Quantization effects in digital filters

ABSTRACT:

Quantization effects in digital filters can be divided into four main categories: quantization of system coefficients, errors due to A-D conversion, errors due to roundoffs in the arithmetic, and a constraint on signal level due to the requirement that overflow must be prevented in the comparison. The effects of quantization on implementations of two basic algorithms of digital filtering-the first-or second-order linear recursive difference equation, and the fast Fourier transform (FFT) – are studied in some detail. For these algorithms, the differing quantization effects of fixed point, floating point, and block floating point arithmetic are examined and compared. The ideas developed in the study of simple recursive filters and the FFT are applied to analyze the effects of coefficient quantization, roundoff noise, and the overflow constraint in two more complicated types of digital filters – frequency sampling and FFT filters. Realizations of the same filter design, by

means of the frequency sampling and FFT methods, are compared on the basis of differing quantization effects. All the noise analyses in the report are based on simple statistical models for roundoff and A-D conversion errors. Experimental noise measurements testing the predictions of these models are reported, and the empirical results are generally in good agreement with the statistical predictions

INTRODUCTION:

Digital filters are widely used in modern signal-transmission systems. The first-order filters are used for extracting lower-frequency or upper-frequency signals. Quantization errors due to the finite number of binary digits in the representation of numbers are typical of digital filters.

Quantization is a representation of data samples with a certain number of bits per sample after rounding to a suitable level of precision. Quantization errors in a Digital Signal Processing (DSP) system can be introduced from three sources; one source is input quantization, a second is coefficient quantization and the third is the finite precision in the arithmetic operations.

The quantization error in the arithmetic operations can be controlled by carefully selecting the size of buffer registers according to the input word length. Quantization errors from input and filter samples are considered in this article. The effects of quantization errors and the tradeoffs required between precision and hardware resources are discussed in relation to the implementation of the DSP in Field Programmable Gate Array (FPGA).

This article is divided into three main sections; quantization effects for upconversion, quantization noise due to rounding off arithmetic and quantization effects for digital beamforming (DBF). Fixed length samples cause reduction in the filter dynamic range and gain resolution.

Quantization

In digital signal processing, quantization is the process of approximating (“ mapping”) a continuous range of values (or a very large set of possible discrete values) by a relatively small (“ finite”) set of (“ values which can still take on continuous range”) discrete symbols or integer values. For example, rounding a real number in the interval [0, 100] to an integer 0, 1, 2,...100.

In other words, quantization can be described as a mapping that represents a finite continuous interval I = [a, b] of the range of a continuous valued signal, with a single number c, which is also on that interval. For example,

rounding to the nearest integer (rounding ½ up) replaces the interval [c – . 5, c + . 5) with the number c, for integer c. After that quantization we produce a finite set of values which can be encoded by say binary techniques.

A. QUANTIZATION EFFECTS ON UPCONVERSION:

In multirate systems, upconvcersion can be achieved with oversampling and filtering techniques. For the proposed digital TIGER system, input Gaussian pulses are upsampled to produce higher order Nyquist zones. A high pass FIR filter is employed to acquire a spectral zone at the expanded band edge. In this case, higher efficiency is possible by exploiting filter symmetry. For a higher throughput rate, polyphase implementation of the FIR filters can be employed. Since signal amplification is performed in the analog domain, a high speed 14 bit DAC is used for digital to analog conversion. Finite precision causes similar effects in the input data samples and filter coefficients. Fixed word length effects on filter coefficients, filter length and dynamic range are described in the following sections.

1. Sensitivity of Filter Coefficients to Quantization

Finite precision plays a significant role in the dynamic range of filter gain and DC offset. A large number of quantization levels will decrease the quantization error; on the other hand it requires larger silicon space to implement the design. The quantization affects the input Gaussian pulse and the filter coefficients. The pole and zero maps show perturbations in Figure 1 when samples are restricted to finite word length. The filter coefficients in the lower parts are constrained to 14 bit quantized samples and the length of the filter is 100 taps. This constraint arises from the fast DAC of 14 bit width used for converting a digital signal into the analog domain. Since the

dynamic range of the quantizer is less than that of the filter coefficients, the quantized coefficients are disturbed from the unit circle. The gain of the quantized filter response is displayed in Figure 1 which is distinctly less than that for the infinite precision filter. For these simulations infinite precision representation is regarded as floating point, which provides significantly better precision than the quantization levels discussed here. The zeros around Z = -1 are responsible for passband attenuation and are less displaced. As the dynamic range of the quantizer is increased to match the filter coefficients, the signal to quantization noise ratio (SNR) improves, but at the cost of increased hardware resources. Similar results can be obtained for the input Gaussian pulse when quantized to specified fourteen bit word lengths.

Finite precision is hardware efficient since the system data width is less than the infinite precision (or floating point) case. Quantization reduces a few out of 100 coefficients to zero, which will further ameliorate the memory cell and arithmetic processing requirement. Quantization also reduces the filter gain

compared to infinite precision samples; however this reduction is acceptable as long as it remains within an attenuation limit. The fourteen bit quantizer provides more than 80dB attenuation which is better than the standard of 60dB used by many communication systems.

2. Quantization Effects on Filter Order

For direct conversion transmission, a cascaded design performs better than a single stage. This is because quantization errors are reduced with a lower filter order. Secondly a lower order design requires less logic resources.

Quantization errors vary with the length of a filter and we now study the effects of the filter order on the quantization error. A simulated result is shown in Figure 2, where quantization error is plotted against variable filter order. The quantization is performed by rounding the infinite precision samples to the closest fixed point value. The quantization error increases with increased filter order, since the highest power index in the filter polynomial is the most affected by the rounding. When the quantizer is increased with one more bit in the precision, the error is reduced by approximately 6dB as would be expected.

The lower order filter provides better dynamic range than the higher order for eight and nine bit quantizers. This fact is also evident in Figure 2. At lower filter order of fifty, accumulative quantization error is around -43dB and at higher order of 200, it is -31dB. The 12dB difference is

equivalent to two additional bits in quantization. Non-linear effects of the quantization can be reduced using a smaller filter order in the modulator. Since the cascaded design comprises a filter of lower order, compared with the single model, it introduces less quantization error than the single stage.

3. Quantization and Word length

The dynamic range of the scaled filter depends on the number of bits assigned to the quantizer. For maximum signal power, the quantizer range should be equal to the signal magnitude. An FIR filter with filter variance 2 f s and quantization noise variance 2 n s has a signal to noise ratio of

This expression can be used to estimate the appropriate word length for the FPGA

implementation. A comparison of SNR versus word precision using the above expression has been calculated and is shown in Figure 3. From this graph it is evident that for each bit added to the word length, there is approximately a six decibel improvement in the SNR. For a higher precision level, a system can still be implemented, but at the cost of increased FPGA logic resources.

B. QUANTIZATION NOISE DUE TO ROUNDING OF ARITHMETIC:

In the poly phase filter, like in any other filter, quantization has to be performed on the result of any arithmetic operation. This is because any such operation requires more bits to represent the result than is required for each of the operands. If the

Word length were always to be adjusted to store the data in full precision, this would be impractical, as there would soon be too many bits required to be stored in the available memory.

Therefore, the word length of the internal data, has to be chosen, and the result of any arithmetic operation has to be constrained back to using the quantization scheme chosen from the ones shown in the previous section, as appropriate for the given application.

The quantization operation may cause a disturbance to the result of the arithmetic operation. For normal filtering operations, such a quantization disturbance can usually be successfully considered as white noise and modeled as an additive noise source at the point of the arithmetic operation with the quantization step equal to the LSB of the internal data, . This certainly is not the case for zero-valued or constant input signals. However, modeling the

quantization has-in most cases-the purpose of determining the maximum noise disturbance in the system.

Hence, even if the additive quantization noise model gives overestimated values of the noise for very specific signals, this fact does not decrease the usefulness of the approach. After the shape of the quantization noise power spectral density (NPSD) is found, it can be used to identify regions that might cause overloading or loss of precision due to arithmetic noise shaping; also the required input signal scaling and the required internal arithmetic word length can be estimated for a given noise performance. The standard methods of estimating the maximum signal level at a given node are L1-norm (modulus of the impulse response-worst-case scenario), L2-norm (statistical mean-square), and L -norm (peak in frequency domain giving the effect of the input spectral shaping). These norms can be easily estimated for the given node from the shape of the NPSD.

The quantization noise injected at each adder and multiplier, originally spectrally flat, is shaped by the noise shaping function (NSF), , calculated from the output of the filter to the input of each of the noise sources, i. e., to the output of each of the arithmetic operators. These functions were calculated for all of the all pass filter structures are shown in Fig. 2. The shapes of the nontrivial of the NFS are shown in Fig. 3. The accumulated quantization NPSD transferred to the output, , is obtained by shaping the uniform NPSD from each of the quantization noise sources by the square of

the magnitude of the NFS corresponding to the given noise injection point and can be described by

The results show that all structures perform in a way very distinct from the other ones. Structure (a)has the best performance at dc,

half-Nyquist , and Nyquist , where the NPSD

falls toward minus infinity. Its two maxima are symmetric about and independent of the coefficient value. The peaks are distant from for small coefficient values and approaches it as the coefficient increases. Structure (b) has uniform noise spectral distribution as all the arithmetic operations are either at the filter input-then noise is shaped by the allpass characteristic of the whole filter-or at its

output. Structure (d) also has a minimum at v= 0. 25 . Its average noise power level decreases as the value of the all pass coefficient increases. Structure (c), the best from the point of view of the required guard bits, has its maximum at v= 0. 25 going toward infinity for coefficient values approaching one. This effect is a result of the denominator of the Nth-order all pass filter causing the poles of the filter to move toward the unit circle at normalized frequencies of v= 2pik/N, k= 0…. N-1 for the coefficient approaching one. If there is no counter effect of the numerator, like for the case of P1(Z) for structure (c) and for structure (a), then the function goes to infinity. Even though structure (c) goes to infinity at v= 0. 25 for alfa= 1, it has the lowest average noise power from all the structures. This structure has a big advantage in terms of the number of required guard bits and ease of cascading a number of them into higher order all pass filters. If the filter coefficients approach one, then the increase in quantization noise power could be countered with few additional bits. Using other structures would

only replace the problem of dealing with an increase in the quantization noise with the problem of having to increase the number of guard bits required to deal with an increase of the peak gains. The NPSD of the quantization noise at the output of the poly phase structure can be calculated as the sum of the NPSD at the output of all all pass filters in the filter scaled by the 1/N factor N, being the number of paths. If the filter is cascaded with another filter, the NPSD of the first one will also be shaped by the square of the magnitude of the second filter.

sources. The intention was to check the correctness of the theoretical equations by applying the white noise sources instead of quantization and by performing the quantization after addition and multiplication (rounding and truncating) to verify the shaping of the quantization noise and its level both for white input noise sources and real-life signals. The shape of the output quantization noise accumulated from all arithmetic elements for a wide-band input signal assuming, for simplicity, no correlation between the noise sources, is shown for all considered all pass structures in Fig. 4. The solid curve indicates the theoretical NSF that is very well matching the median of the quantization noise (curves lying on top of each other). The quantization noise power increase calculated for the given coefficient was 8. 5 dB for structure (a), 6 dB for structure (c), 7. 3 dB for structure (d), and 9 dB for structure (b). It is clear that the quantization “ noise” differs from the assumed white noise characteristic. However, the approximation still holds with an accuracy of around 5-10% depending on the structure of the input signal. An example of more accurate modeling of the quantization noise caused by arithmetic operations can be found in (a). The arithmetic

quantization noise certainly decreases the accuracy of the filter output. The value of the arithmetic word length has to be chosen such that the quantization noise power is smaller than the stop band attenuation of the filter and the stop band ripples. In certain cases, the design requirements have to be made more stringent to allow some unavoidable distortion due to the arithmetic word length effects. For the case of decimation filters for the based A/D converters, the quantization noise adds to the one originating from the modulator.

In such a case, each stage of the decimator has to be designed so that it filters out this noise as well. The verification of the peak gain analysis was performed by applying single-tone signals at the characteristic frequencies-where functions from Fig. 2 have their extremes-and by using wideband signals to make sure that the estimates are accurate. The experimental results confirmed the theoretical calculations. The results of the simulation for the white noise input signal of unity power are given in Fig. 8. The simulation was performed for a white noise input signal of unity power in order to have a uniform gain analysis across the whole range of frequencies. The theoretical shape of the gain is shown by a solid line that is very closely matching the median value of the signal at the test points.

C. QUANTIZATION EFFECTS ON DIGITAL BEAMFORMING:

The quantization of infinite precision samples into fixed word length degrades the phased signals. As was discussed in the previous section, the use of more levels for higher precision decreases the quantization error at the expense of larger hardware resources. For a reduced precision level, quantization error is spread to the main beams and to the grating lobes as

well. In this section we present effects of quantization on beam resolution and associated grating lobes.

1. Quantization effects on Beam Pattern

Phased signals have similar quantized effects on main beam resolution as the filter samples. However non-linearity arises in the sidelobes since the quantizer is not of adequate resolution to represent small changes that affect the sidelobe levels. In order to investigate the quantization effects, an example is presented with fixed word length delay samples. The coefficients of the time vector are quantized into four and ten bits; the increased number of bits will reduce the quantization effect. For an actual design the fixed bit width will depend on available hardware resources. The quantized beam in Figure 1 shows that a four bit fixed number does not adequately represent the beam pattern and thus introduces quantization noise. The ten bit numbers will also introduce quantization error, but at a lower level as shown in Figure 1(b). As can be seen from this simple example, the four bit quantization compromises the sidelobes at the 20dB level, while the ten bit quantization provides a reasonably faithful reconstruction of the theoretical sidelobes at this level. Therefore we conclude that for the 14 bit DAC of the proposed system, the sidelobe level will be essentially unaffected by the quantization at the -20dB level.

2. Sensitivity of Sidelobe Levels to Quantization

Quantization causes gain errors in sidelobe levels. Higher resolution in quantization introduces lower quantization error. The graph in Figure 1 shows that the four bit samples result in

a quantization error which reduces the first

sidelobe gain while producing a gain error in the second sidelobe. The quantization error changes the dynamic range of the grating lobes and degrades the adjacent beam resolution for multiple beam systems. A simulated graph is displayed in Figure 2 to demonstrate non-linear behavior of the quantizer in the sidelobe resolution.

For a lower order quantizer, the quantization step is not perfectly matched with the sidelobe levels. For the first sidelobe, the quantized resolution is less than the infinite precision case, although it approaches the floating point value with increasing quantized levels. Figure 2(a) shows that for a three bit quantizer, the first sidelobe resolution is at -18dB, while at ten bits it approaches the infinite precision value of -13. 5dB. Unlike the first sidelobe, the second sidelobe exhibits higher resolution error at a lower precision level, since the quantizer can not represent the dynamic range adequately. Again, quantization error reduces with an increase in the number of bits.

CONCLUSION:

In this paper, effect of fixed word lengths on signal upconversion, quantization noise due to round of arithmetic and quantization effects on

digital beam forming have been discussed. For the digital up conversion process, the quantization error can be described using pole/zero filter and frequency response plots. Filter resolution and stop band attenuation are degraded when quantization is introduced. For an increase in filter order, the quantization error increases as the highest order in filter polynomial is effected the most. To overcome this limitation, the number of precision levels of a quantizer can be increased, however this will require increased

logic resources for FPGA implementation. Quantization effects in phasing are more complex than in the filter quantization since finite precision degrades the side lobe resolution. For lower precision levels, the quantization error exhibits non-linear behavior in the second side lobe. The quantization error is higher for lower precision levels. In order to overcome these non-linear effects, a precision level of more than eight bits is required. Performance of the proposed digital system will be effectively unaffected by the fixed word length limitations since a system data bus of at least 14 bits is suggested.

REFERENCES:

1. A. B. Sripad and D. L. Snyder, ‘ A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White
2. P. P. Vaidyanathan, “ On coefficient-quantization and computational roundoff effects in lossless multirate filter banks.
3. Google. com