International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 03, March 2016
www.ijates.com
ijates
ISSN 2348 - 7550

DESIGN OF QUANTIZED FIR FILTER USING COMPENSATING ZEROS

Nivedita Yadav, O.P. Singh, Ashish Dixit

Department of Electronics and Communication Engineering, Amity University, Lucknow Campus, Lucknow, (India)

ABSTRACT

Quantization of filter coefficients are involved in real world implementation of digital filters i.e. coefficients are approximated using fixed point mathematics. The quantization of filter coefficients varies the pole-zero plot and frequency response of quantized filter from the originally desired unquantized filter's pole-zero plot and frequency response respectively. This paper presents a design method of quantized finite impulse response (FIR) filters to address design and implementation issues of real world implementation. Using Canonical Signed Digit (CSD) representation, the floating point FIR filter coefficient is converted to fixed point multiplier-less FIR filter coefficient. The fixed point filter coefficients can be implemented in smaller and faster hardware than floating point filter coefficients. The filter is realized using cascade form and the effect of quantization of filter coefficients in one cascade section is compensated in other cascade section by pushing the zeros. The proposed method concludes that quantized FIR filter using CSD technique is most efficient and matches with the desired frequency response keeping small and fast hardware.

KeywordsL: CSD, IIR filter, quantization

I. INTRODUCTION

The coefficient of a filter designed is floating point number and any challenges arising in implementation of deigned filter in fixed point hardware coefficient are not considered. For real world implementation, the hardware as well as signal processing aspects must be considered simultaneously to obtain an optimum filter implementation. The frequency response of the final fixed point filter depends on the approximations made during the conversion from floating point to fixed point. A digital FIR filter having floating point filter coefficient can be designed using the windowing method and the Parks-McClellan method [1,2]. For a real time application, the floating point coefficients of filter must be converted to fixed point to perform more quickly in hardware. A multiplier-less implementation of a filter is used for embedded system applications, multiplications are replaced with the faster and cheaper shifts and additions. The filter coefficients are converted to a fixed point, multiplier-less by quantizing the original floating point filter coefficients. Quantization alters the zeros of the original filter as the filter coefficients are either truncated or rounded off during quantization and consequently the frequency response of original filter is also altered.This paper presents a method for conversion of a finite impulse response (FIR) floating point filter design into a fixed point multiplier-less filter design. FIR filters are often preferred over Infinite Impulse Response (IIR) filters since they exhibit no stability

problems and can be designed with exact linear phase. However, FIR filter suffers more computational complexity as compared to IIR filters for equivalent magnitude response. Thus an approach is proposed to reduce the complexity of the FIR filters by ensuring that the frequency response of quantized filter closely matches with the unquantized frequency response (in magnitude and phase both). The method developed here take towards the approach of quantizing the cascaded sections so that the finite word length compensation in one section do not effects the other section. In multiplier-less filter, all mathematical operations are represented by shifts and additions that can be achieved by reducing the number of non-zero bits in every multiplier coefficient to a very small number. This simple method is called as 'compensating zeros'. Several other techniques have also been proposed to improve the efficiency of FIR filter in terms of the requirement of computations. Some of which include IFIR technique that reduce the use of multipliers and adders at the cost of large system delay and other technique of implementing by rounding operation for efficient FIR filters. By coefficient rounding in FIR filter, we may design multiplier-less filters which can be applied in many signal processing applications in the field of both uniform and non-uniform filter bank. Multiplier-less filter design can minimize the reduction in performance by several optimization techniques like genetic algorithms, simulated annealing and integer programming. However, optimization techniques are complex, require long time for process run and also not guarantees for the performance. Therefore, compensating zeros technique is an intuitive method that relinquishes unnecessary optimization by involving solutions of linear system of equations.

II. FIGURES OF MERIT

The performance of fixed point filter is evaluated using the following figure of merits [3]:

- Mean-Squared Error (MSE): It is the average of squared error i.e. difference between the magnitude frequency response of the quantized filter and the unquantized filter. It should be as low as possible.
- Hardware complexity: The hardware size is determined by the total number of logic cells used on the FPGA. In most of the FIR filter applications, the number of multipliers is excessively required when compared to IIR filter which increases the hardware complexity. Thus it is required that the coefficient of quantized filter should be indicated as sum and differences of powers of two using a minimum number of terms. CSD representation uses minimum number of nonzero terms thus the coefficient are represented using canonical signed digit (CSD) representation [4-6]. Before implementing quantized filter design into actual hardware, the complexity of hardware can be estimated from of all filter coefficients in CSD format in terms of T, total number of non-zeros terms used for representing filter coefficients. In general, for faster and smaller hardware implementation, the T is small.
- Throughput: The rate of generation of output samples, in samples per second.
- Latency: The time taken for obtaining the first filter output after applying the first filter input.
- Power consumption: The average power required for calculating one output sample.

The goal of the proposed method for design of quantized FIR filter is to achieve a small magnitude MSE while keeping other hardware perspective: hardware size, throughput and latency into consideration with low cost. To achieve small magnitude MSE and closer coefficient of quantized and unquantized filters, the number of non-zero bits used for representing coefficients of filter, T, should be high. Conversely higher

value of T makes the magnitude MSE worse. Hence, between the performance and hardware cost there is always a trade-off.

III. FILTER STRUCTURES

Filter designs has three basic structures: direct, cascade, and lattice by which it can be implemented in hardware. If the zeros of the FIR filters are not clustered but very uniformly distributed then the direct structure performs well. However, performance degrades quickly with direct structure. In general, when cascade and lattice structures are used, infinite impulse response (IIR) filters, pole-zero are more robust to quantization effects.The lattice structure cannot be employed because most of the FIR filters have linear phase i.e. the coefficients are symmetric which equals ±1. Though direct structure performs well but cascade structure is preferred because quantization of one coefficient of FIR filter affects all of the zeros of filters in direct form structure. While using cascade structure, the quantization of coefficients effects only in one cascaded section leaving the zeros in other section unaffected. Here, compensating zeros method is performed using cascade structure. However, it takes this idea of 'simple quantization' technique a step further that uniformly divides up the given T non zero terms across coefficient in the cascaded section in CSD format. Figure 1 indicates the block diagram to direct form of *h(n)* and its equivalent cascaded form. For implementation of filter in cascade structure, *h(n)* is divided into two subsections $c_1(n)$, $c_2(n)$ and gain *k*.

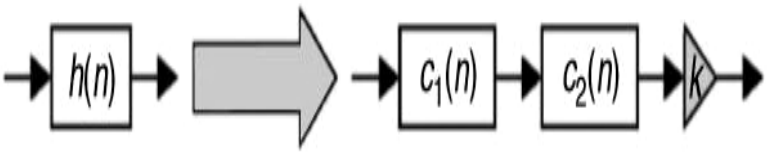

Figure 1. Direct form of *h(n)* and the equivalent cascade form using $c_1(n)$, $c_2(n)$ and *k*

The zeros of the original filter are divided into two groups by scanning the pole zero plot from $\omega = 0$ to $\omega = \pi$ in anticlockwise direction. First zero encountered, its conjugate and their reciprocals are placed in one group and next zero encountered, its conjugate and their reciprocal are placed in another group. Similarly the pole zero plot is scanned and zeros are divided into groups. The group with lesser number of zeros is considered as group 1, c_1, and group with larger number of zeros is considered as group 2, c_2 [3]. Pole zero plot and grouping of zeros in two groups is shown in figure 2 for a 31 length FIR filter.By dividing the zeros into two section using this method keeps the zeros of both the section spread out in the entire pole zero plot and thus it minimizes the quantization effect [7].

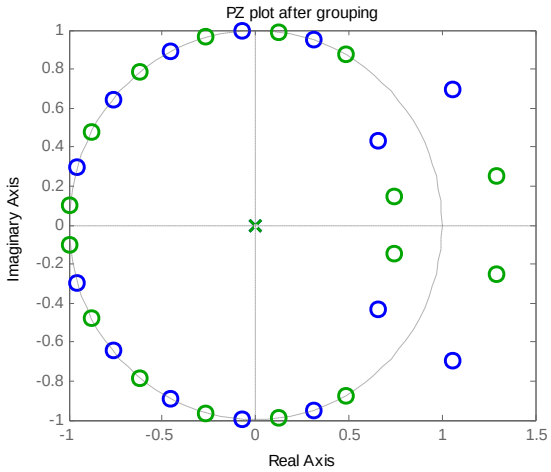

Figure 2. Pole-zero plot for 31 length FIR filter. The zeros of the filter are divided into two cascaded sections, blue zeros represent first section $c_1(n)$, and the green zeros represents second section $c_2(n)$

IV. SIMPLE QUANTIZATION

The quantizing process involves conversion of fixed point cascaded coefficient and gain to floating point coefficient and allocating T CSD terms to all the coefficients of two unquantized cascaded sections and the unquantized gain factors. All the reasonable distributions are examined while distributing a fixed number of CSD terms T to single cascade sections with *n* coefficients. Reasonable distributions should be mostly uniform i.e. all the coefficient should receive atleast one CSD terms and allocating extra terms to those coefficients that are different from their unquantized values. Choosing the closest value expressed in CSD by allocating number of terms, fixed point value of each coefficient is found. All reasonable distributions can also help in finding the one that best results in small magnitude MSE of the frequency response of the resulting system.

Table 1: Unquantized (c_1 & c_2) and simple quantized coefficients (c_1' & c_2') and their CSD representation

n	c_1	c_1'	*CSD*	*T*	*n*	c_2	c_2'	*CSD*	*T*
0,14	1.00000	1	1	1	0,16	1.00000	1	1	1
1,13	0.21382	0.375	0.10$\underline{1}$	2	1,15	0.21382	0.21875	0.0100$\underline{1}$	2
2,12	-1.04815	0.625	0.101	2	2,14	-1.04815	-1.0625	$\underline{1}$.000$\underline{1}$	2
3,11	-1.23282	0.375	0.10$\underline{1}$	2	3,13	-1.23282	-1.25	$\underline{1}$.0$\underline{1}$	2
4,10	-1.11120	1.0625	$\underline{1}$.000$\underline{1}$	2	4,12	-1.11120	-1.125	$\underline{1}$.00$\underline{1}$	2
5,9	-0.21548	1.5	10.$\underline{1}$	2	5,11	-0.21548	-0.21875	0.0$\underline{1}$001	2
6,8	0.83675	2.5	10.1	2	6,10	0.83675	0.75	1.0$\underline{1}$	2
7	1.84215	2.5	10.1	2	7,9	1.84215	1.75	10.0$\underline{1}$	2
Sub-Total				**28**	8	2.16289	2.125	10.001	2
k	0.0212	0.015625	0.000001	1	**Sub-Total**				**32**
									Total *T* = 61

To ensure each cascade section to be efficient, atleast two of the coefficients in each section must be represented in CSD format [8]; this requires coefficients $c_1(n)$ and $c_2(n)$ to be normalized so that in each section first and last section are one and also it is necessary to include gain factor k.When windowed FIR filter is applied with simple quantization method, the unquantized cascade coefficients $c_1(n)$ and $c_2(n)$ are quantized independently to the simple quantized cascade coefficient $c_1'(n)$ and $c_2'(n)$. The unquantized filter $h(n)$ is compared to simple quantized $c_1'(n)$ and $c_2'(n)$ frequency response. However, after simple quantization a linear phase response remains same but the magnitude response is significantly different from the original. Figure 3 indicates frequency response of an unquantized filter coefficient and simple quantified filter coefficient for a 31 length FIR filter.

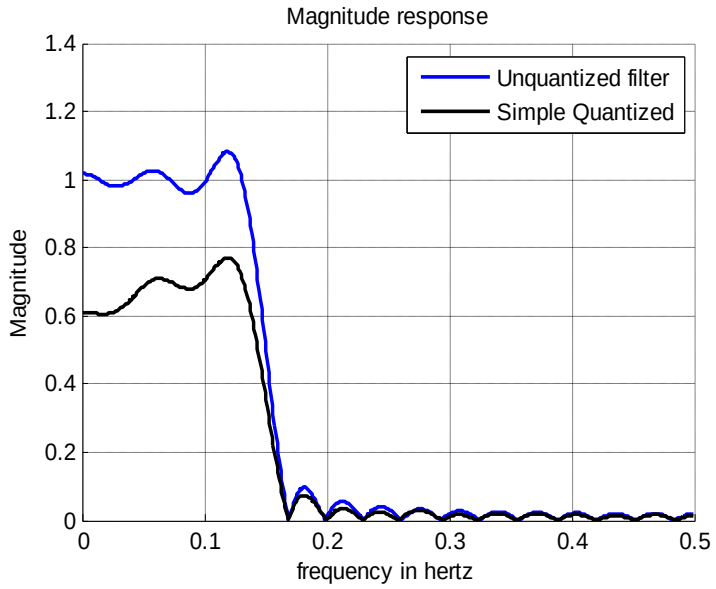

Figure 3. Frequency response of 31 length FIR filter *h(n)* in blue and frequency response after simple quantization in black

After quantization of the two cascaded section the coefficients are not exactly the same hence poles and zeros of the filter will be different from original filter. Consequently the filter response of the filter is altered. It has observed that if the zeros are clustered the sensitivity of zero location is very high due to quantization [9].

V. REFINEMENT: COMPENSATING ZEROS

The refinement to the cascaded design called 'compensating zeros' generates an alternative fixed point filter design to closely match with floating point frequency response. This quantization method estimates quantization error of the first cascade section while quantizing the other section [10].

The compensating zero quantization method starts with quantizing first cascade section $c_1(n)$ to $c'_1(n)$ and the gain factor k to k'. Now $c_2(n)$ is quantized to $c_{comp}(n)$ instead of quantizing $c_2(n)$ to $c'_2(n)$ so that when $c_1'(n)$ is cascaded with $c_{comp}(n)$, it achieves original filter $h(n)$. Here $c_{comp}(n)$ is called compensating section since it compensate the degrading performance while quantization of $c_1'(n)$.

If $c_1(n)$, $c_2(n)$, $c'_1(n)$, $c'_2(n)$ and $c_{comp}(n)$ has the transfer function $C_1(z)$, $C_2(z)$, $C'_1(z)$, $C'_2(z)$ and $C_{comp}(z)$ respectively then the unquantized cascade filter $H(z)$ is expressed as

$$H(z)=kC_1(z)C_2(z) \quad (1)$$

Where k is the gain factor. The semi-quantized filter using compensating zero method transfer function is given by [10]

$$H'_{comp}(z)=k'C'_1(z)C_{comp}(z) \quad (2)$$

The goal is to achieve unquantized and semi-quantized transfer function to be equal, thus [10]

$$H(z)=H'_{comp}(z) \quad \text{i.e.}$$

$$k'C'_1(z)C_{comp}(z)=kC_1(z)C_2(z) \quad (3)$$

In (3), on the right hand side k, $C_1(z)$ and $C_2(z)$ are known unquantized filters. On the left hand side k' and $C_1(z)$ are known after quantizing k and $C_1(z)$. Thus $c_{comp}(n)$ can be obtained by solving (3) for M unknown frequencies [11], where M is the number of unknown terms in the $c_{comp}(n)$. For a 17 length compensating section the number of unknown terms are 8, since coefficients are symmetric and one of the coefficient is 1.

The coefficients obtained after solving (3) i.e. $c_{comp}(n)$ are quantized and represented in CSD terms using same number of T used for representation of $c_2'(n)$ i.e. 32. The frequency response of the filter obtained after replacing the second quantized cascaded section $c'_2(n)$ with compensating zero section $c'_{comp}(n)$ is shown in figure 4. The frequency response of this designed filter closely matches with the frequency response of original filter. Figure 5 shows the comparison between pole zero plot of the original filter and compensating zeros quantized method. The zeros of first section have been shifted from its original position by quantization of coefficients and this shifting is being compensated by zeros of second section i.e. compensating zeros section by shifting the zeros of second cascade section in opposite direction. As the name of the method depicts 'compensating zeros', the frequency response of the quantized filter is compensated to match with frequency response of original desired filter by compensating the zeros of the second section.

Table 2: Unquantized Compensating zeros coefficients (c_{comp}), its quantized coefficients (c_{comp}') and its CSD representation

n	c_{comp}	c_{comp}'	***CSD***	***T***
0,16	1.00000	1	1	1
1,15	1.33116	1.25	1.01	2
2,14	-2.25159	-2.25	10.01	2
3,13	-1.89096	-1.875	10.001	2
4,12	-0.89474	-0.875	1.001	2
5,11	-0.83662	-0.75	1.01	2
6,10	1.48971	1.5	10.1	2
7,9	2.52829	2.5	10.1	2
8	2.79878	2.75	100.01	2
			Sub-Total	**32**

The MSE of simple quantized method 3.088e-2 has been reduced to 1.079e-3 for compensating zeros method.

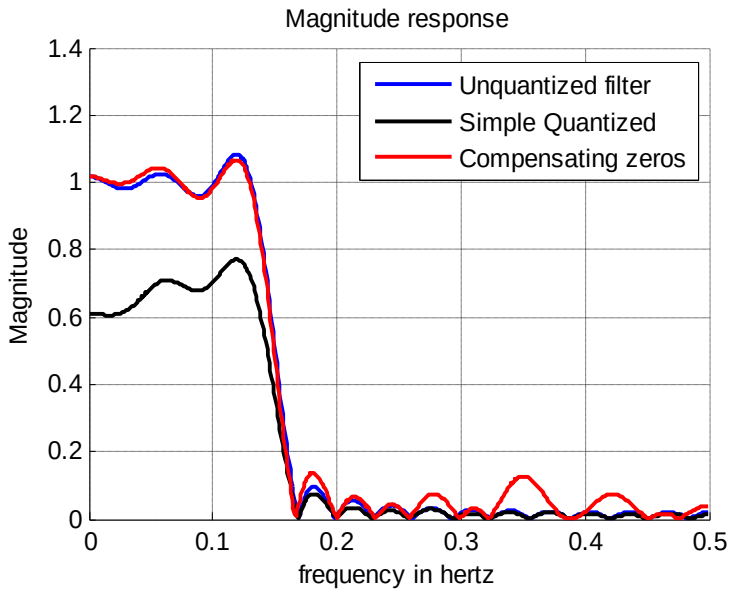

Figure 4. Frequency response of 31 length FIR filter *h(n)* in blue, frequency response after simple quantization in black and frequency response after compensating zero quantization in red.

VI. CONCLUSION

The effect of quantization of first cascade section due to approximation of coefficients is compensated by the second cascaded section by using compensating zeros method. This method can be used to quantize any FIR filter and the method does not require optimization as other methods. The result of this method is obtained by solving a linear system of equations. Moreover, compensating zeros quantization ensures that the frequency response of designed quantized filter closely matches with the frequency response of original desired filter both in magnitude and phase. The MSE for compensating zeros quantization has been reduced than the MSE for simple quantization and other, the hardware requirement for real world implementation of compensating zeros quantized filter is smaller and faster than the simple quantized filter.

REFERENCES

[1]. J.G. Proakis and D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Application*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1995, pp. 500–652.

[2]. A.V. Oppenheim, R.W. Schafer, and J.R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1999, pp. 366–510.

[3]. K.A. Kotteri, A.E. Bell, and J.E. Carletta, "Quantized FIR Filter Design: A Collaborative Project for Digital Signal Processing and Digital Design Courses," in *Proc. American Society for Engineering Education (ASEE) Annual Conf.*, Salt Lake City, Utah, June 2004.

[4]. C. Lim, R. Yang, D. Li, and J. Song, "Signed power-of-two term allocation scheme for the design of digital filters," *IEEE Trans. Circuits Syst. II*, vol. 46, no. 5, pp. 577–584, May 1999.

[5]. R.M. Hewlitt, and E.S. Swartzlander, "Canonical signed digit representation for FIR digital filters," *IEEE Workshop on Signal Processing Systems*, pp. 416-426, Oct. 2000.

[6]. R. Guo, and L.S. DeBrunner, "A Novel Fast Canonical-Signed-Digit Conversion Technique for Multiplication," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Prague, pp. 1637-1640, May 2011.

[7]. G. D'Antona, and A. Ferrero, *Digital Signal Processing for Measurement Systems: Theory and Applications*, USA: Springer, 2006, pp. 179-220.

[8]. R.G.Lyons, *Streamlining Digital Signal Processing: A Tricks of the Trade Guidebook*, 2nd ed. Hoboken, NJ: Wiley-IEEE Press, 2012, pp. 11-24.

[9]. T.A.C.M. Claasen, W.F.G. Mecklenbrauker, and J.B.H. Peek, "Effects of Quantization and Overflow in Recursive Digital Filters," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 6, pp. 517-529, Dec. 1976.

[10]. K.A. Kotteri, A.E. Bell, and J.E. Carletta, "Quantized FIR Filter Design Using Compensating Zeros" *IEEE Signal Processing Magazine*, vol. 20, no. 6, pp. 60-67, Nov. 2003.

[11]. IEEE Signal Processing Magazine Resources, [Online]. Available: http://www.signalprocessingsociety.org/publications/periodicals/spm/columns-resources-archive/2003-columns-resources/