COEFFICIENT QUANTIZATION ERROR FREE FIXED-POINT IIR POLYNOMIAL PREDICTOR DESIGN

Jarno M. A. Tanskanen

Helsinki University of Technology, Institute of Intelligent Power Electronics
P.O.Box 3000, FIN-02015 HUT, FINLAND
Phone: +358-9-451 2446, Fax: +358-9-460 224, E-mail: jarno.tanskanen@hut.fi

ABSTRACT

In this paper, roundoff noise properties of fixed-point IIR polynomial predictors (FIPPs) and polynomial-predictive differentiators (FIPPDs) are investigated. These filters are designed by augmenting the corresponding FIR basis filters with magnitude response shaping feedbacks. Here we use ideally quantized coefficient (coefficient quantization error free) polynomial FIR predictors (PFPs) or predictive differentiators (PPFDs) as basis filters. Also, sufficient conditions for designing coefficient quantization error free direct form FIPPs are given for completeness, even though augmented FIR implementations of FIPPs and FIPPDs are preferred since they are coefficient quantization error free by their nature, and offer greater flexibility for magnitude response shaping without affecting the desired group delay properties set forth by the underlying PFP or PPFD.

1. INTRODUCTION

Finite calculation precision [1] may have a crucial effect on filter properties. This has been found to be the case with polynomial FIR predictors (PFP) [2] and polynomial-predictive FIR differentiators (PPFD). There exist design methods to produce PFPs and PPFDs, which provide for exact prediction and/or differentiation with short coefficient word lengths, [3], [4], respectively. Though these filters function exactly as desired even under coefficient quantization to six bits, for many applications it is desirable to be able to shape their frequency responses. This can be done efficiently by applying a simple feedback extension [5,6,7] which shapes the frequency response but does not affect the prediction and/or differentiation properties. If the FIPP or FIPPD is implemented using a direct form IIR structure with finite computation precision, the IIR extension may not be exact, and the prediction and/or differentiation properties are destroyed even if the FIRs were originally quantization error free. In this paper, the sufficient conditions for ideal IIR coefficient quantization are given for the filter length $N = 2$, so that the desired properties are exactly preserved under direct form IIR coefficient quantization. On the other hand, the augmented FIR structure, Figure 1, by its nature preserves the prediction and/or differentiation properties even under feedback coefficient quantization, and is thus preferred over the direct form IIR implementation. A few magnitude responses of the designed FIPPs are shown, roundoff noise effects are analyzed and a cure to roundoff noise effects is proposed.

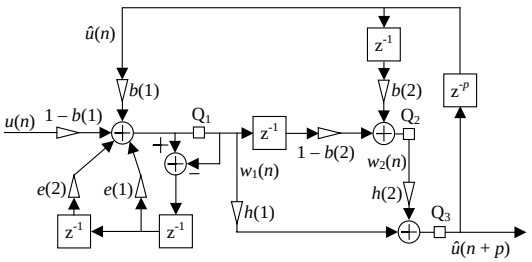

Figure 1. Structure of the augmented FIR [6] of length $N = 2$ with one error feedback installed with input signal $u(n)$, basis FIR coefficient vector $\mathbf{h} = [h(1)\ h(2)]$, feedback coefficient vector $\mathbf{b} = [b(1)\ b(2)]$, quantization error feedback coefficient vector $\mathbf{e} = [e(1)\ e(2)]$, hat denoting an estimate, squares Q_i, $i = 1, 2, 3$, denoting quantizers, and prediction step p.

For many control applications it would be highly beneficial if control loop delays could be avoided, likewise in many signal processing applications, signal processing delays may be harmful. In these applications, polynomial predictors and predictive differentiators may be successfully applied to overcome or lessen the effects of the delays while also attenuating noise.

In the examples in this paper, 8-bit two's complement number system with magnitude truncation is used. Multiply-accumulate (MAC) operations are calculated with 16-bit precision, and the results are truncated to 8 bits at the quantizers Q_i, $i = 1, 2, 3$. A filter is considered ideally quantized if it exactly fulfills the design criteria even under coefficient quantization.

In section 2, PFPs and PPFDs, and their ideal quantization design methods are shortly reviewed along with the augmented FIR structure. In section 3, direct form IIR implementations of FIPPs and FIPPDs are considered, and sufficient conditions for ideally quantized designs of direct form IIR implementations are given. In section 4, augmented FIR implementation of FIPPs is described with examples, and their roundoff noise analysis is performed and a cure is proposed in section 5. The roundoff noise robust designs are overviewed in section 6, and section 7 concludes the paper.

2. PFPs, PPFDs, AND THE CORRESPONDING AUGMENTED FILTER STRUCTURES

2.1. Polynomial FIR Predictors

PFPs, derived in [8], assume a low-degree polynomial input signal contaminated by white Gaussian noise. Filter output at a discrete time instant n, is defined to be a p-step-ahead predicted input,

$$u(n+p)=\sum_{k=1}^{N}h(k)\,u(n-k+1) \tag{1}$$

where $u(n)$ is input signal sample, $h(k)$ are filter coefficients, N is filter length, and p is prediction step. After providing for exact prediction, the rest of the degrees of freedom are used to minimize the white noise gain,

$$NG=\sum_{k=1}^{N}|h(k)|^2\ . \tag{2}$$

A set of linear constraints can be derived from the definition of the filter output (1) [2]:

$$g_0=\sum_{k=1}^{N}h(k)-1=0\,,\ldots,g_I=\sum_{k=1}^{N}k^I h(k)=0 \tag{3}$$

The constraints (3) give the prediction of the polynomial degrees 0, …, I, and from them can closed form solutions for the FIR coefficients for low-degree polynomial input signals be solved by the method of Lagrange multipliers [9]. The closed form solutions for FIR coefficients for the 1st, 2nd, and 3rd degree polynomial input signals can be found in [8].

2.2. Polynomial-Predictive FIR Differentiators

PPFDs are derived in the similar way as the PFPs. For the PPFDs, the filter input-output relation is written as [5,7]

$$\dot{u}(n+p)=\sum_{k=1}^{N}h(k)\,u(n-k+1) \tag{4}$$

where the dot denotes time derivative. The linear constraints on the filter coefficients are now given by [5]

$$g_I=\sum_{k=1}^{N}(N-k)^I h(k)=I(N-1+p)^{I-1}. \tag{5}$$

The closed form solution for FIR coefficients for the second-degree polynomial input signals is obtained from the constraints (5) and is given in [5].

It is worth noting that, for example, the coefficients of the first-degree $p=1$ filter of length $N=2$ and second-degree, $p=1$, $N=3$, are still exact if quantized to six bits or more for both PFPs and PPFDs, but the noise gains (2) of these filters are impractically high. Since these filters are coefficient quantization error free by their nature, they are good basis filters for the feedback extension [6] which relieves the noise gain problem. Otherwise, longer filters are to be used for achieving acceptable noise gains, and the method described in [3,4] is to be used to obtain correctly functioning fixed-point coefficient filters.

2.3. Ideal FIR Coefficient Quantization

There exists methods for designing ideally quantized-coefficient PFPs [3] and PPFDs [4] that function exactly correctly in short word length environments. The method is based on finding quantized filter coefficients that exactly fulfill the constraints (3), or (5) through a search algorithm. As the constraints (3), or (5), will be exactly satisfied, the prediction step p at zero frequency is exactly preserved in coefficient quantization, likewise is the unity magnitude gain at zero frequency. All thus designed filters are natural choices for basis filters for fixed-point IIR extension since their desired properties are not affected by coefficient quantization.

2.4. Augmented PFPs and PPFDs

To meet design specifications, which require polynomial signal prediction and/or differentiation and good noise attenuation, would require PFPs or PPFDs of the length of the order of several tens of taps. Also, it is not even possible to design very long, e.g. $N=100$, ideally quantized coefficient PFPs or PPFDs since the quantized coefficients of long filters tend to zero, and also, designing very long ideally quantized PFPs and PPFDs is computationally difficult. Shaping the magnitude response of an FIR filter with desired group delay properties is possible via an IIR extension [6], shown in Figure 1 for the PFP. Effectively, the IIR extension in Figure 1 introduces a smoothing feedback to the FIR basis filter. The overall transfer function of a augmented FIR is given in [6] and yields the transfer function of a feedback augmented 2-tap FIR as [6]

$$\begin{aligned}H(z)&=\frac{(h(1)-h(1)b(1))+(h(2)-h(2)b(1)-h(2)b(2)+h(2)b(1)b(2))z^{-1}}{1-h(1)b(1)z^{-p}-(h(2)b(1)+h(2)b(2)-h(2)b(1)b(2))z^{-p-1}}\\&=\frac{B(1)+B(2)z^{-1}}{1-A(2)z^{-p}-A(3)z^{-p-1}}\end{aligned} \tag{6}$$

with IIR coefficient vectors $\mathbf{B}=[B(1)\ B(2)]$ and $\mathbf{A}=[1\ A(1)\ A(2)]$, FIR coefficient vector $\mathbf{h}=[h(1)\ h(2)]$, and feedback coefficient vector $\mathbf{b}=[b(1)\ b(2)]$. Let us denote the corresponding quantized coefficient vectors as $\mathbf{B_q}$, $\mathbf{A_q}$, $\mathbf{h_q}$, and $\mathbf{b_q}$, respectively, and the quantized coefficient space by H. The transfer function (6) is used to calculate the magnitude response of the augmented structure. Since the feedback coefficients $\mathbf{b}$ of the augmented structure do not affect the desired prediction and/or differentiation properties but only shape the magnitude response, they can be freely selected from H to yield coefficient quantization error free FIPPs and FIPPDs as long as also $1-b(i)\in$ H, $i\in[1,2]$.

3. IDEALLY QUANTIZED COEFFICIENT DIRECT FORM FIPPs AND FIPPDs

In this section, it is shown possible to design direct form IIR implementations of FIPPs and FIPPDs with quantized coefficients even though the direct form IIR implementations are extremely sensitive to coefficient quantization and the augmented FIR structure, Figure 1, offers much greater design possibilities. Assume that the original basis FIR is such that quantization of a given $\mathbf{h}$ does not affect it, $\mathbf{h}=\mathbf{h_q}$. For preserving the exact prediction and/or differentiation properties, it is necessary that $\mathbf{h}$ remains untouched in FIR augmentation and coefficient quantization. Thereafter, quantizing $\mathbf{B}$ and $\mathbf{A}$ in (6) may obviously result in a situation in which (6) does not hold anymore, i.e., calculating $b(1)$ and $b(2)$ with given $h_q(1)$ and $h_q(2)$ from $B(1)$, $B(2)$, $A(2)$, and $A(3)$, does not necessarily yield unique values of $b(1)$ and $b(2)$. To guarantee that the direct form IIR implementation is coefficient quantization error free it is sufficient to ensure that

$$b(i),\ 1-b(i),\ h_q(i)b(j),\ h_q(i)b(j)b(k)\in \text{H},\ i,j,k\in[1,2]. \tag{7}$$

Thereafter, it is necessary to check the poles of the resulting filters to ensure that the poles remain sufficiently inside the unit circle. It turns out that it is possible to find such feedback coefficient vectors $\mathbf{b}$ that (7) and thus (6) hold. For example, for di-

rect form implementation of the first-degree, one-step-ahead predictive $p = 1$, FIPP of order $N = 2$ with eight bit coefficients and maximum accuracy of 0.03125, i.e., with 5 fractional bits, there are 237 possible vectors **b** that fulfill the constraints (7) and have poles inside the unit circle. An exemplary filter with **h** = [2 –1], **b** = [0.9375 0.5] is shown in Figure 2.

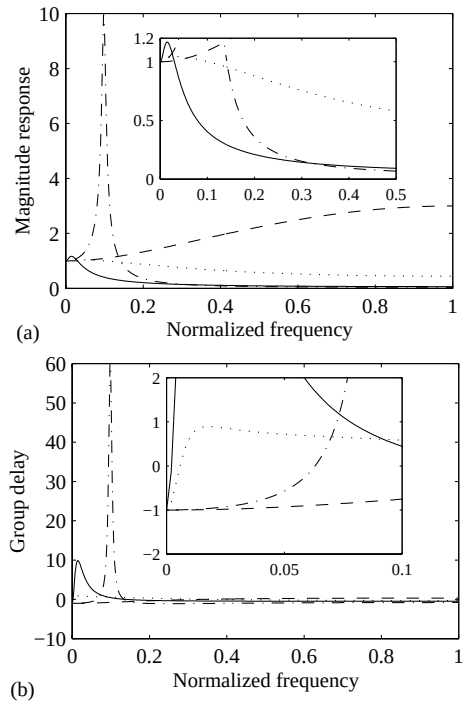

Figure 2. Magnitude responses (a) and group delays (b) of a direct form implementation of a roundoff error free one-step-ahead predictive first-degree FIPP with the feedback coefficients **b** = [0.9375 0.5] (dash-dot), two augmented type FIPPs with **b** = [0.6875 –0.9375] (dotted) and **b** = [0.9375 –0.9375] (solid), along with their basis filter basis FIR **h** = [2 –1] (dashed). For exact implementation of these FIPPs, 8-bit coefficient precision with 4 fraction bits is sufficient.

4. AUGMENTED FIR FORM DESIGN OF FIPPs AND FIPPDs

FIPP and FIPPD implementation using the augmented FIR structure, shown in Figure 1, offers design flexibility without having to be concerned with the coefficient quantization effects to the desired magnitude response and group delay properties. The magnitude response can be freely shaped using all possible feedback coefficients **b** such that $b(i), 1 - b(i) \in$ H, $i = 1, \ldots, N$, without the feedback coefficients affecting the desired group delay properties of the underlying FIR with the coefficients $\mathbf{h} = \mathbf{h}_q$.

In Figure 2, two examples of one-step-ahead predictive first-degree FIPPs with the basis FIR of length $N = 2$, implemented using the augmented structure, Figure 1, with coefficient precision of eight bits, are shown with **h** = [2 –1] with **b** = [0.6875 –0.9375], and **b** = [0.9375 –0.9375]. The one with **b** = [0.6875 –0.9375] exhibits a little wider prediction band but less noise attenuation at high frequencies. The second example with **b** = [0.9375 –0.9375] exhibits narrow prediction band with low passband peak and good noise attenuation at high frequencies. It is worth stressing that filters in Figure 2 are quantization error free since now $h(i), b(i), 1 - b(i) \in$ H, $i \in [1,2]$, and the augmented FIR implementation is used. The FIR basis filter **h** = [2 –1] is also seen in Figure 2.

5. ROUNDOFF NOISE OF AUGMENTED FORM FIPPs AND FIPPDs

As we are concerned with quantized coefficient IIRs, it is necessary to perform the roundoff noise analysis [10] of the designed IIRs. Even though the designed FIPPs and FIPPDs are coefficient error free, quantization of MAC outputs introduces roundoff noise into the system.

Modeling quantization as noise added to each summation node, and assuming two's complement arithmetic, the average roundoff noise power at each quantizer σ_0^2, and thereafter the total average roundoff noise power of a filter σ_e^2, can be expressed, respectively, as [10]

$$\sigma_0^2 = \Delta^2/12\,,\ \sigma_e^2 = \sigma_0^2 \sum_j k_j S_j\,,\ S_j = \sum_{n=0}^{\infty} \left| g_j(n) \right|^2 \quad (8,9,10)$$

where Δ is quantization step, j number of summation nodes, k_j products feeding into the jth node, and $g_j(n)$ is impulse response from the ith summation node to the filter output. Transfer functions for calculating $g_j(n)$, $j = 1, 2, 3$, in (10) for the FIPP structure seen in Figure 1 can be found in [11].

Individual noise power contributions S_j, $j = 1, 2, 3$, (10) of the summation nodes in Figure 1, for the three filters shown in Figure 2, are listed in Table 1 with coefficient scaling of 8 in order to make S_j converge. Thereafter, it is still to be observed that since **h** = [2 –1], and signals $w_1(n), w_2(n) \in$ H, Q_3 does not produce any roundoff error since now $w_1(n)h(1) + w_2(n)h(2) \in$ H, taking that the dynamic range is not exceeded. Thus we can set $S_3 = 0$ in (9). For quantizers Q_1 and Q_2, almost all quantizer inputs within the dynamic range do not belong to H.

Table 1. Roundoff noise contributions S_j, $j = 1, 2, 3$, (10) of the individual summation nodes in Figure 1 for the filters seen in Figure 2. For this **h** = [2 –1], $S_3 = 0$ when calculating (9).

b	S_1	S_2	S_3
[0.9375 0.5]	0.077	0.016	1.032
[0.6875 –0.9375]	0.081	0.016	1.00
[0.9375 –0.9375]	0.080	0.016	1.00

6. ROUNDOFF NOISE ROBUST FIPPs AND FIPPDs

In FIPPs and FIPPDs, roundoff noise can be a problem but, at least to some extend, it can be reduced by error feedback [10,12]. Observing S_i, $i = 1,2,3$, in Table 1, it is clear that the roundoff errors produced by Q_1 are to be combated (since with **h** = [2 –1], Q_3 does not contribute to the roundoff error). In Table 2, a few examples showing that even short error feedbacks

with unity gains, $\mathbf{e} = [1\ 1]$, or $\mathbf{e} = [1\ 0]$, c.f. Figure 1, can be successfully applied in augmented PFP design, although the feedback structure is generally very depended on the filter coefficients and input signal statistics. This selection of $\mathbf{e}$ is naturally the simplest and least costly to implement, and thus employed here. In Table 2, the mean square errors (MSE) when using the exemplary filters in Figure 2 to filter a ramp $u(n) = 0.01\,n$, $n = -100, 99, \ldots, 100$, and a sinusoid $u(n) = \sin(0.01\,n)$, $n = 0, 1, \ldots, 5\pi$, are shown. To let the filters settle, the ramp is preceded by 200 samples of –1, and the sinusoid by 200 zero samples. Improvement with a very simple error feedback is approximately from 40 % to 90 % in half of the cases shown in Table 2. Performance of the FIPP $\mathbf{h} = [2\ -1]$, $\mathbf{b} = [0.9375\ -0.9375]$, is much improved with regard to both signals by simple error feedbacks. However, an improper error feedback may easily make filtering perform worse than without error feedback. The filter $\mathbf{h} = [2\ -1]$, $\mathbf{b} = [0.9375\ 0.5]$ benefits from the error feedback when operating on a ramp signal, but cannot be improved when filtering a sinusoid. The filter $\mathbf{h} = [2\ \ -1]$, $\mathbf{b} = [0.6875\ -0.9375]$, cannot be improved by these types of error feedback for either of the signals. For practical applications rigorously designed error feedbacks should be applied [12].

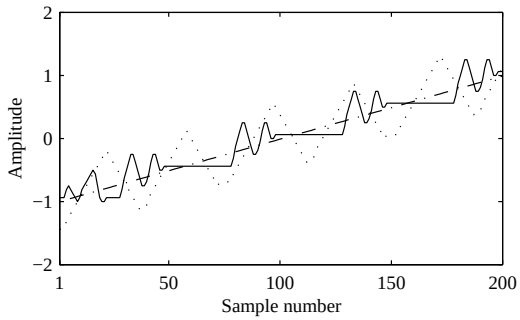

Figure 3. Effect of the installed error feedback on filtering the ramp with the FIPP $\mathbf{h} = [2\ -1]$, $\mathbf{b} = [0.9375\ 0.5]$. Shown are the desired one-step-predicted filter input (dash-dot), and filter outputs with (solid) and without (dotted) the 3-bit error feedback of length 1, c.f. Table 2.

7. CONCLUSIONS

In this paper, it has been shown that polynomial-predictive FIRs and polynomial-predictive FIR differentiators can be augmented to IIR filters with feedback extensions to shape their magnitude responses even in short word length fixed-point environments. It is shown possible to design coefficient quantization error free polynomial-predictive and differentiative IIRs. With the designed coefficient quantization error free filters, a roundoff noise problem is identified, and error feedback is shown effective in solving the roundoff noise problem.

8. ACKNOWLEDGMENT

The financial support granted by Jenny and Antti Wihuri Foundation, Finland; Walter Ahlström Foundation, Finland; The Finnish Society of Electronics Engineers; and by Foundation of Technology, Finland; is greatly acknowledged.

Prof. Seppo J. Ovaska of the Institute of Intelligent Power Electronics, Helsinki University of Technology (HUT), Finland, is thanked for reviewing and commenting the manuscript. Mr. Petri Jehkonen from Atmel Corp., and Mr. Aki Suikonen from the Signal Processing Laboratory of HUT, are thanked for enlightening processor technology discussions.

9. REFERENCES

[1] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*. New York, NY, USA: Macmillan Publishing Company, 1992.

[2] J. M. A. Tanskanen and S. J. Ovaska, "Coefficient sensitivity of polynomial-predictive FIR differentiators: Analysis," in *Proc. The 42nd IEEE Midwest Symp. on Circuits and Systems*, NM, USA, Aug. 1999, in press.

[3] J. M. A. Tanskanen and V. S. Dimitrov, "Round-off error free fixed-point design of polynomial FIR predictors," in *Proc. The 33rd Asilomar Conf. on Signals, Systems, and Computers*, CA, USA, Oct. 1999, pp. 1317–1321.

[4] V. S. Dimitrov and J. M. A. Tanskanen, "Round-off error free fixed-point design of polynomial-predictive FIR differentiators," in *Proc. the IASTED Int. Conf. Intelligent Systems and Control*, CA, USA, Oct. 1999, pp. 199–204.

[5] S. Väliviita, S. J. Ovaska, and O. Vainio, "Polynomial Predictive filtering in control instrumentation: A review," *IEEE Trans. Industrial Electronics*, vol. 46, pp. 876–888 Oct. 1999.

[6] S. J. Ovaska, O. Vainio, and T. I. Laakso, "Design of predictive IIR filters via feedback extension of FIR forward predictors," *IEEE Trans. Instrumentation and Measurement*, vol. 46, pp. 1196–1201, Oct. 1997.

[7] S. Väliviita and S. J. Ovaska, "Delayless recursive differentiator with efficient noise attenuation for control instrumentation," *Signal Processing*, vol. 69, pp. 267–280, Sept. 1998.

[8] P. Heinonen and Y. Neuvo, "FIR-median hybrid filters with predictive FIR substructures," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, pp. 892–899, June 1988.

[9] D. Bertsekas, *Constrained Optimization and Lagrange Multipliers Methods*. New York, NY, USA: Academic Press, 1982, Chapter 1.

[10] L. B. Jackson, *Digital Filters and Signal Processing*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996.

[11] P. T. Harju, "Roundoff noise properties of IIR polynomial predictive filters," in *Proc. IEEE Instrumentation and Measurement Conf.*, Ottawa, Canada, May 1997, pp. 66–71.

[12] T. I. Laakso and I. O. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," *IEEE Trans. Signal Processing*, vol. 40, pp. 1096–1107, May 1992.

Table 2. Mean square errors (MSE) when filtering a ramp or a sinusoid with the exemplary FIPPs seen in Figure 2 with and without error feedback installed over the quantizer Q_1. Superscipt "1" denotes the minimum MSE found over the cases listed.

Signal	Ramp	Sinusoid	Ramp	Sinusoid	Ramp	Ramp
Feedback coefficients	$\mathbf{b} = [0.9375\ -0.9375]$	$\mathbf{b} = [0.9375\ -0.9375]$	$\mathbf{b} = [0.9375\ 0.5]$	$\mathbf{b} = [0.9375\ 0.5]$	$\mathbf{b} = [0.6875\ -0.9375]$	$\mathbf{b} = [0.6875\ -0.9375]$
Error feedback length	1	2	1	1 or 2	1 or 2	1 or 2
Error feedback coefficients	$\mathbf{e} = [1\ 0]$	$\mathbf{e} = [1\ 1]$	$\mathbf{e} = [1\ 0]$	$\mathbf{e} = [1\ 0]$ or $\mathbf{e} = [1\ 1]$	$\mathbf{e} = [1\ 0]$ or $\mathbf{e} = [1\ 1]$	$\mathbf{e} = [1\ 0]$ or $\mathbf{e} = [1\ 1]$
Error feedback accuracy (bits)	3	1	3	1, 2 or 3	1, 2 or 3	1, 2 or 3
MSE without error feedback	0.1196	0.1360	0.1036	0.1033^1	0.0017^1	0.0017^1
MSE with error feedback	0.0091	0.0699	0.0284	> 0.1033	> 0.0017	> 0.0017