A Novel Approach for Diminishing and Predicting the Error Dynamic Range in Finite Wordlength FIR Based Architectures

A Benkrid, K Benkrid and D Crookes
School of Computer Science, The Queen's University of Belfast, Belfast BT7 1NN, UK

ABSTRACT

This paper analyses the effects of fixed-point arithmetic in FIR filter based architectures, based on the roundoff statistical noise model. A novel approach, which allows diminishing the error dynamic range and predicting its value according to the wordlength precision, is suggested. This permits the user to preset the fraction precision according to the sought architecture's precision. The efficiency of this approach is demonstrated through the 2-D DWT Biorthogonal 9&7 transform.

1. INTRODUCTION

Usually, DSP algorithms are initially developed without regard to the effects of the quantisation errors introduced when implemented in digital systems. When implementing a DSP system in hardware, truncation or rounding operations are often undertaken to limit the growth in precision of the intermediate calculations and thus excessive hardware requirements [1]. However, wordlength reduction does introduce error into the data architecture's, so the designer must balance the need for an efficient implementation with acceptable output quality.

The two sources of error in fixed-point arithmetic are the [2]:

- **Signal Wordlength Quantisation (SWQ)** due to the truncation or rounding carried on the internal signals as well as the filter output, which results in what is called the *roundoff noise*.
- **Filter coefficient quantisation (FCQ)** due to the quantisation applied on the filter coefficients values. The resulting filter impulse response will be then different from the ideal.

In this paper, we will analyse the effect of those two factors on the architecture's precision. In section 2, we give the roundoff noise statistical model. The validity of this model is then assessed in section 3. Section 4 is devoted to our novel approach owing to diminish and predict the error dynamic range in FIR based architectures. A case study is reported in section 5. Finally, a conclusion is drawn.

2. ROUNDOFF NOISE STATISTICAL MODEL

The effect of using finite wordlength precision has been studied for some time. Rabiner in [2] lays down standard models for quantisation errors and error propagation through linear time-invariant systems, based on a linearisation of the truncation or rounding operation. We linearise the truncation or rounding operation by replacing it by an addition with an error signal, *e*, as illustrated in Fig 1. The multiplier is modelled as infinite precision multiplier followed by an adder where the quantisation error(noise), *e*, is added to the product. To statistically model this quantisation error, often the following assumptions are made [2]:

1 e(n) is uniformly distributed white noise.
2 e(n) is a wide-sense stationary random process
3 e(n) is uncorrelated to all other signals such as input and other noise signals.

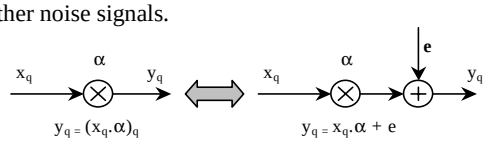

Figure 1. Fixed point quantisation noise model, the letter q stands for quantisation

When implementing a filter with finite wordlength precision, the critical information from the user's perspective is the error variance at the filter output. We therefore propagate the SWQ errors, $e_i(t)$, associated with each multiplier to the filter output. This allows us to estimate their effect on the overall system precision. Doing so, the error signal E(t) at N-tap filter output is $E(t) = \sum_{i=0}^{N-1} e_i(t)$ .Under the above model's statistical assumptions, the total output error variance (noise power) can be shown to be the addition of the individual noise powers contributed by each of the error sources $e_i(t)$. Then, we need to find the statistical parameters of each roundoff error $e_i(t)$.

Unlike with the rounding operation, the truncation introduces a bias (its mean is not equal to zero), and therefore it leads to lower precision. In the following, we will limit ourselves to the rounding operation.

According to [3], when rounding b_1-bit precision to b_2-bit precision, the mean of the rounding error satisfies:

$$m_{qR} = 0 \qquad (1)$$

and the variance is given by:

$$\sigma_{qR}^2 = \frac{1}{12}(2^{-2b_2} - 2^{-2b_1}) \qquad (2)$$

To calculate the noise variance and mean at the outputs of a multistage FIR filters, we use the following theorem

Theorem[4]

If an FIR filter, h, is fed with a stationary zero mean white noise of variance σ^2, then the noise at its output is of zero mean and of variance, σ_0^2, equal to:

$$\sigma_0^2 = \sigma^2 \sum_n |h(n)|^2 \qquad (3)$$

where h(n) are the filter's coefficients. The term $\sum_n |h(n)|^2$ represents the filter power gain. Therefore by applying the above theorem and the previous statistical noise assumptions on a two cascading FIR filters (h_0 and h_1 respectively), the error noise variance at the second filter σ^2 is:

$$\sigma^2 = \sigma_{h_1}^2 + K_{h_1}\sigma_{h_0}^2 \qquad (4)$$

and its mean is equal to zero. K_{h_1} is the $h_1$filter's power gain. For a multistage FIR module, this formula has simply to be extended to the number of cascading FIR filters.

It is well known that when adding two uncorrelated random variables, the resulting error Probability Distribution Function (PDF) is the convolution of the two input PDFs [5]. We therefore expect triangle-shaped error PDFs for the sum of two rounding errors e_i(recall having uniform distribution, i.e a rectangle-shape PDF). If a third random variable is added, the resulting PDF error gets a bell-shaped graph. By adding more uniform random variables, the resulting error, E, tends to have a Gaussian distribution according to the central limit theory [5]. Figure 2 shows the roundoff error PDF shape (histogram) evaluated at the output of the low biorthorgonal 9&7 FIR filter (9taps!) when rounding at 2-bit precision. This is deduced by taking the same filter implemented with double precision (64 bits) filter coefficients and intermediate results as a reference. A 256x256 Lena standard image has been chosen as a testing input. Therefore, we will assume that the filter output roundoff noise be following a Gaussian distribution.

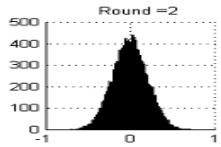

Figure 2. Noise histogram at the output of 9-taps low Biorthogonal 9&7 filter due to 2-bit rounding order

A Gaussian distribution function with *mean* (m) and *standard deviation* σ (SDV), is defined from minus to plus infinity. However, mathematical computation shows that **99.994%** of the values fall within 4 standard deviations of the mean, that is, between m-4σ and m+4σ[5]. Thus, almost all values of a Gaussian distribution function lie *within 4 standard deviations of the mean*. Throughout, we refer to this property by the *Gaussian Range (GR) property*. This property is so important since it permits us to predict the dynamic error range at any FIR filter output by using its estimated theoretical variance. However, simulations need first to be undertaken to assess to what extent the model assumptions are valid in practice. For this purpose, we need to address the following question: *Is the predicted mean and the variance accurate (even for multistage FIR modules)?*

3. ROUNDOFF NOISE STATISTICAL MODEL ACCURACY

To answer the previous question, we use 3-Stage (thus 6 cascading filters!) Biorthogonal 9&7 2-D DWT transform. This family has been chosen due to its widespread use in compression based system [6]. Figure 3 shows the basic one stage transform. It produces four images – three detailed images along the horizontal(LH), vertical(HL) and diagonal(HH), and one coarse approximation of the original image(LL). The Latter can be further decomposed by using the same block diagram transform. Three stages of decomposition are usually considered sufficient for most applications.

A 256x256 Lena standard image will be used as a test image. *Uniform rounding* is carried on the architecture's wordlength. The resulting transform image is then subtracted from the full precision transform (where even the architecture's wordlength are represented with full precision) to produce the error transform image. Because of the paper size limit, we will be reporting the simulations results performance at only the L,H,LL,LH bands (see fig 3) of the 3-stage transform. Those have been chosen because they infer higher cascading FIR filters modules. We will use also these abbreviations: *Round(I)* to refer to the I-bit rounding, *Pred* and *Sim* stands respectively for the predicted and the simulated value, *CWL* to the filter coefficients wordlength and Bio 9&7 to the Biorthogonal 9&7.

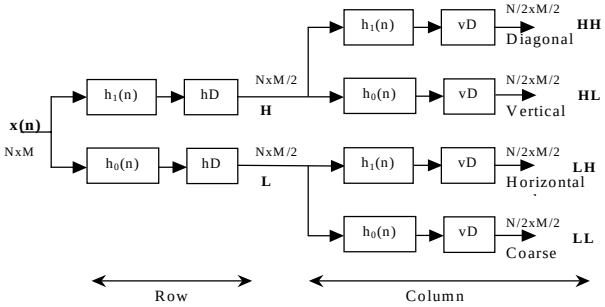

Figure 3. Basic stage of the 2-D wavelet transform using the filter bank structure

h_0/h_1 stands for low pass/ high pass filter
hD: horizontal decimation, i.e. keeps one column from every two
vD: vertical decimation, i.e. keeps one row from every two

a. SWQ factor

By using full filter coefficients representation and different rounding order on the intermediate results, the simulation results show that the statistical noise model gives a reasonably accurate prediction of the error standard deviation and mean values. Table 1.a&b show the values obtained for rounding order of 2.

Round(2)	Stage 1		Stage 2		Stage 3	
	L	LL	L	LL	L	LL
Pred	0	0	0	0	0	0
Sim	0.012	0.016	0.021	0.029	0.047	0.055

(a) The mean

Round(2)	Stage 1		Stage 2		Stage 3	
	L	LL	L	LL	L	LL
Pred	0.216	0.309	0.378	0.4399	0.490	0.547
Sim	0.221	0.321	0.404	0.483	0.577	0.654

(b) The standard deviation

Table 1. The predicted and the simulated mean and standard deviation values at 3-stages Bio 9&7 2D-DWT transform using full CWL

Being able to predict the mean and the subbands' SDV values, the GR property can be used to preset the fractional precision needed for the sought error dynamic range. Table 2 gives for rounding order of 2, the dynamic range of the predicted and the actual (simulated) quantisation errors at the 2-D DWT Bio9&7 low sub-bands.

Round(2)	Stage 1		Stage 2		Stage 3	
	L	LL	L	LL	L	LL
Pred	-0.866 0.866	-1.237 1.237	-1.510 1.510	-1.759 1.759	-1.961 1.961	-2.188 2.188
Sim	-0.831 0.832	-1.160 1.164	-1.443 1.840	-1.770 2.134	-2.240 2.325	-2.445 2.184

Table 2 Comparison between the predicted and the simulated error dynamic range at the 2-D DWT Bio9&7 subbands using full precision CWL.

The above results validate clearly the statistical model assumption when only quantisation is carried out on the

intermediate results. Further investigation should be carried out to see the FCQ factor effect.

b. FCQ factor

As stated previously, the rounding noise statistical model has been introduced without taking into account the FCQ factor. If the filter coefficients are quantified (limited wordlength), the accuracy of the predicted error dynamic range risks degrading. Table 3 highlights this problem when the filter coefficients are implemented with just 8-bits(as it is often adopted in hardware implementation).

Round(2)	Stage 1		Stage 2		Stage 3	
	L	LL	L	LL	L	LL
Pred	-0.866 0.866	-1.237 1.237	-1.510 1.510	-1.759 1.759	-1.961 1.961	-2.189 2.189
Sim	-2.376 0.278	**-5.731** **0.015**	**-10.80** **-1.234**	**-19.21** **-3.158**	**-33.51** **-7.36**	**-56.05** **-13.331**

Table 3. Comparison between the predicted and the simulated error dynamic range at the 2-D DWT Bio 9&7 subbands using 8-bits CWL

We can easily see from table 3 (compared to table 2) how the accuracy of the predicted error dynamic range has remarkably dropped.

To try to recover (some of) the accuracy, which has been lost through FCQ, and therefore maintaining the validity of the previous statistical model, we have developed the error cancellation approach in order to cancel the FCQ contribution on the overall error.

4. FILTER ERROR CANCELLATION APPROACH

When implemented in hardware, the filter coefficients are quantified, and not exactly represented. Thus some error, Δe, is introduced. If for instance, a two taps filter with coefficients $h_0^{(t)}, h_1^{(t)}$ (t stands for theoretic) has to be implemented in fixed-point arithmetic, the coefficients will be quantified to $h_0^{(h)}, h_1^{(h)}$ (h stands for hardware). Such quantisation produces two errors Δe_0, Δe_1 associated respectively with $h_0^{(t)}$ and $h_1^{(t)}$. Subsequently, the filter implemented in hardware can be modelled with a full precision coefficient filter and an *error coefficient filter* (see fig.4).

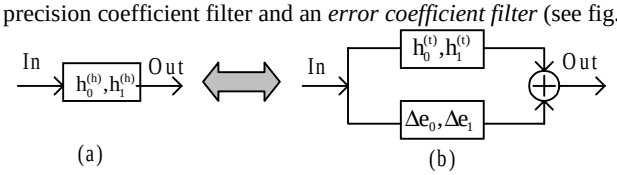

Figure 4 Hardware implementation filter Model

From Figure 4(a), the output $out_0^{(h)}$ satisfies:

$$out_0^{(h)} = In_0\, h_0^{(h)} + In_1\, h_1^{(h)}$$

which can be written, using the model Figure 4(b) as:

$$out_0^{(h)} = In_0\,(h_0^{(t)} + \Delta e_0) + In_1\,(h_1^{(t)} + \Delta e_1)$$

$$out_0^{(h)} = (In_0\, h_0^{(t)} + In_1\, h_1^{(t)}) + (In_0\, \Delta e_0 + In_1\, \Delta e_1)$$

and so

$$out_0^{(h)} = out_0^{(t)} + \Delta e$$

where $out_0^{(t)}$ denotes the theoretic filter output, and Δe is the error associated with its hardware implementation.

We can notice then that the error Δe can be *cancelled* if the term $(In_0.\Delta e_0 + In_1.\Delta e_1)$ is made equal to zero. If not, we try to get as close as possible to this equality by carefully choosing the quantified representation of the coefficients (and hence Δe_0, Δe_1) so to cancel as much of the error as possible. Note that in most natural signals (such images), the values of neighbouring samples are strongly correlated. Also, there is often some correlation between the filtered signal samples [4]. If for instance, we pass a relatively smooth image through the above 2-elements filter, we can assume therefore that $In_0 \approx In_1 \approx In$. The error Δe will be equal to $In\,(\Delta e_0 + \Delta e_1)$. The reader can deduce easily that this error would tend to zero if $\Delta e_0 \approx -\Delta e_1$, and so the error Δe will be equal to $(In_0 - In_1)\,\Delta e_0$. Since the image features are contained in the edges (which show a big fluctuation, $In_0 >> In_1$ or the inverse), filtering through those pixels will increase the error Δe drastically. Nevertheless, by knowing the input dynamic range, one can still ensure that the error Δe is bounded by a value η (at any position of the filtered image) by choosing enough bits in order that Δe_0 and Δe_1 are less than $(\eta / (Max - Min))$, where Max and Min denote, respectively, the maximum and minimum input filter values.

This scheme can be expanded to any filter length and most of the practical signals. The cancellation just has to be done by considering pairs of coefficients (or sometimes triples if the filter length is odd). By following this approach, the quantisation errors are sought to be limited to just the SWQ factor which is statistically well modelled.

5. CASE STUDY

In the following, the error filter coefficients will be computed by subtracting the quantified coefficients filter from the real filter values one. Figure 5(a) shows the low Bio 9&7 error filter coefficients due to the quantisation at 8-bit CWL. As shown in this figure, we can see that although the error filter coefficients values are at 10^{-3} order, they almost all have the same sign increasing then the dynamic error range of this error filter. This clearly explains why the accuracy of the error dynamic range prediction has so dropped in table 3.

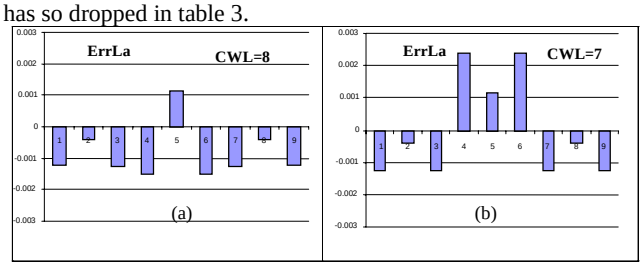

Figure 5. Low error Bio 9&7 filter coefficients using (a) 8-bit and (b) 7-bit CWL.

In[7], we have shown that for any arbitrary input image, the dynamic range is [114.10,898.13] at the LL output of the second stage, therefore for $\eta \approx 2$, Δe_i should be of 10^{-3} order(see section 4). We have found rounding at 7-bit CWL enough to deliver such individual coefficient precision (see fig 5(b)). In the following, we refer to the resulting I-coefficient low(high) filter with $La_7(I)(Ha_7(I))$. For each $La_7(I)(Ha_7(I))$, we associate the coefficient quantisation error $e_{La7}(I)(e_{Ha7}(I))$. The real values of the low and the high Bio 9&7 coefficients can be found in [6]. We also note by *LsbV(W)* the LSB value associated with the W-bits representation. It is equal to $1/2^W$ and for example LsbV(9)= 0.00195, LsbV(8)=0.00390 and LsbV(7)= 0.00781.

To enhance the precision performance of the 7-bit CWL low Bio 9&7, our error cancellation approach can be applied by choosing instead a non-uniform coefficients wordlengths (CWL).

For this 9-tap filter, the cancellation could be applied on a subset of 3 samples, or alternatively on the first 2 coefficients and then the subsequent 5 coefficients. If the last option is adopted, the reader can verify easily that the individual FCQ error associated with the first two coefficients are respectively $e_{La7}(1)$=-0.00123 and $e_{La7}(2)$=-0.00041. The sum of those individual errors is equal to "-0.00164" which is close to "-LsbV(9)" value. Therefore, the LsbV(9) value can be added to $e_{La7}(1)+e_{La7}(2)$ to ensure error cancellation between those two coefficients. Since the quantisation error associated with $La_7(2)$ is minimal (0.00041), we subtract LsbV(9) from $La_7(1)$ (=10/256). The resulting updated coefficient $La_7(1)$ is then equal to 19/512. Now for the {3,4,5,6,7} order set coefficients, the associated sub-filter dynamic should be minimised while ensuring the overall error filter dynamic being small. The user can verify that $[e_7(3)+e_7(4)+e_7(6)+e_7(7)]+e_7(5)$=0.00344, which is nearly equal to LsbV(8). Therefore if we add the LsbV(8) value to the fifth coefficient(109/128), we get 109/128+LsbV(8)=219/256. This update has not been applied on the other coefficients since there was already a partial error cancellation between them (see fig 5(b)). The resulting filter coefficients are [19/512,-3/128,-7/64,3/8,219/256,3/8,-7/64,-3/128,19/512]. Fig 6 shows the resulting low Bio 9&7 error filter coefficients, whereas table 4 gives the improvement achieved in predicting the error dynamic range at the low subbands outputs compared to table 3.

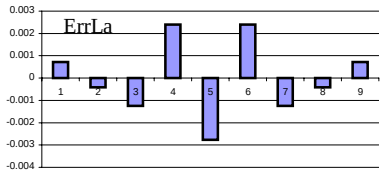

Figure 6. Low error Bio 9&7 filter coefficients using hybrid CWL

Round(2)	Stage 1		Stage 2		Stage 3	
	L	LL	L	LL	L	LL
Pred	-0.866 0.866	-1.237 1.237	-1.510 1.510	-1.759 1.759	-1.961 1.961	-2.189 2.189
Sim	-1.003 1.000	-1.807 1.234	-2.179 1.920	-3.027 2.276	-3.600 2.334	-3.083 2.381

Table 4. Comparison between the predicted and the simulated error dynamic range at the 2-D DWT Bio 9&7 subbands using Hybrid CWL

The same approach can be applied on the high Bio 9&7 filter (7 taps). Table 5 gives the predicted and the simulated error dynamic range at the high bands of the transform using 8-bit CWL.

Round(2)	Stage 1		Stage 2		Stage 3	
	H	LH	H	LH	H	LH
Pred	-0.764 0.764	-1.165 1.165	-1.454 1.454	-1.709 1.709	-1.918 1.918	-2.148 2.148
Sim	-0.336 2.229	-0.520 3.200	-0.267 4.409	-0.209 5.792	0.458 7.217	0.443 10.498

Table 5. Comparison between the predicted and the simulated error dynamic range at the 2-D DWT Bio 9&7 subbands using Hybrid CWL

On the other side Fig 7(a) shows the error filter using 7-bit CWL. We can see from fig 7(a), an already partial error cancellation taking place with the two first coefficients. Therefore, we can focus on the subset coefficients order {3,4,5}. The user can verify that $e_{Ha7}(3)+ e_{Ha7}(4)+ e_{Ha7}(5)$=-0.00699,which is close to -LsbV(7) value. Therefore, we can minimise this sum by substracting LsbV(7) to some of the high filter coefficients. The third and the fifth sample can be updated with 0.5*LsbV(7) i.e. LsbV(8). The resulting filter high filter coefficients is [-1/16,5/128,107/256,-101/128,107/256,5/128,-1/16].

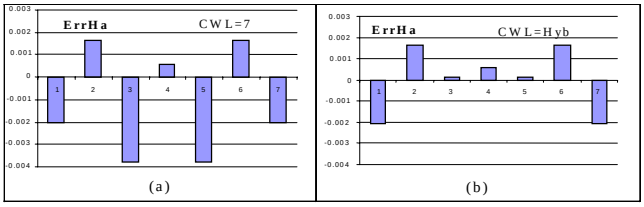

Figure 7. High error Bio 9&7 filter coefficients using (a) 7-bit CWL and (b) hybrid CWL

Table 6 shows the improvement reached comparing to table 5.

Round(2)	Stage 1		Stage 2		Stage 3	
	H	LH	H	LH	H	LH
Pred	-0.764 0.764	-1.165 1.165	-1.454 1.454	-1.709 1.709	-1.918 1.918	-2.148 2.148
Sim	-1.057 0.787	-1.571 1.148	-1.673 1.671	-1.942 2.422	-2.856 2..313	-3.083 2.381

Table 6 Comparison between the predicted and the simulated 3-stage Bio9&7 2-D DWT subbands' error noise dynamic range using hybrid CWL

The previous results confirm clearly the validity of our error cancellation approach. i.e. the best way is to have the error dynamic filter nearly equal to zero with neighbouring error coefficients cancelling each others. Moreover, for the Bio 9&7 family, our approach gave an average number of bits per low and high filter coefficient of 6.4 and 6.42 respectively comparing to an average of 7 and 7.52 with the 8 bits coded. Though, using fewer bits, it gives better precision!

6. CONCLUSION

In this paper we have given a thorough analysis of filter coefficient quantisation and roundoff noise in FIR based architectures. A novel approach, the error cancellation, has been presented. This approach allows the designer to reduce the error dynamic range due to quantisation, and especially to predict its value. The approach tends to use hybrid wordlength representation for the different filter coefficient in order to reduce the error filter dynamic range. This allows limiting the quantisation effects to just the roundoff error which is well modelled statistically. The model made an extra assumption of a Gaussian distribution output noise. The simulations results demonstrate clearly the validity and the efficiency of our approach

REFERENCES

[1]. P.Fiore, "Efficient Wordlength Reduction Techniques for DSP Applications",Journal of VLSI Signal Processing Systems 24, 9–18 (2000)

[2]. Lawrence R. Rabiner, "Theory and application of digital signal processing" Prentice-Hall, 1975.

[3]. G.Constantinides et al , "Truncation Noise in Fixed-Point SFGs" IEE Electronics Letters Vol 35, No23, pp 2012-2014. November 1999

[4]. A.V. Oppenheim and R.W. Schafer," Digital Signal Processing",*Prentice-Hall*, 1975.

[5]. Papoulis, "Probability, Random Variables, and StochasticProcesses", *McGraw-Hill*, 1984.

[6]. Vetterli M, Kovacevic M, ''Wavelets and Subband Coding,'' *Prentice Hall, New Jersey*, USA, 1995

[7]. A.Benkrid et al, ''Dynamic range of the1-D and 2-D forward and inverse DWT transform'' ISCA2003.