FLOATING-POINT TO FIXED-POINT CONVERSION WITH DECISION ERRORS DUE TO QUANTIZATION

Changchun Shi and Robert W. Brodersen

Berkeley Wireless Research Center, Department of EECS, University of California, Berkeley

ABSTRACT

Most existing analyses of quantization effects are given under the condition that all decision-making blocks, if exist in a system, produce identical decisions in both fixed-point and infinite-precision (IP) implementations. However, in doing floating-point to fixed-point conversion (FFC), a fixed-point design with occasional decision errors may still be an acceptable approximation of the IP system. We study the effect of this decision error, and relate its probability to the fixed-point data types. Our previous FFC methodology is then extended to include systems with possible decision errors due to quantization. The extended approach is applied to both CORDIC and BPSK transceiver.

1. INTRODUCTION

To lower hardware costs, most implementations of digital systems rely on binary fixed-point (FP) number systems—either 2's complement or unsigned-magnitude—with roundoff and truncation quantization [1-6]. Existing work studies the effect of this quantization on systems that have no decision-making blocks, a term that is to be defined in section 2, or based on the assumption that there is no decision error [2-6]. In particular, an automated infinite-precision (often also referred as floating-point) to fixed-point conversion (FFC) method has been proposed in [3] based on a perturbation theory that uses this assumption. The rigorous proof of theory is given in [2]. An implementation of the tool has shown inspiring results tested on systems when this condition applies. However, in many complicated communication and DSP systems, decision errors in a fixed-point system are acceptable as long as its probability is small; then, the system is still a fair approximation of its IP correspondence. Other FFC methods based on unguided optimization and recursive estimations without understanding of the effects of these decision errors, on the other hand, require a large number of long simulations [1]. This becomes especially time-consuming when each simulation takes minutes to hours in bit-error-rate (BER) type of estimation.

Based on a study of the types of decision making blocks and the probability of decision errors as a function of fixed-point data-types in a system, we extend the FFC method proposed in [3] to include possible decision errors. The updated FFC problem formulation looks similar to [3] with additional constraints, such that each requires one BER type of estimation for coefficient fitting—itself a well-defined task. Finally, we show two examples, BPSK transceiver with root-raised-cosine-filter and CORDIC, to support our analytical results.

2. ERROR OVER DECISION MAKING BLOCKS

2.1. Categorizing signals and blocks

Let us first categorize signals and blocks in a digital system and give the necessary definitions. Digital signal processing systems are constructed by the interconnection of functional operators such as adders and multiplexers. Quantizers in a fixed point (FP) system can be used to reduce the accuracy of some signals associated with these functional units from infinite-precision (IP) to limited-precision. These signals which are allowed to have reduced accuracy will be called *arithmetic* signals. Signals which are already discrete and are not modified by quantizers will be termed *logical* signals.

Assume each operator generates one output. Operators in an IP system can be separated into different types,

1. *Arithmetic* operator—the output is an arithmetic signal, such as adder and delay in an FIR or LMS.
2. *Logical* operator—all the inputs and outputs are logical, such as an AND gate in control logic.
3. *Decision-making* operator—some of the inputs are arithmetic and the output is logical, such as the final slicer in a communication system.

For example, the slicer operator, denote as **SL**, is a decision-making operator. It has one arithmetic signal input x and one logic signal output y. The slicer function is given as

$$y = f_{\mathbf{SL}}(x) = \begin{cases} 1, \text{if } x \geq 0 \\ -1, \text{if } x < 0 \end{cases}. \quad (1)$$

All decision-making operators can be equivalently modeled as a combination of arithmetic operator, slicers (a basic decision-making operator), and control operators as shown in Fig. 1.

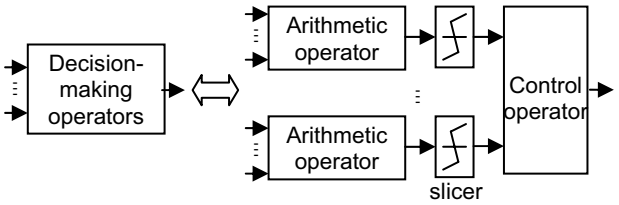

Fig. 1. Using slicers to represent any decision-making operators

For example, consider a comparator **A** with input x_1 and x_2 that outputs 1, if $x_1 > x_2$, and 0 otherwise. Its transfer function can be rewritten as,

$$f_{\mathbf{A}}(x_1, x_2) = \mathbf{NAND}(f_{\mathbf{SL}}(x_2 - x_1), 1),$$

where **NAND** is a logic operator that gives 0 if two inputs are the same, and 1 otherwise. Then, comparator **A** is equivalent to a combination of a subtractor, a slicer and a NAND operator. So, the analysis in the subsequent discussion concentrates on slicer as the typical decision-making block.

2.2. Weak and strong decision errors

Quantization errors propagating through arithmetic operators that have smooth transfer functions give a small perturbation on IP output [2]. This small perturbation, however, can propagate into possible different decisions between IP and FP system at a decision-making operator—an effect that is explicitly ignored as an assumption in [2]. We define decision differences between the corresponding decision-making operators in FP and IP systems as *decision errors* of the FP system.

Decision errors of a slicer usually happen when the magnitude of the input IP signal to a slicer is compatible to its accumulated quantization noises. Therefore, it is not acceptable to assume that decision error events at a signal node are independent to all the signals in IP system. For example, Fig. 2. shows an architecture of absolute value function. A slicer judges the sign of the input x, possibly corrupted by some quantization noise e. If the sign is positive, a following multiplexer will select the input directly; otherwise, the negated value is selected.

Without losing insight, let's assume IP signal x and quantization noise e zero-mean independent Gaussian distribution with variances σ_x^2 and σ_e^2, respectively. The actual power of output difference between IP and FP system at output can be calculated straightforward as

$$E[(|x+e|-|x|)^2] = \sigma_e^2 + \frac{4}{\pi}\sigma_x^2\left(-\frac{\sigma_e}{\sigma_x} + \frac{\pi}{2} - \tan^{-1}\left(\frac{\sigma_x}{\sigma_e}\right)\right). \quad (2)$$

When $\sigma_e << \sigma_x$, it gives $\sigma_e^2 - \frac{4\sigma_e^3}{3\pi\sigma_x}$. The first term of this approximation is the regular perturbation theory result in [2]. The second part is a higher order term on noise power, thus can be ignored. This example indicates that when an operator is continuous over its input and quantization noise, the decision-making errors will only cause higher order adjustments comparing with the perturbation result in [2] and [3]. We call this kind of decision errors *weak decision errors*, and the associated decision-making blocks are called *weak decision-making blocks*.

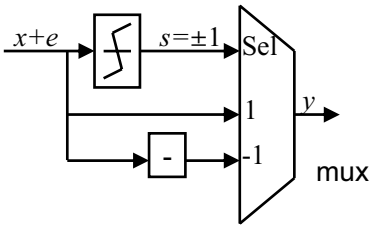

Fig. 2. An implementation of absolute value function.

We define those decision errors that have non-negligible effects on the system performance as *strong decision errors*, and their associated blocks as *strong decision blocks*. Examples of strong decision blocks include final slicers used in digital communication systems, comparators to decide which frequency offset should be used in a frequency synchronization unit, and so on. It is normally these blocks that system specifications should be set.

Quantization effects caused by strong decision errors are difficult to analyze in general. Fortunately, due to the similar difficulty of analyzing their effects in IP system in the presence of physical noise, IP system are usually designed to have only a very limited number of strong decision blocks, and they can be clearly identified by the designers. We just need to make the probability of decision error at any of these blocks much smaller than that due to physical noise or other imperfections.

2.3. Probability of decision errors due to quantization noises

Denote the IP input of a slicer as x, and the difference between FP and IP version of x as θ, that is,

$$\theta = x_{\mathbf{FP}} - x_{\mathbf{IP}} = x_{\mathbf{FP}} - x. \quad (3)$$

Furthermore, in a complicated system, both the dependence of θ and x on input signals are so mixed that it often suffices to consider them independent. In addition,

magnitude of θ is small comparing with x since we need the FP system an approximation of the IP system.

Let the probability density of x and θ be p_x and p_θ respectively. Then the FP decision may differ from IP decision according to the following formula

$$P(f_{\mathbf{SL}}(x_{\mathbf{FP}}) = -1, f_{\mathbf{SL}}(x_{\mathbf{IP}}) = 1)$$
$$= P(x + \theta < 0, x \geq 0).$$

With preceding assumption, the probability above can be written as a double integral over x and θ,

$$P(f_{\mathbf{SL}}(x_{\mathbf{FP}}) = -1, f_{\mathbf{SL}}(x_{\mathbf{IP}}) = 1)$$
$$= \iint_{\substack{x+\theta<0 \\ x\geq 0}} p_x(x)p_\theta(\theta)dxd\theta = \int_{-\infty}^{0}\left(p_\theta(\theta)\int_{0}^{-\theta} p_x(x)dx\right)d\theta.$$

Under the assumption that error magnitude is small comparing with signal x, the integral regarding $p_x(x)$ is around $p_x(0)$ in the integral, the probability becomes

$$P(f_{\mathbf{SL}}(x_{\mathbf{FP}}) = -1, f_{\mathbf{SL}}(x_{\mathbf{IP}}) = 1)$$
$$\cong p_x(0)\int_{-\infty}^{0}(-\theta \cdot p_\theta(\theta))d\theta = -p_x(0)\cdot E_\theta[\theta \cdot \mathbf{1}(\theta < 0)], \quad (4)$$

where the last step follows directly from definition of expectation value.

Similarly, the probability of error from decision of -1 in IP system to decision of 1 in FP system is given by

$$P(f_{\mathbf{SL}}(x_{\mathbf{FP}}) = -1, f_{\mathbf{SL}}(x_{\mathbf{IP}}) = 1)$$
$$= p_x(0)\cdot E_\theta[\theta \cdot \mathbf{1}(\theta \geq 0)]. \quad (5)$$

Sum (4) and (5) together, we get the probability of decision error event between IP and FP system as

$$P(f_{\mathbf{SL}}(x_{\mathbf{FP}}) \neq f_{\mathbf{SL}}(x_{\mathbf{IP}}))$$
$$= p_x(0)\cdot(E_\theta[\theta \cdot \mathbf{1}(\theta \geq 0)] + E_\theta[-\theta \cdot \mathbf{1}(\theta < 0)]).$$

The two expectations in the parenthesis can be combined together as the expectation of $E_\theta[|\theta|]$. Therefore, the proceeding equation can be written as

$$P(f_{\mathbf{SL}}(x_{\mathbf{FP}}) \neq f_{\mathbf{SL}}(x_{\mathbf{IP}})) = p_x(0)\cdot E_\theta[|\theta|]. \quad (6)$$

However, due to Cauchy-Swartz inequality,

$$E_\theta[|\theta|] \leq (E_\theta[|\theta|^2])^{1/2}.$$

So (6) can finally be written into a form

$$P(f_{\mathbf{SL}}(x_{\mathbf{FP}}) \neq f_{\mathbf{SL}}(x_{\mathbf{IP}})) = \gamma \cdot p_x(0)\cdot\sqrt{E_\theta[\theta^2]}. \quad (7)$$

where $\gamma \leq 1$. Furthermore, γ is usually between 0.7 and 1 for practical distribution of θ. For example, $\gamma = \sqrt{2/\pi} \cong 0.8$, $\gamma = \sqrt{3}/2 \cong 0.87$, and $\gamma = 1$ for cases that θ has zero-mean Gaussian distribution, zero-mean uniform distribution, and two point masses symmetric around 0, respectively.

Equation (7) shows that the decision difference between IP and FP system is proportional to the square-root of the accumulated quantization error power $E[\theta^2]$, also called mean-squared error (MSE) of θ. The coefficients may vary a little in real system depending on how well the independence assumption at the beginning of this subsection applies. This quantity has been related directly to the fixed-point data types [2-3].

3. APPLICATION IN FLOATING-POINT TO FIXED-POINT CONVERSION

The accumulated quantization noise power MSE(θ), is related to fixed-point data-types, namely fractional word-lengths $W_{Fr,1}, W_{Fr,2}, \ldots$ and quantization modes $q_1, q_2, \ldots$, are the following [2-3],

$$\text{MSE}(\theta) = \bar{\mu}^T \text{B}\bar{\mu} + \sum_{i\in\{\text{Data Path}\}} \text{C}_i 2^{-2W_{Fr,i}}, \quad (8)$$

where coefficients B is a positive semi-definite matrix, denoted as $\text{B} \succeq 0$, and $\text{C}_i \geq 0$. Vector $\bar{\mu}$ is related to fixed-point data-types deterministically as shown in [2-3]. These coefficients in (8) can be found using simulations [3] in an FFC problem with careful setups of fixed-point data types. From (7) and the discussion in previous section, using large word lengths for the setups avoid strong decision errors in the simulations; thus, the coefficients can be obtained in the same way.

Furthermore, since weak-decision-errors can be neglected in both simulation and analysis, we just need to regulate the chance of strong decision-errors. The quantization effects further caused by a strong decision-error do not affect the system performance in an avalanche effect, because the IP system is tested to be robust under physical noise. So the probability of decision error at a strong decision-making block needs to be smaller than those caused by physical noise, which usually corresponds to BER specification, that is,

$$P(f_{\mathbf{SL}}(x_{\mathbf{FP}}) \neq f_{\mathbf{SL}}(x_{\mathbf{IP}})) < \alpha \cdot BER,$$

where design parameter α is a positive guard fractional number. Substituting (7) into this inequality, we get

$$\gamma \cdot p_x(0)\cdot\sqrt{E_\theta[\theta^2]} < \alpha \cdot BER.$$

Here $E_\theta[\theta^2]$ is the same as MSE(θ) in (8) since the effects of previous strong decision errors, which happen long-time ago, have faded away. Rewriting this equation, we get

$$\text{MSE} < \left(\alpha \cdot BER \Big/ \gamma \cdot p_x(0)\right)^2. \quad (9)$$

A stronger version of (9) is by substituting the fractional number γ by 1. Furthermore, $p_x(0)$ can be directly obtained by estimating the probability of decision difference between the IP system and an otherwise identical system, but with an additive noise n of power MSE_n added at the input of the decision-making block. Denote this probability as $P(f_{\mathbf{SL}}(x_{\mathbf{IP}\text{ with }n}) \neq f_{\mathbf{SL}}(x_{\mathbf{IP}}))$, from (7), we get

$$p_x(0) = \frac{P(f_{\mathbf{SL}}(x_{\mathbf{IP}\text{ with }n}) \neq f_{\mathbf{SL}}(x_{\mathbf{IP}}))}{\gamma_n \cdot \sqrt{\mathrm{MSE}_n}}, \qquad (10)$$

where γ_n depends only on the noise shape of n, as explained shortly after (7). With (10), the right side of (9) is completely determined, denoted as A; therefore, (9) reduces to $\mathrm{MSE} - A < 0$. This condition, associated with (8), again gives a constraint function on FFC problem in exactly the same form of those showed in [3], where no decision-making blocks have been considered. Thus, with a condition for each strong decision-making block, the FFC problem is re-formulated in the same form in [3]. The only change is some additional constraint functions. One BER type estimation is needed for each of this strong decision-making blocks—a very well-defined task.

4. BPSK AND CORDIC EXAMPLES

Our first example of weak decision errors whose quantization effects can be neglected are those happened in a CORDIC system with large number of rotation stages [4-5]. In fact, the errors at CORDIC output caused by decision errors can be essentially bounded by the residue error caused by finite rotation stages—one type of architecture imperfection that vanishes as the number of stages becomes large [5]. Furthermore, these errors shown in section 2.3, happen with very small probability. These two reasons ensure that the noise power at CORDIC output can be accurately predicted regardless of the possible internal decision errors [4].

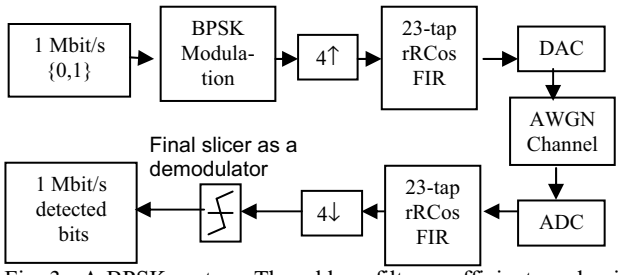

Fig. 3. A BPSK system. The adders, filter coefficients and gain output of the root-raised-cosine filters, as well as ADC, suffer quantization noises.

Second, we validate our central result (7) and (10) using the binary-phase-shift-keying (BPSK) base-band transceiver in Fig. 3. Two root-raised-cosine FIR filters, each with 23 taps, act as band limiter and matched filter, respectively [6]. The slicer, as a demodulator, makes decisions on transmitted data based on the signal polarity of its input. Fig. 4 shows that the probability of decision errors in FP system, calculated as a function of MSE of quantization noise using the (7) and (10), indeed agrees well with simulation results with various word length realizations of all the fixed-point operators in the system.

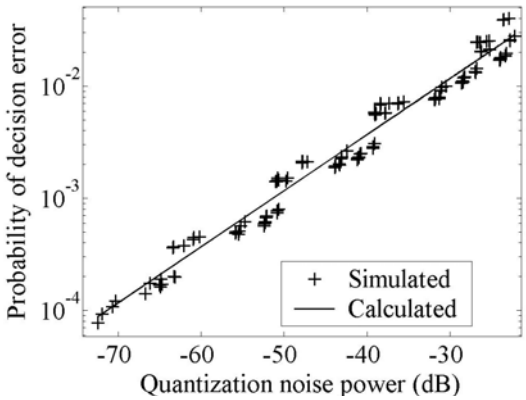

Fig. 4. Calculated curve is from (7), where γ=1 and $p_x(0)$ is obtained from (10) with one BER type estimation using an additive i.i.d. sequence {0.1, -0.1} with equal probability.

5. CONCLUSION

Two examples were given to illustrate and support our analysis of the effect and probability of a decision error. Based on the result, we have extended our previous FFC methodology to include decision making blocks and decision errors due to quantization.

6. ACKNOWLEDGEMENTS

This work was sponsored by DARPA and the SIA under the MARCO focus centers program as well as the sponsors of the Berkeley Wireless Research Center.

7. REFERENCES

[1] M. A. Cantin, Y. Savaria, and P. Lavoie, "A comparison of automatic word length optimization procedures," *IEEE Int. Sym. Circs. and Sys, 2002*, vol. 2, pp. 612 -615.

[2] C. Shi, and R. W. Brodersen, "A perturbation theory of statistical quantization effects in DSP systems with non-stationary inputs," Submitted to *Int. Sym. on Circs. Sys.* 2004.

[3] C. Shi, and R. W. Brodersen, "An automated floating-point to fixed-point conversion methodology," *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Vol. 2, pp. 529-532, April 2003.

[4] S. Y. Park, and N. I. Cho, "Fixed point error analysis of CORDIC processor based on the variance propagation," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 565-568, Apr. 2003.

[5] X. Hu, S. C. Bass, "A neglected error source in the CORDIC algorithm," *IEEE Int. Sym. on Circs. Sys.,* vol. 1, pp. 766 -769, May 1993.

[6] J. Proakis, *Digital Communication.* 4th Edition, Boston: McGraw-Hill, 2001.