17th Symposium IMEKO TC 4, 3rd Symposium IMEKO TC 19 and 15th IWADC Workshop
Instrumentation for the ICT Era
Sep. 8-10, 2010, Kosice, Slovakia

Errors in floating-point DSP calculations

Vilmos Pálfi, and István Kollár

Department of Measurement and Information Systems, Budapest University of Technology and Economics,
Email: palfi.vilmos@upcmail.hu, kollar@mit.bme.hu, phone: +36 1 463-1774, Fax: +36 1 463-4112

***Abstract*-Quantization errors like ADC errors and roundoff errors are similar in the sense that they are difficult to analyze and predict, and therefore need careful tests/measurements. This paper discusses roundoff errors: analyzes the needs and summarizes the possibilities. It deals primarily with floating-point roundoff, analyzed in the literature less than fixed-point roundoff. As an example, it illustrates the examination of the error and discusses it in one of the most common signal processing steps, the fast Fourier Transform.**

I. Introduction

In digital signal processing, quantization errors have two main sources: ADC imperfections, and roundoff in the computer, usually a digital signal processor. Therefore, to properly characterize the data processing chain, both need to be investigated. The usual problem is, however, that analytic results are rare. Most of these are of statistical nature, and they are based on the noise model of roundoff: for fixed-point, the roundoff is noise-like, with variance $LSB^2/12$, it is independent of the signal, and it is usually white. Floating-point error is less regular, but also has general approximation rules: its variance is about $\mathrm{var}\{v\} \approx 0.180 \cdot 2^{-p} \cdot \mathrm{var}\{x\}$, with p the precision and x the signal ([1], page 257, Eq. (12.24)), and it is uncorrelated with the signal, and it is white. These properties are usually all valid, but need careful experimentation because they are usually only approximations with no guarantee. However, experimentation is not easy: tools are often not available, and if available, they need to be cross-checked because even small deviations from the theory can cause noticeable errors.

II. Novelty in the paper

The paper enumerates, discusses, uses and compares simulation methods for IEEE single precision number representation (often used in DSPs). These are:

- calculations using variables of class "single" in Matlab,
- calculations using single precision variables in C,
- using a DSP simulator,
- using a DSP itself for calculation,
- using a roundoff simulation package in Matlab,
- using an object-based finite precision package in Matlab.

These possibilities offer different analysis tools which have different advantages/disadvantages. Direct calculations can only simulate IEEE single precision. This is proper, however, the sensitivity to precision (the dependence to the bit number) cannot be easily investigated, only the error of the given calculation can be evaluated. When the algorithm depends on the signal, statistical analysis is often impossible (it is very difficult to run the same algorithm with a statistically meaningful set of signals) unless statistical investigation using the same input signal is supported by the simulation package. Compiler-based methods may be sensitive to optimizations during compilation: implementation of an algorithm may well depend on the compiler, and its settings. One of our goals is to analyse what conditions must be met to produce (almost) the same roundoff errors on different platforms with the same algorithm. Effects of the different implementations on the roundoff error will be also discussed with examples.

The paper will enumerate these possibilities, analyse the error sources, and suggest solutions, using floating-point FFT on a DSP as an example.

References

[1] B. Widrow and I. Kollár (2008), "Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications," Cambridge University Press, Cambridge, UK. Home page: http://www.mit.bme.hu/books/quantization/

[2] Pálfi Vilmos, Kollár István, Roundoff Errors in Fixed-Point FFT. In: WISP'2009, IEEE International symposium on Intelligent Signal Processing. Budapest, Hungary, 2009.08.26-2009.08.28. (IEEE) pp. 87-91. Paper W1B5. http://www.mycite.omikk.bme.hu/search/docres.php?sid=login&filter=5&SCTrue=-1&SCFalse=-1&SCNull=-1&lang=1&DocumentID=74336