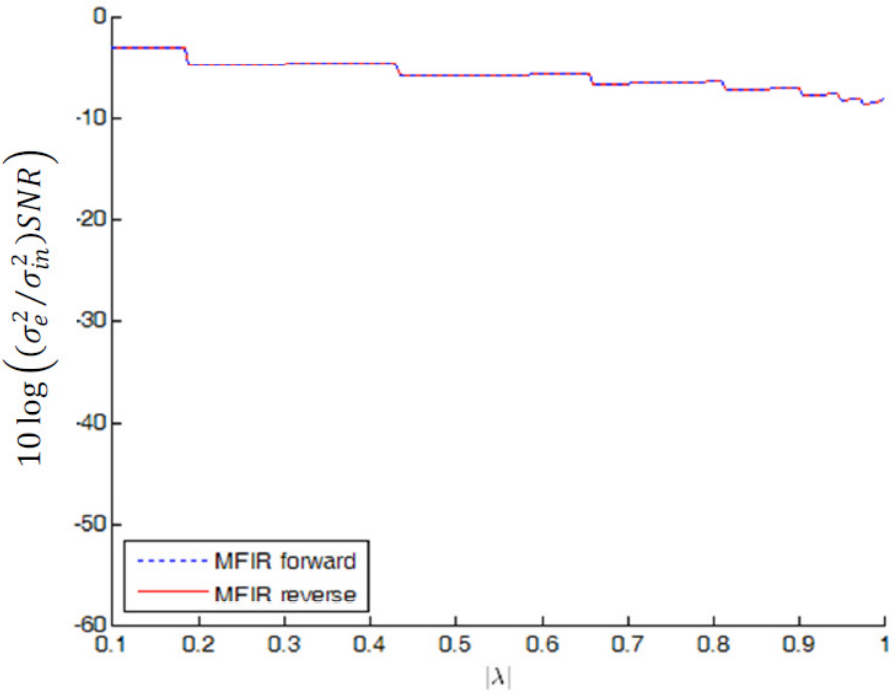


Figure 9. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ of MFIR filters approximating the squared magnitude response of real poles $|\lambda|$ in the interval [0.1, 1) in case of absolute and/ or infinity bound scaling.

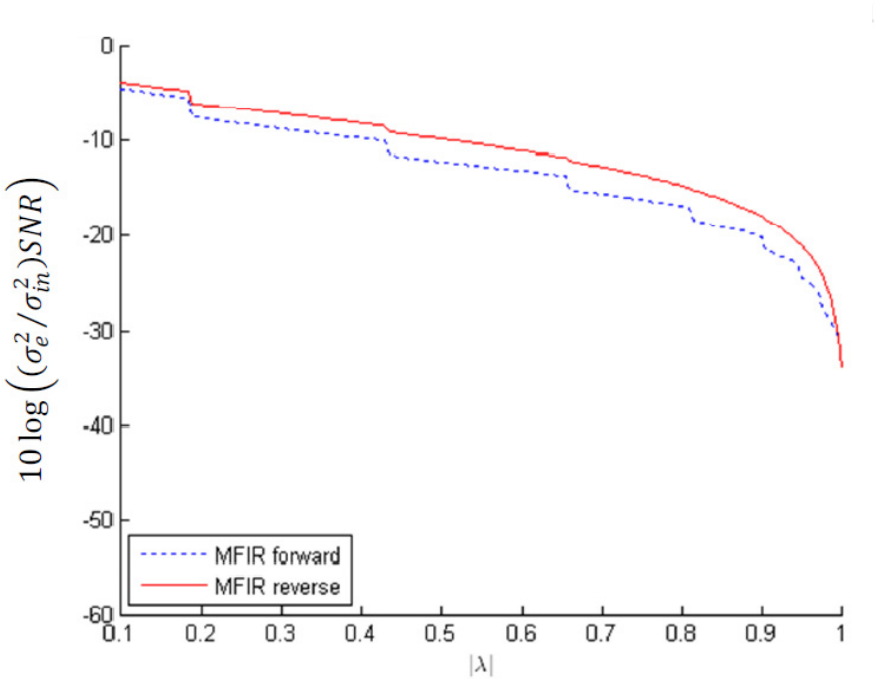


Figure 10. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ for complex-conjugate pole pairs with magnitude 0.9 and L_2 bound scaling; The MFIR approximation uses $P = 7$ stages. (The MFIR forward and MFIR reverse results are superimposed).

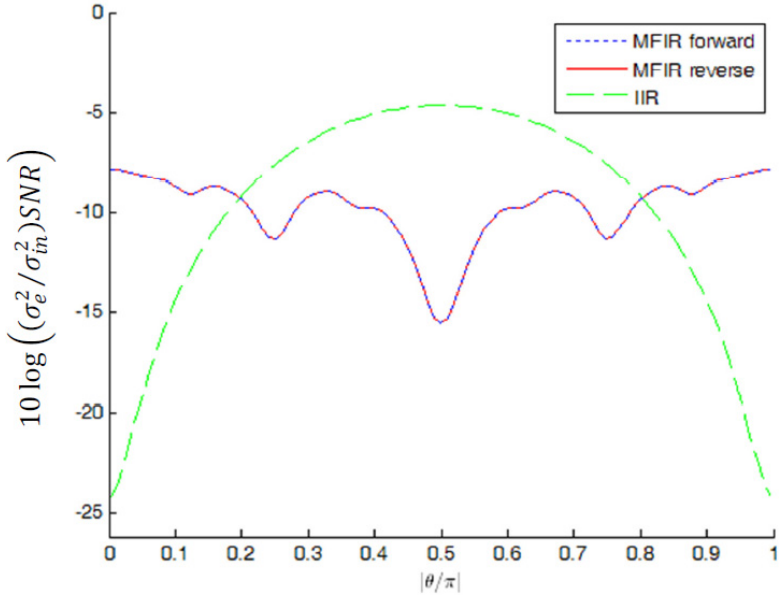


Figure 11. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ for complex-conjugate pole pairs with magnitude 0.9 and infinity bound scaling; The MFIR approximation uses $P = 7$ stages.

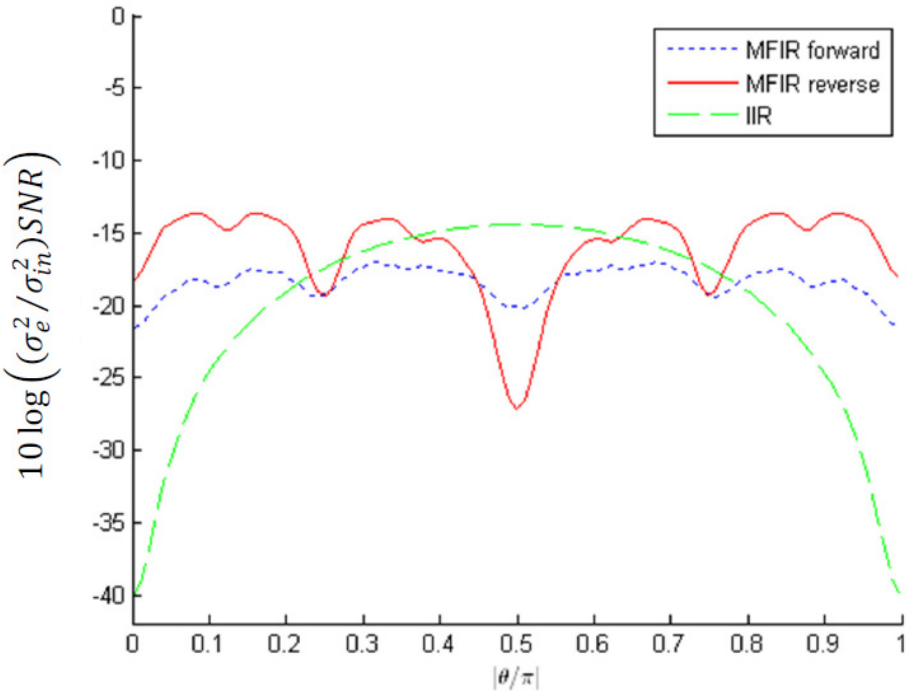


Figure 12. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ for complex-conjugate pole pairs with magnitude 0.9 and absolute bound scaling; The MFIR approximation uses $P = 7$ stages.

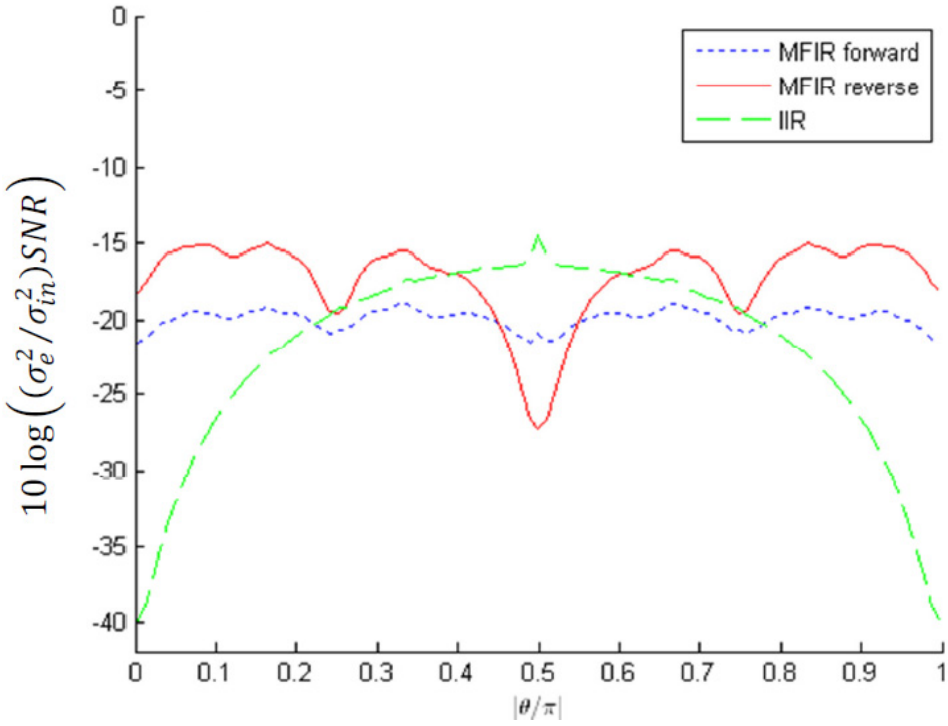


Figure 13. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ in function of pole magnitude r and pole angle θ, in case of infinity bound scaling and reverse ordering. (surface = MFIR; discrete curves = IIR filter).

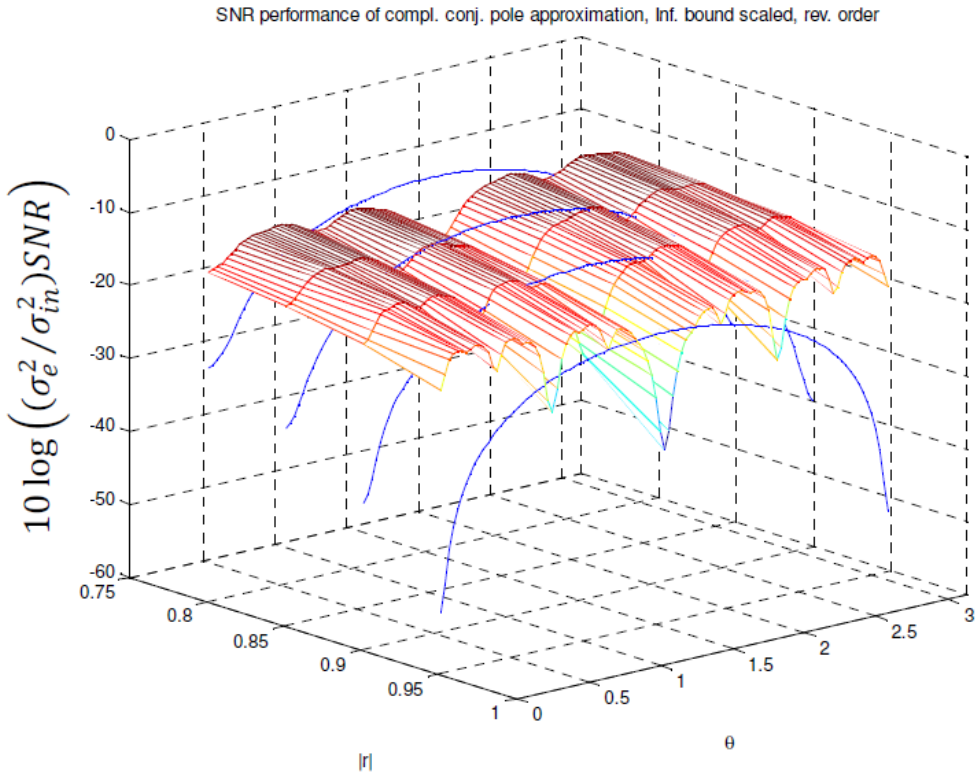


Figure 14. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ of an MFIR filter approximating a complex-conjugate pole pair with magnitude 0.9, and $P = 7$.

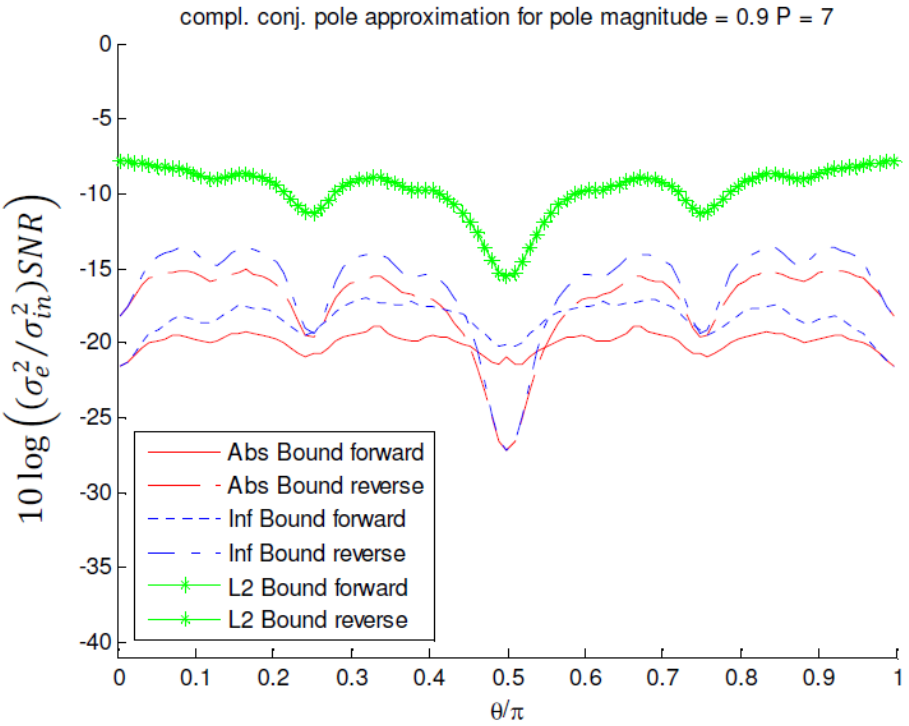


Figure 15. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ of a linear phase MFIR filter approximating the squared magnitude response of a complex-conjugate pole pair with magnitude 0.9 and $P = 7$ in forward ordering.

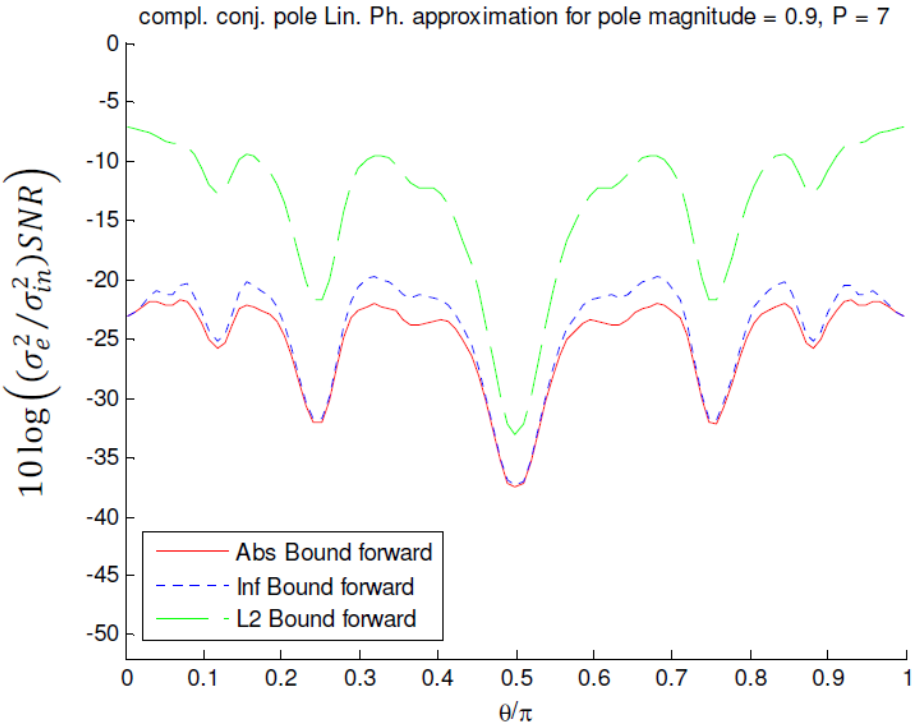


Figure 16. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ of a linear phase MFIR filter approximating the squared magnitude response of a complex-conjugate pole pair with magnitude 0.9 and P = 7 in reverse ordering.

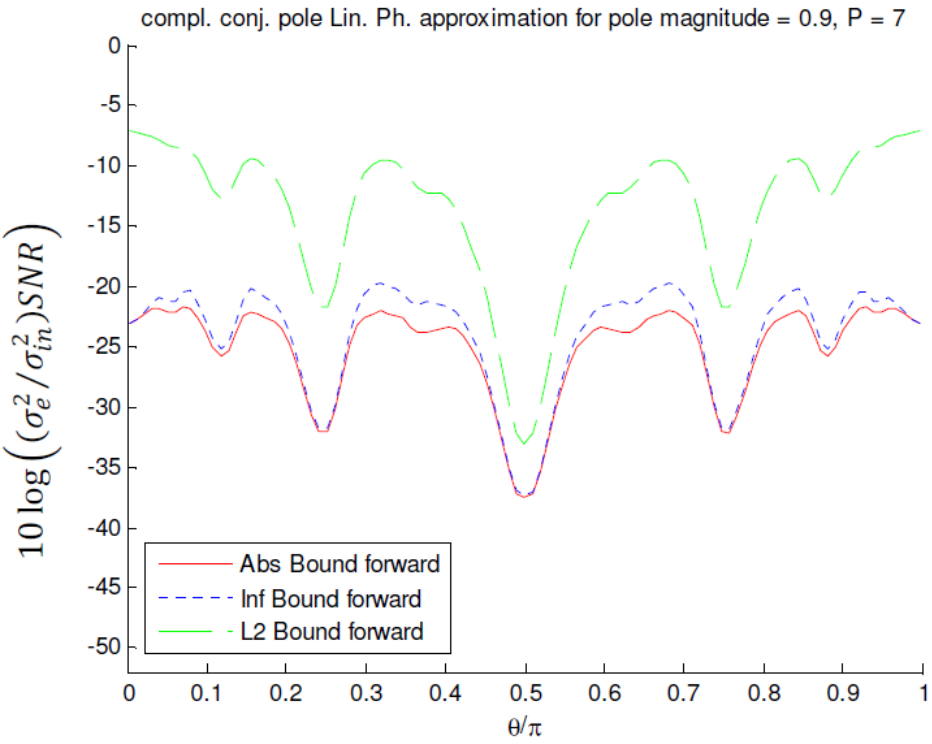
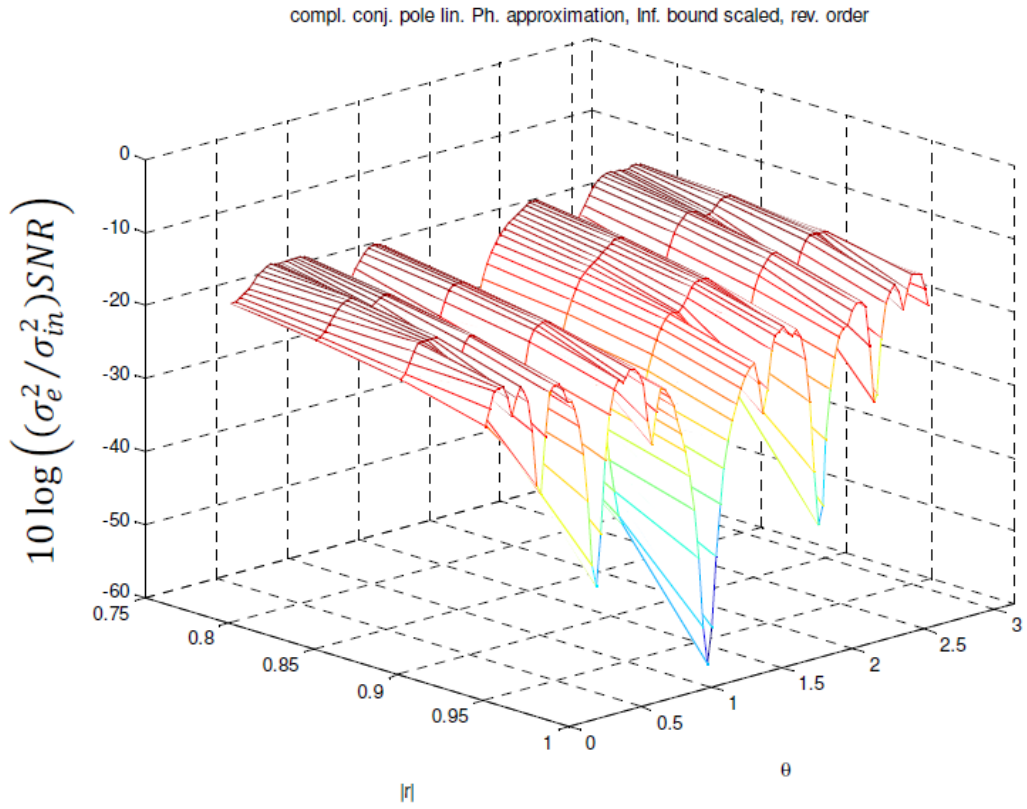


Figure 17. $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ of a linear phase MFIR filter in function of the pole angle and magnitude in case of infinity bound scaling and reverse ordering.



As explained in Section 4.2, the ordering has no impact on the noise performance when L_2 bound scaling is used. Absolute bound scaling and infinity bound scaling yield the same noise performance and reverse ordering is better (max. 4 dB) than forward ordering. L_2 bound scaling always yields a better noise performance than absolute and infinity bound scaling.

In case of absolute or infinity bound scaling, the MFIR filter has a better signal to noise performance than the approximated real pole filter for $|\lambda| \geq 0.861$. For L_2 bound scaling, only for $|\lambda| \geq 0.925$ the MFIR filter has a better performance than the approximated IIR filter.

In case $|\lambda| < 0.9$, the worst-case differences between the MFIR filter and the IIR filter are 3.6 dB for absolute (and infinity) bound scaling and 5 dB for L_2 bound scaling. This is, however, not dramatic because in the range of interest ($|\lambda| > 0.9$), in general the MFIR filter has a better signal to noise performance than its corresponding IIR filter (except for the small region between $|\lambda| = 0.9$ and $|\lambda| = 0.925$ for L_2 bound scaling where the difference is maximum 1.2 dB in favor of the IIR filter).

The more $|\lambda|$ approaches the unit circle, the better the noise performance of the MFIR filter compared to the IIR filter. For example, for $|\lambda| = 0.999$, the MFIR approximation is 27 dB better for absolute (and infinity) bound scaling, for L_2 bound scaling the MFIR filter is 17 dB better than the approximated IIR filter.

5.3.3. SNR Performance of a Linear Phase MFIR Filter Approximating the Squared Magnitude Response of a Real Pole Filter

Figures 8 and 9 show the SNR performance of linear phase MFIR filters approximating the squared magnitude response of real pole filters with $|\lambda|$ in the interval [0.1, 1). For each $|\lambda|$ value, the number of stages, P, is kept the same as in Section 5.3.2.

There is no objective comparison possible between this MFIR approximation and the IIR filter, since the IIR filter is not a linear phase filter and the MFIR filter approximates the *squared* magnitude response of the pole indicated on the horizontal axis. However, it is clear that the round-off noise performance is comparable with the real pole non-linear phase MFIR approximations.

The ordering has no impact when L_2 bound scaling is used and the noise performance with L_2 bound scaling is always better than absolute and infinity bound scaling. Infinity bound scaling and absolute bound scaling have the same performance. Reverse ordering yields again better results than forward ordering. However, the maximum difference between the two orderings is rather small (3.3 dB).

Compared to the non-linear phase approximation, the difference between L_2 bound scaling and the other scaling methods are somewhat larger.

5.3.4. SNR Performance of an MFIR Filter Approximating a Complex-Conjugate Pole Pair Filter in Cascade

Figure 10 shows the $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ values when realizing pole pairs with a magnitude $r = 0.9$ and angles θ in the interval $[0, \pi)$ in case of L_2 bound scaling. Figures 11 and 12 show the results in case of infinity bound scaling and absolute bound scaling respectively. In Figure 13 the $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ values are calculated for the pole magnitudes: 0.8, 0.85, 0.9 and 0.95 approximated with P = 5, 6, 7, and 7

stages respectively in case of infinity bound scaling and reverse ordering. The surface plots are the $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ values for the MFIR filters. The half circle shaped curves are the $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ values for the corresponding IIR filters. Figure 14 compares the noise performances of the approximation of the complex-conjugate pole pair with a magnitude r = 0.9 and any angle between 0 and π, for several scaling methods and orderings.

All pole magnitudes between |0.8| and |0.99| in steps of 0.01 have been calculated but are not all shown here. The results shown in the figures are however representative for all combinations of scaling methods, pole magnitudes and angles that were calculated.

After extensive analysis of the data, the following conclusions can be drawn for the MFIR approximation of a complex-conjugate pole pair filter realized using the cascade structure. The round-off noise performance of an MFIR filter approximating a complex-conjugate pole pair filter

- is significantly better (up to 20 dB) than the noise performance of its corresponding IIR filter when the approximated poles are situated in the neighborhood of the real axis;
- is far less pole angle θ dependent in comparison with the corresponding IIR filter;
- is up to 2.5 dB better for infinity bound scaling than for absolute bound scaling (using reverse ordering);
- is always better for L_2 bound scaling than for the other scaling methods (obeys Equation (44));
- is pole magnitude dependent, but not that much as the corresponding IIR filter;
- is fairly insensitive to an extra MFIR filter stage (typically 1 dB);
- is very sensitive to the stage ordering for absolute and infinity bound scaling;
- is ordering independent in case of L_2 bound scaling;
- is in general better in reverse ordering than in forward ordering, except for pole angles in the neighborhood of $\pi/2$;
- is for pole angles in the neighborhood of $\pi/2$, for absolute and infinity bound scaling, better in forward ordering than in reverse ordering (It was already remarked in Section 4.2, from a theoretical point of view, that this situation could occur.);
- can be up to 6 dB worse than the corresponding IIR filter for pole angles in the neighborhood of $\pi/2$. The width of this region and the magnitude of the difference decreases however with increasing pole magnitude ;

In the normal range of pole magnitudes that are considered for MFIR approximations ($|r| > 0.9$), the noise performance is in general (depending on the pole angle) better than for the corresponding IIR filter.

5.3.5. SNR Performance of a Linear Phase MFIR filter Approximating the Squared Magnitude Response of a Complex-Conjugate Pole Pair Filter

Figure 15 shows the $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ values for pole pairs with magnitude r = 0.9 and angles θ in the interval [0, π) for forward ordering. Figure 16 shows the values for reverse ordering. In Figure 17, the $\left(\sigma_e^2/\sigma_{in}^2\right)SNR$ in case of reverse ordering and infinity bound scaling for the pole magnitudes 0.8, 0.85, 0.9 and 0.95 approximated with P = 5, 6, 7 and 7 stages respectively, is shown.

Compared with the non-linear phase approximation, the $\left(\sigma_e^2/\sigma_{in}^2\right)_{SNR}$ is more pole angle dependent. The dips at the pole angles $\pi/4$, $\pi/2$ and $3\pi/4$ are also deeper. The figures show that the ordering of the stages has no effect when L_2 bound scaling is used. In case of absolute or infinity bound scaling, reverse ordering performs better than forward ordering except in the regions where $\theta = \pi/4$, $\pi/2$, and $3\pi/4$. However, the width of these regions is pole magnitude dependent. Consequently, in practice both orderings will have to be considered when poles with angles in these regions are to be approximated.

Unfortunately in the neighborhood of the pole angles $\theta = \pi/4$, $\pi/2$ and $3\pi/4$ the linear phase approximation clearly performs worse than the non-linear phase approximation. The non-optimal performance for pole angles $\theta = \pi/4$, $\pi/2$ and $3\pi/4$ can be explained by using an example. In case $r = 0.9$, $\theta = \pi/2$ and $P = 9$, $M_0(z)$ is given by (using Equation (5)):

$$M_0(z) = 1 + 2.045z^{-2} + z^{-4} \tag{62}$$

and $M_8(z)$ is given by:

$$M_8(z) = 1 + 1.035*10^{12}z^{-256} + 2.678*10^{23}z^{-512} + 1.035*10^{12}z^{-768} + z^{-1024}. \tag{63}$$

In Section 4.1 it is shown that in case of infinity bound scaling, the stages with the largest peak gains (in the frequency domain), should fall most often in $G_k(z)$ to reduce the output round-off noise variance. In case of absolute bound scaling, the stages with the largest coefficients should fall most often in $G_k(z)$ to reduce the output round-off noise variance.

It is clear from Equations (62) and (63) that forward ordering ($M_8(z)$ most often in $G_k(z)$) is in this case much better than reverse ordering. There is a combined effect involved in the linear phase approximation of complex-conjugate pole pairs with angles in the neighborhood of $\pi/4$, $\pi/2$ and $3\pi/4$:

- $1/r^{2^i}$ factors in Equation (5) can have very large values when i is large,
- $\cos\left(2^i\theta\right)$ factors in Equation (5) are for most stages close to unity implying the coefficients are not reduced by the cosine functions.

Even in case of forward ordering, the SNR performance for these angles is not good. Indeed the stages with larger i values, still have very large coefficients implying that Equation (45) or Equation (46) will never be very small.

6. Conclusions

An approach to model round-off noise in general cascade filter structures has been studied. This round-off noise depends on the used scaling method and on the ordering of the stages. These general results are used to study and optimize the round-off noise behavior of MFIR filters.

It has been shown that the round-off noise performances of the MFIR pole approximations indeed depend on the used scaling method. L_2 bound scaling results in the best performance, followed by infinity and absolute bound scaling.

In general, it can be concluded that in the region of interest (approximating pole behaviors with magnitudes $r > 0.9$) the MFIR approximations perform better than the approximated IIR filters. Even

outside the region of interest, the performance generally does not differ too much from the corresponding IIR filters (max 6 dB).

The analysis presented in [1] suggests forward ordering as an optimal ordering in case no scaling is applied. In case of L_2 bound scaling the ordering has no impact on the round-off noise performance. The analysis presented here, extends these results to other practical scaling methods and concludes that in these cases, reverse ordering performs better than forward ordering for most pole approximations. However, it should be noted that special attention to the stage ordering is required

- when approximating a complex-conjugate pole pair having a pole angle in the neighborhood of $\pi/2$ and the cascade structure has been used;
- when approximating the squared magnitude response of a complex-conjugate pole pair filter in case the pole angles are situated in the neighborhood of $\pi/4$, $\pi/2$ and $3\pi/4$ and the linear phase cascade structure has been used.

Further research is required to determine if alternative orderings can be found which would yield a better noise performance. Note that the orderings considered here are not the only possible orderings. For best performance, one must determine the "declining amplification" order for every pole pair that is approximated. At the moment, no systematic approach has been found, so trial and error is required. Research will have to prove if an optimal ordering can be calculated. If not, a heuristic approach as in [8] for the pole zero pairing or an iterative optimization algorithm [17] or another near optimal ordering technique [16] could also be interesting.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Fam, A.T. MFIR filters: Properties and applications. *IEEE Trans. Acoust. Speech Signal Process*. **1981**, *29*, 1128–1136.
2. Vandenbussche, J.-J.; Lee, P.; Peuteman, J. Analysis of time and frequency domain performance of MFIR filters. In Proceedings of the 2008 International Conference on Embedded Systems and Applications, Las Vegas, NV, USA, 14–17 July 2008; Arabnia, H.R., Mun, Y., Eds; CSREA Press: Las Vegas, NV, USA, 2008; pp. 323–329.
3. Rademacher, H. *Topics in Analytic Number Theory*; Springer Verlag: New York, NY, USA, 1973; Chapter 12, pp. 213–214.
4. Vandenbusschem, J.-J.; Leem, P.; Peutemanm, J. Linear phase approximation of real and complex pole IIR filters using MFIR structures. In Proceedings of the 5th European Conference on the Use of Modern Information and Communication Technologies, Gent, Belgium, 22–23 March 2012; De Strycker, L., Ed.; Nevelland: Gent, Belgium, 2012; pp. 221–231.
5. Vandenbussche, J.-J.; Lee, P.; Peuteman, J. An FPGA based digital lock-in amplifier implemented using MFIR resonators. In Proceedings of the International Conference on Signal Processing, Pattern Recognition and Applications, Crete, Greece, 18–20 June 2012; Petrou, M., Sappa, A.D., Triantafyllidis, G.A., Eds.; Acta Press: Calgary, AB, Canada, 2012; Volume 778–034, pp. 92–99.

6. Vandenbussche, J.-J.; Lee, P.; Peuteman, J. Design of an FPGA based TV-Tuner test bench using MFIR structures. *Annu. J. Electron.* **2013**, *7*, 21–25.
7. Vandenbussche, J.-J.; Lee, P.; Peuteman, J. On the coefficient quantization of multiplicative FIR filters. *Digit. Signal Process.* **2013**, *23*, 689–700.
8. Jackson, L.B. On the interaction of roundoff noise and dynamic range in digital filters. *Bell Syst. Tech. J.* **1970**, *49*, 159–184.
9. Chan, D.S.K.; Rabiner, L.R. An algorithm for minimizing roundoff noise in cascade realizations for finite impulse response digital filter. *Bell Syst. Tech. J.* **1973**, *52*, 347–385.
10. Mitra, S.K. *Digital Signal Processing: A Computer-Based Approach*, 3rd ed.; Mc Graw Hill Higher Education: New York, NY, USA, 2006; pp. 665–738.
11. Jackson, L.B. Round-off noise analysis for fixed point digital filters realized in cascade or parallel form. *IEEE Trans. Audio Electroacoustics*. **1970**, *18*, 107–122.
12. Chan, D.S.K.; Rabiner, L.R. Analysis of quantization errors in the direct form for finite impulse response digital filters. *IEEE Trans. Audio Electroacoustics*. **1973**, *21*, 354–366.
13. Fettweis, A. Roundoff noise and attenuation sensitivity in digital filters with fixed-point arithmetic. *IEEE Trans. Circuit Theory* **1973**, *20*, 174–175.
14. Mondal, K.; Mitra, S.K. Roundoff noise upper bounds for cascaded recursive digital filter structures. *IEE Proc. Electron. Circuit Syst.* **1982**, *129*, 250–256.
15. Lim, Y.C.; Liu, B. Design of cascade form FIR filters with discrete valued coefficients. *IEEE Acoust. Speech Signal Process*. **1988**, *36*, 1735–1739.
16. Montgomery Smith, L.; Henderson, M.E. Roundoff noise reduction in cascade realizations of FIR digital filters. *IEEE Trans. Signal Process.* **2000**, *48*, 1196–2000.
17. Dehner, G.F. Noise optimized IIR digital filter design-tutorial and some new aspects. *Signal Process.* **2003**, *83*, 1565–1582.
18. Shi, D.; Yu, Y.J. Design of discrete-valued linear phase FIR filters in cascade form. *IEEE Trans. Circuits Syst.* **2011**, *58*, 1627–1636.
19. Vandenbussche, J.-J. Analysis and Implementation of MFIR Filters in FPGA Technology. Ph.D. Thesis, University of Kent, Canterbury, UK, September 2012.