
**Abstract**

When a digital filter is implemented on a computer or with special-purpose hardware, errors and constraints due to finite word length are unavoidable. These quantization effects must be considered, both in deciding what register length is needed for a given filter implementation and in choosing between several possible implementations of the same filter design, which will be affected differently by quantization.

Quantization effects in digital filters can be divided into four main categories: quantization of system coefficients, errors due to analog-digital (A-D) conversion, errors due to roundoffs in the arithmetic, and a constraint on signal level due to the requirement that overflow be prevented in the computation.

The effects of these errors and constraints will vary, depending on the type of arithmetic used. Fixed point, floating point, and block floating point are three alternate types of arithmetic often employed in digital filtering.

A very large portion of the computation performed in digital filtering is composed of two basic algorithms -- the first- or second-order, linear, constant coefficient, recursive difference equation; and computation of the discrete Fourier transform (DFT) by means of the fast Fourier transform (FFT). These algorithms serve as building blocks from which the most complicated digital filtering systems can be constructed.

The effects of quantization on implementations of these basic algorithms are studied in some detail. Sensitivity formulas are presented for the effects of coefficient quantization on the poles of simple recursive filters. The mean-squared error in a computed DFT, due to coefficient quantization in the FFT, is estimated. For both recursions and the FFT, the differing effects of fixed and floating point coefficients are investigated. Statistical models for roundoff errors and A-D conversion errors, and linear system noise theory, are employed to estimate output noise variance in simple recursive filters and in the FFT. By considering the overflow constraint in conjunction with these noise analyses, output noise-to-signal ratios are derived. Noise-to-signal ratio analyses are carried out for fixed, floating, and block floating point arithmetic, and the results are compared.