LIMITED DYNAMIC RANGE OF SPECTRUM ANALYSIS DUE TO ROUNDOFF ERRORS OF THE FFT

Quang Hung Nguyén and István Kollár

Department of Measurement and Instrument Engineering
Technical University of Budapest
H-1521 Budapest, Hungary
Fax: + 36 1 166-4938, email: kollar@mmt.bme.hu

Abstract

Roundoff errors of the block-float Fast Fourier Transform (FFT) are treated. Special emphasis is given to the case when signals containing sine waves are analyzed. In the detection and analysis of sine waves, rms values and overall signal-to-noise ratios do not provide adequate information. An analysis of the maximum values is suggested, and the achievable dynamic range is given. It is shown that, in contrast to the common conviction, the dynamic range does not significantly depend on the point number of the FFT, when the roundoff errors originate dominantly from the arithmetic roundings within the FFT.

Keywords: roundoff error, Fast Fourier Transform, FFT, block-float FFT, quantization error, spectrum analysis, sine wave, dynamic range.

I. Introduction

Though floating-point signal processors are commercially available on the market, block-float FFT is still the standard means for obtaining spectra of measured signals. This means that the sample record is represented by fixed-point numbers (mantissas), and a common scale factor (exponent) is assigned to the whole block.

Frequency domain is very selective to periodic components, however, rounding errors of the FFT put an important limitation to the dynamic range, that is, to the maximum power ratio of sine waves that can be simultaneously analyzed.

Roundoff errors of the FFT have been extensively studied in the literature [1-11]. The usual approach is to assume that roundings or truncations to a given bit number can be modelled by additive independent white noises, and an rms error measure can be derived.

The above model is quite bad in the case when a dominant sine wave is present in the signal. Since the maximum value in the spectrum is proportional to $N/2$, where N is the number of samples, rescaling is necessary at almost each stage of the FFT, therefore roundings in the last stages will dominate. Moreover, the error at points different from the spectral peak, will be large, coloured, and correlated to the spectral content. Therefore, the case of a rather large sinusoidal component should be considered with special care, and results concerning the dynamic range, without using the stochastic model of the roundoff errors, are required.

II. Aspects of Error Analysis

The roundoff errors of the FFT depend on several factors. Let us briefly enumerate these.

1. FFT algorithm. There are several FFT algorithms that can be applied [12-13]. The behaviour of roundoff errors slightly depend on the algorithm actually implemented. In this paper we are going to deal with Radix-2 decimation-in-time (DIT) and decimation-in-frequency (DIF) algorithms.

2. Bit numbers in the representation of numbers. Three different bit numbers have to be considered:

a) Input bit number - this is essentially the bit number of the A/D converter. In this paper, this bit number will be denoted by b_1.

b) Bit number in the representation of the complex exponentials (b_2).

c) Bit number of the samples (b).

Each bit number is understood in this paper *without* the sign, that is, for example the samples are represented by altogether b_1+1 bits.

The bit number of the block characteristics is usually chosen large enough not to introduce number representation problems.

In the paper, all three bit numbers are addressed.

3. Record length, that is, the point number of the FFT (N throughout this paper).

4. Representation of fixed-point numbers, that is, two's complement, one's complement, sign and magnitude etc. We are going to deal with the most common two's complement representation.

5. Rounding strategy. Quantization is necessary when the bit number of any intermediate result is higher than the bit number of the target register, or an element of the block becomes too large to be represented as a fixed-point number, and the whole block has to be downscaled. Here rounding and truncation are usually applied. Before quantization a uniform or a triangular dither may be added to the numbers to be quantized.

Since downscaling is often performed when analyzing a signal which contains a relatively large sine wave, it may often happen that values with a fraction exactly equal to 0.5 have to be rounded. Therefore, the strategy how to handle these values is of importance. Such possibilities: common rounding, that is, round positive values upwards, negative values downwards (in order to keep the mean value); round randomly upwards or downwards; or convergent rounding [14] (decide on the basis of the *second least significant bit*).

In this paper we will treat common rounding, with no dither applied.

0-7803-1229-5/93/$3.00 © 1993 IEEE

6. Windowing. When a special window function is applied (Hanning, Blackman-Harris, Flat Top etc, [15-18]), the power of the sine wave is somewhat "smeared" in the frequency domain. As a consequence, the maximum value of the spectrum is lower, and less downscalings are necessary.

Because of the numerous aspects and influencing quantities and circumstances, it is extremely difficult to give generally applicable results for all cases. Also, theoretical calculations quickly become very lengthy and complex [11], so it is impossible to present them in this paper. Rather, we are going to present some interesting results for a few selected cases.

III. Dynamic Range

The dynamic range can be defined as the ratio of the amplitudes of the maximal and the minimal sine waves, simultaneously detectable and measurable in the spectrum. The minimal amplitude of a such sine wave is determined by the noise level. In our case, we will consider roundoff noise caused by the block-float FFT. A sine wave may be considered to be detectable when the peak belonging to it is significantly higher than the maximum peak of the noise. In simulations this is easy to check, however, in theoretical calculations not any more. In order to obtain a theoretical limit, the approximate standard deviation of the noise was evaluated, and a conservative 95% confidence limit of $k\cdot\mathrm{std}\{n\}$, with k=4.5 (obtained from the Chebyshev inequality) was used a a maximum peak value. The minimum detectable sine peak was taken as double of this maximum value.

IV. Theoretical and Simulation Results

1) DIF FFT, rounding, rectangular window

In this case the theoretically calculated standard deviation of the roundoff error is

$$\mathrm{std}\{n\} =$$

$$\sqrt{\frac{N}{12}\,2^{-2b_1} + \frac{11N^2+234N-1136}{72}\,2^{-2b} + \frac{N^2+108N-496}{64}\,2^{-2b_2}}\,,$$

where N is the point number, and $M=\log_2 N$.

The dynamics obtained from the above standard deviation are summarized in Table 1.

Table 1 Dynamics of DIF FFT, rounding, rectangular window

Bit numbers	b_1	15	11	15	11	11	7	7	7
	b	15	15	15	15	11	15	15	11
	b_2	15	15	11	11	11	15	11	11
FFT Point numbers N	16	76,9	70,5	57,6	57,3	52,8	47,5	45,7	46,4
	32	78,0	73,0	59,3	59,1	53,9	50,5	48,5	48,9
	64	79,0	75,2	61,1	61,0	54,7	53,5	51,1	51,1
	128	79,7	77,1	62,6	62,5	55,6	56,5	54,3	53,0
	256	80,1	78,5	63,6	63,6	56,0	59,5	56,9	54,4
	512	80,3	79,4	64,3	64,3	56,2	62,5	59,3	55,3
	1024	80,4	79,9	64,7	64,7	56,3	65,4	61,3	55,9
N=256 Simulated		81,1	79,2	62,4	61,7	56,2	59,1	58,6	56,0

2) DIF FFT, rounding, flat-top window

$$\mathrm{std}\{n\} =$$

$$4.47\cdot\sqrt{\frac{N}{12}\,2^{-2b_1} + \frac{11N^2+72N-896}{1152}\,2^{-2b} + \frac{N^2+72N-448}{2304}\,2^{-2b_2}}$$

Table 2 Dynamics of DIF FFT, rounding, flat top window

Bit numbers	b_1	15	11	15	11	11	7	7	7
	b	15	15	15	15	15	15	15	11
	b_2	15	15	11	11	11	15	11	11
FFT Point numbers N	16	77,4	56,8	61,7	55,7	53,3	32,9	32,9	32,8
	32	78,2	59,4	62,6	58,0	54,1	35,9	35,9	35,5
	64	78,8	62,0	63,9	60,3	54,7	38,9	38,9	38,0
	128	79,3	64,3	64,9	62,3	55,2	41,9	41,9	40,2
	256	79,5	66,2	65,6	63,9	55,4	44,8	44,8	42,1
	512	79,7	67,6	66,0	65,0	55,6	47,7	47,7	43,5
	1024	79,7	68,5	66,3	65,7	55,6	50,5	50,5	44,4
N=256 simulated		80,4	66,8	65,2	61,0	59,2	44,4	44,0	42,8

3) DIT FFT, rounding, rectangular window

$$\mathrm{std}\{n\} = \left[\frac{N}{12}.2^{-2b_1}+\frac{7N^2-16N}{24}.2^{-2b}+\frac{3N^2-24N}{32}.2^{-2b_2}\right]^{1/2}$$

Table 3 Dynamics of DIT FFT, rounding, rectangular window

Bit numbers	b_1	15	11	15	11	11	7	7	7
	b	15	15	15	15	11	15	15	11
	b_2	15	15	11	11	11	15	11	11
FFT Point numbers N	16	75.4	68.2	60.0	57,3	51.4	45.9	45.7	44.1
	32	74.0	69.4	58.2	59,1	49.9	48.8	48.4	45.3
	64	72.9	70.1	57.4	61,0	48.9	59.7	50.7	46.0
	128	72.1	70.2	57.0	62,5	48.0	54.5	52.7	46.1
	256	71.3	70.0	56.8	63,6	47.3	57.1	54.2	46.0
	512	70.7	69.7	56.6	64,3	46.6	59.5	55.1	45.6
	1024	70.0	69.2	56.4	64,7	46.0	61.7	55.6	45.1
N=64 Simulated		71.8	71.4	58.3	61,7	52.0	54.9	54.8	48.5

4) DIT FFT, rounding, flat-top window

$$\mathrm{std}\{n\} = 4.47\cdot\left[\frac{N}{12}\,2^{-2b_1} + \frac{7N^2-40N}{384}\,2^{-2b} + \frac{N^2-4N}{384}\,2^{-2b_2}\right]^{1/2}$$

Table 4 Dynamics of DIT FFT, rounding, flat top window

Bit numbers	b_1	15	11	15	11	11	7	7	7
	b	15	15	15	15	11	15	15	11
	b_2	15	15	11	11	11	15	11	11
FFT Point numbers N	16	75.7	56.8	60.9	55.5	51.6	32.9	32.9	32.8
	32	74.3	59.5	60.1	57.0	50.2	35.9	35.9	35.4
	64	73.2	62.0	59.5	57.9	49.1	38.9	38.9	38.0
	128	72.3	64.1	59.2	58.4	48.2	41.9	41.8	40.0
	256	71.4	65.8	59.0	58.5	47.4	44.8	44.7	41.7
	512	70.7	67.0	58.8	58.6	46.6	47.7	47.4	42.9
	1024	70.0	67.6	58.6	58.5	45.9	50.5	50.0	43.5
N=64 Simulated		76.0	62.2	58.6	56.0	50.5	40.0	39.9	39.1

The simulations well support the theoretical results in all four cases.

V. Conclusions and Practical Consequences

From the above results the following conclusions can be drawn.

a) When the DIT FFT algorithm is used, the noise spectrum is somewhat larger (and a more detailed examination shows that it is also more uniform) than that in the case of the DIF FFT. This can be explained as follows. The number of multiplications for the DIT FFT is twice of that for DIF FFT. Hence, in general, the dynamics of DIF FFT is usually better than that of DIT FFT. This means an advantage of DIF FFT versus DIT FFT.

b) The dynamics has advantageous features:

- When the DIF FFT is used, the dynamics does not decrease with the increase of N while the SNR significantly decreases. The explanation for this is that the maximal error is approximately proportional to N, and the spectrum of sine wave is directly proportional to N, so the dynamics is approximately unchanged. But the total noise power is proportional to N^3 while the power spectrum is proportional only to N^2, thus the SNR decreases when N increases.

- When DIT FFT is used, the maximal error is approximately proportional to NM, therefore the dynamics slightly decreases when N increases.

c) When the input wordlength (b_1) is much less than the arithmetic wordlength (b), the error of the input quantization dominates and the noise spectrum is more uniform. In this case the maximal error is proportional to $\sqrt{N}$, therefore the dynamics increases by 3 dB when N is doubled.

VI. Summary

Rather than dealing with rms errors, the paper presents upper bounds of the error spectra of sinusoidal input signals, and determines the achievable dynamic range of spectral analysis. It is shown and it is also illustrated by simulation results that, on the contrary to common conviction, the dynamic range only slightly depends on the number of samples, when the bit number of the block-float FFT dominates. The effect of spectral windows is also taken into account. DIT and DIF algorithms are compared from the point of view of spectral error components.

References

[1] P. D. Welch, "A fixed point fast Fourier transform error analysis," IEEE Tr*ans. Audio and Electroacoustics*, Vol. AU-17, pp. 151-157, June 1969.

[2] W. R. Knight and R. Kaiser, "A simple fixed-point error bound for the fast Fourier transform' " *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-217, No. 6, pp. 615-620, Dec. 1979.

[3] T. Kanenko and B. Liu, "Accumulation of roundoff error in fast Fourier transform," *J. ASS. Comput. Math.* pp. 637-654, Vol. 17, Oct. 1970.

[4] C. I. Weinstein, "Roundoff noise in floating point fast Fourier transform computation," *IEEE Trans. Audio Electroacoustics*, Vol. AU-17, pp. 128-142, Sept. 1969.

[5] G. U. Ramos, "Roundoff error analysis of the fast Fourier transform," *Math. Comput.*, Vol. 25, pp. 757-768, Oct. 1977.

[6] T. Thong and B. Liu, "A fixed point fast Fourier transform error analysis," *IEEE Trans. Audio Speech and Signal Processing*, Vol. ASSP-24, No. 6, pp. 563-573, Dec. 1976.

[7] P. Furon, D. Bloyet, "Error Analysis of a Floating Block FFT Processor: Model and Experiment," *Traitement du Signal*, Vol. 5, No. 1, pp. 27-40, 1988. (In French)

[8] U. Heute, "Results of a Deterministic Analysis of FFT Coefficient Errors," *Signal Processing*, Vol. 3, No. 4, pp. 321-331. 1981.

[9] R. Meyer, "Error Analysis and Comparison of FFT Implementation Structures," *Proc. ICASSP-89: 1989 International Conference on Acoustics, Speech and Signal Processing* (89CH2673-2), Glasgow, UK, May 23-26. Vol. 2, pp. 888-891.

[10] D. Calvetti, "A Stochastic Roundoff Error Analysis for the Fast Fourier Transform," *Mathematics of Computations*, Vol. 56, No. 194, pp. 755-74, April 1991.

[11] Q. H. Nguyén, "Investigation of roundoff errors in the fast Fourier transform," Ph.D. thesis, Budapest, Hungarian Academy of Sciences, 1990. (In Hungarian).

[12] E. O. Brigham, The Fast Fourier Transform. Englewood Cliffs, NJ, Prentice-Hall, 1974.

[13] H. J. Nussbaumer, Fast Fourier Transform and Convolution Algorithms. 2nd corrected and updated edition. Berlin, Heidelberg, New York, Springer-Verlag, 1982.

[14] DSP56000 Digital Signal Processor User's Manual. Motorola Corp.

[15] D. C. Rife and G. A. Vincent, "The Use of the Discrete Fourier Transform in the Measurement of Frequencies and Bands of Tones," *Bell System Technical Journal,* Vol. 49, No. 2, pp. 197-228. 1970.

[16] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proc. IEEE,* Vol. 66, No. 1, pp. 51-83, Jan. 1978.

[17] A. H. Nuttall, "Some Windows with Very Good Sidelobe Behavior," *IEEE Trans. on Acoustics, Speech and Signal Processing,* Vol. 29, No. 1, pp. 84-89. Feb. 1981.

[18] I. Kollár and F. Nagy, "On the Design and Use of FFT-Based Spectrum Analyzers," *Periodica Polytechnica Ser. Electrical Engineering,* Vol. 26, Nos 3-4, pp. 295-315. 1982.