On the Fixed-point Rounding in the DFT

Miloš Daković, Ljubiša Stanković, Budimir Lutovac, Isidora Stanković,

University of Montenegro, Faculty of Electrical Engineering
81000, Podgorica, Montenegro
{milos, ljubisa, budo, isidoras}@ac.me

***Abstract*—Rounding error in the discrete Fourier transform calculated with fast fixed-point algorithms is considered. It is shown that the variance of the rounding error depends on frequency index. Theoretically obtained results for error variance are statistically checked on decimation-in-time fast Fourier transform with two rounding methods.**

***Keywords*—DFT, fixed point arithmetics, rounding error**

I. Introduction

Algorithms for fast Fourier transform (FFT) calculation are intensively developed and analyzed [1]–[6]. Various implementations are considered. Fixed-point arithmetics implementations are very important for low-cost energy efficient devices with small computational power. Rounding error in the fixed-point FFT implementations is analyzed and presented in almost all textbooks covering digital signal processing area [2]–[5].

In most cases, the rounding error is modeled as random variable with uniform distribution. This model is correct only for rounding in the first stage where arbitrary input signal values are stored in the fixed-point registers. In the next stages rounding error probability density function is discrete and cannot be accurately modeled with uniform distribution. This implies that the results obtained by simulations are close, but not equal to the theoretically obtained results.

In this paper, we will provide rounding error variance derivation that is very close to the results obtained by statistical simulations. The main difference from the previous works on this topic [7]–[11] are assumptions that the error variance is dependent on the frequency index, and that rounding error probability density function is of the discrete nature except in input FFT stage. A similar approach is used in [12] where an approximative relation for error variance is derived. Here, we will provide exact recursive relation for error variance. Since the rounding error is almost signal independent, the error variance is used instead of the signal to quantization noise ratio.

The exact formula for rounding error variance is derived in Section II. It is shown that the error variance is frequency dependent. Presented theory is statistically checked in Section III with high agreement between theoretical and simulation results. Two rounding methods are considered and error variance bounds are derived.

This research is supported by the Montenegrin Ministry of Science, project grant CS-ICT "New ICT Compressive sensing based trends applied to: multimedia, biomedicine and communications".

II. Variance derivation

Let us consider fixed point implementation of the FFT algorithm with decimation in time. Assume that signal length is a power of two $N = 2^r$.

In order to avoid fixed-point overflows, the coefficients are divided by 2 in each stage. Single FFT butterfly is presented in Fig. 1. The FFT is calculated at stages $p = 1, \dots, r$. At each stage there is 2^{r-p} FFT blocks with 2^{p-1} butterflies in each block.

Denote with $x^{(p)}(n)$, $n = 0, 1, \dots, N-1$ outputs of the each FFT stage. Additionally, denote with $x^{(0)}(n)$ FFT inputs at the first stage. They are equal to the input signal samples with bit-reversed index. Inputs to the butterfly at stage p are

$$\begin{aligned} f &= x^{(p-1)}(m2^p + k) \\ g &= x^{(p-1)}(m2^p + 2^{p-1} + k), \end{aligned} \tag{1}$$

where $m = 0, 1, \dots, 2^{r-p} - 1$ and $k = 0, 1, \dots, 2^{p-1} - 1$. Coefficients C_p^k and S_p^k are defined as

$$\begin{aligned} C_p^k &= \frac{1}{2}\cos\left(\frac{2\pi}{2^p}k\right) \\ S_p^k &= -\frac{1}{2}\sin\left(\frac{2\pi}{2^p}k\right). \end{aligned} \tag{2}$$

The outputs are

$$\begin{aligned} F &= x^{(p)}(m2^p + k) \\ G &= x^{(p)}(m2^p + 2^{p-1} + k). \end{aligned} \tag{3}$$

Relations between inputs and outputs are

$$\begin{aligned} \Re[F] &= \frac{1}{2}\Re[f] + C_p^k\Re[g] - S_p^k\Im[g] \\ \Im[F] &= \frac{1}{2}\Im[f] + S_p^k\Re[g] + C_p^k\Im[g] \\ \Re[G] &= \frac{1}{2}\Re[f] - C_p^k\Re[g] + S_p^k\Im[g] \\ \Im[G] &= \frac{1}{2}\Im[f] - S_p^k\Re[g] - C_p^k\Im[g]. \end{aligned} \tag{4}$$

Let us now analyze discretization error propagation in a single butterfly. Additions and subtractions do not cause

978-1-5090-3843-5/17/$31.00

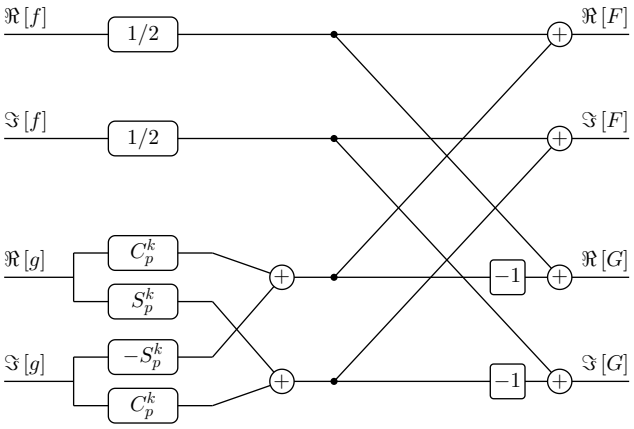

Fig. 1. Fixed point FFT butterfly

discretization error. The discretization error is present in a division by 2. Denote variance of this error as σ_h^2. Multiplication with coefficients C_p^k and S_p^k also cause discretization error with variance σ_{cs}^2.

Now we can write

$$\begin{aligned}
\sigma^2_{\Re[F]} &= \frac{1}{4}\sigma^2_{\Re[f]} + (C_p^k)^2\sigma^2_{\Re[g]} + (S_p^k)^2\sigma^2_{\Im[g]} + \sigma_h^2 + 2\sigma_{cs}^2 \\
\sigma^2_{\Im[F]} &= \frac{1}{4}\sigma^2_{\Im[f]} + (S_p^k)^2\sigma^2_{\Re[g]} + (C_p^k)^2\sigma^2_{\Im[g]} + \sigma_h^2 + 2\sigma_{cs}^2 \\
\sigma^2_{\Re[G]} &= \frac{1}{4}\sigma^2_{\Re[f]} + (C_p^k)^2\sigma^2_{\Re[g]} + (S_p^k)^2\sigma^2_{\Im[g]} + \sigma_h^2 + 2\sigma_{cs}^2 \\
\sigma^2_{\Im[G]} &= \frac{1}{4}\sigma^2_{\Im[f]} + (S_p^k)^2\sigma^2_{\Re[g]} + (C_p^k)^2\sigma^2_{\Im[g]} + \sigma_h^2 + 2\sigma_{cs}^2,
\end{aligned} \tag{5}$$

or

$$\begin{aligned}
\sigma_F^2 &= \frac{1}{4}\sigma_f^2 + \left((C_p^k)^2 + (S_p^k)^2\right)\sigma_g^2 + 2\sigma_h^2 + 4\sigma_{cs}^2 \\
\sigma_G^2 &= \frac{1}{4}\sigma_f^2 + \left((C_p^k)^2 + (S_p^k)^2\right)\sigma_g^2 + 2\sigma_h^2 + 4\sigma_{cs}^2.
\end{aligned} \tag{6}$$

Having in mind that coefficients C_p^k and S_p^k satisfy

$$(C_p^k)^2 + (S_p^k)^2 = 1/4, \tag{7}$$

we can write

$$\begin{aligned}
\sigma_F^2 &= \frac{1}{4}\sigma_f^2 + \frac{1}{4}\sigma_g^2 + 2\sigma_h^2 + 4\sigma_{cs}^2 \\
\sigma_G^2 &= \frac{1}{4}\sigma_f^2 + \frac{1}{4}\sigma_g^2 + 2\sigma_h^2 + 4\sigma_{cs}^2.
\end{aligned} \tag{8}$$

The above relations are derived having in mind that coefficients C_p^k and S_p^k does not have special properties. They are correct for each k except for $k = 0$ and for $k = 2^{p-2}$. Let us analyze these special cases. For $k = 0$ we have $C_p^0 = 1/2$ and $S_p^0 = 0$ resulting in output variances

$$\begin{aligned}
\sigma^2_{\Re[F]} &= \frac{1}{4}\sigma^2_{\Re[f]} + \frac{1}{4}\sigma^2_{\Re[g]} + 2\sigma_h^2 \\
\sigma^2_{\Im[F]} &= \frac{1}{4}\sigma^2_{\Im[f]} + \frac{1}{4}\sigma^2_{\Im[g]} + 2\sigma_h^2 \\
\sigma^2_{\Re[G]} &= \frac{1}{4}\sigma^2_{\Re[f]} + \frac{1}{4}\sigma^2_{\Re[g]} + 2\sigma_h^2 \\
\sigma^2_{\Im[G]} &= \frac{1}{4}\sigma^2_{\Im[f]} + \frac{1}{4}\sigma^2_{\Im[g]} + 2\sigma_h^2,
\end{aligned} \tag{9}$$

or

$$\begin{aligned}
\sigma_F^2 &= \frac{1}{4}\sigma_f^2 + \frac{1}{4}\sigma_g^2 + 4\sigma_h^2 \\
\sigma_G^2 &= \frac{1}{4}\sigma_f^2 + \frac{1}{4}\sigma_g^2 + 4\sigma_h^2.
\end{aligned} \tag{10}$$

The same variances are obtained for $k = 2^{p-2}$ when $C_p^k = 0$ and $S_p^k = -1/2$.

Now we can write recursive formula for discretization error variance at each stage as

$$\sigma_p^2(k) = \frac{1}{4}\sigma_{p-1}^2(k) + \frac{1}{4}\sigma_{p-1}^2(2^{p-1} + k) + \delta_{k,p} \tag{11}$$

$$\delta_{k,p} = \begin{cases} 4\sigma_h^2 & \text{for } k = 0 \text{ or } k = 2^{p-2} \\ 2\sigma_h^2 + 4\sigma_{cs}^2 & \text{otherwise,} \end{cases} \tag{12}$$

where $p = 1, 2, \ldots, r$ and $k = 0, 1, \ldots, 2^{p-1} - 1$. Note that each $\sigma_p^2(k)$ is periodic in k with period 2^{p-1}.

Discretization error at input stage is caused by simple rounding input values and it will be denoted with σ_r^2 producing initial conditions for the previous recurrence in the form

$$\sigma_0^2(k) = \sigma_r^2, \tag{13}$$

for each $k = 0, 1, \ldots, N - 1$.

III. Variance analysis and verification

Results given in the previous section will be analyzed and verified through examples. In the first example, we will assume that each division by 2 is performed by binary shift-right of the magnitude. In this case, LSB of the input value is discarded, and the result of the division is always rounded toward zero.

In the second example, we will assume that division by 2 is performed with random tie-breaking. In this case, if the LSB of the input value is equal to one the result is rounded up or down with equal probabilities.

A. Example 1

Assume that real and imaginary parts of the input signal are limited to range $(-1, 1)$ and that all operations are performed with $b + 1$ bits precision. The discretization step is $\Delta = 2^{-b}$.

Consider a case when division by 2 is performed by the binary shift-right operation. If negative numbers are saved in sign-magnitude representation, then the discretization variance σ_h^2 will be

$$\sigma_h^2 = \frac{1}{16}\Delta^2 = 2^{-2b-4}. \tag{14}$$

Note that this rounding is biased with bias equal to $\Delta/4$ for positive and $-\Delta/4$ for negative numbers.

Error caused by discretization after signal multiplication with non-integer factors is uniformly distributed over the interval from $-\Delta/2$ to $\Delta/2$ producing variance

$$\sigma_{cs}^2 = \frac{1}{12}\Delta^2 = \frac{1}{3}2^{-2b-2}. \tag{15}$$

This rounding is unbiased.

By replacing these values into (11) and (12), we obtain recursive relation

$$\sigma_p^2(k) = \frac{1}{4}\sigma_{p-1}^2(k) + \frac{1}{4}\sigma_{p-1}^2(2^{p-1}+k) + \delta_{k,p} \tag{16}$$

$$\delta_{k,p} = \begin{cases} \frac{1}{4}\Delta^2 & \text{for } k=0 \text{ or } k=2^{p-2} \\ \frac{11}{24}\Delta^2 & \text{otherwise,} \end{cases} \tag{17}$$

where $p = 1, 2, \ldots, r$ and $k = 0, 1, \ldots, 2^{p-1}-1$.

Discretization error at input stage is caused by simple rounding input values and it is equal to

$$\sigma_0^2(k) = \frac{1}{6}\Delta^2, \tag{18}$$

for each $k = 0, 1, \ldots, N-1$.

The lowest variance is obtained when, for the considered sample, we have $k = 0$ or $k = 2^{p-2}$ at each stage. It is equal to

$$\begin{aligned}\sigma_{min}^2 &= 2^{-r}\frac{\Delta^2}{6} + \frac{\Delta^2}{4}\left(1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{r-1}}\right) \\ &= \Delta^2\left(\frac{1}{2} - \frac{1}{3}2^{-r}\right).\end{aligned} \tag{19}$$

This variance is obtained for frequency index $k = 0$, $k = N/4$, $k = N/2$, and $k = 3N/4$.

The highest variance is obtained when special cases $k = 0$ or $k = 2^{p-2}$ in (17) are avoided whenever it is possible. Note that this is not possible for $p = 1$ or $p = 2$. We get

$$\begin{aligned}\sigma_{max}^2 &= 2^{-r}\frac{\Delta^2}{6} + \frac{11}{24}\Delta^2\left(1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{r-3}}\right) \\ &+ \frac{1}{4}\Delta^2\left(\frac{1}{2^{r-2}} + \frac{1}{2^{r-1}}\right) = \Delta^2\left(\frac{11}{12} + \frac{23}{16}2^{-r}\right).\end{aligned} \tag{20}$$

This variance is obtained for every odd frequency index k.

For large $N = 2^r$ we can approximate output variance bounds as

$$\frac{1}{2}\Delta^2 \le \sigma^2 \le \frac{11}{12}\Delta^2. \tag{21}$$

The results are presented in Fig. 2 for $N = 32, 64, 128$. Calculations are performed with $b = 12$ bits precision. Statistical results are averaged over 5000 realizations of the complex-valued random input signal with uniformly distributed real and imaginary parts from -1 to 1. The results are compared with statistically estimated error variances. It can be concluded that theoretically derived variances are very close to the statistically obtained results.

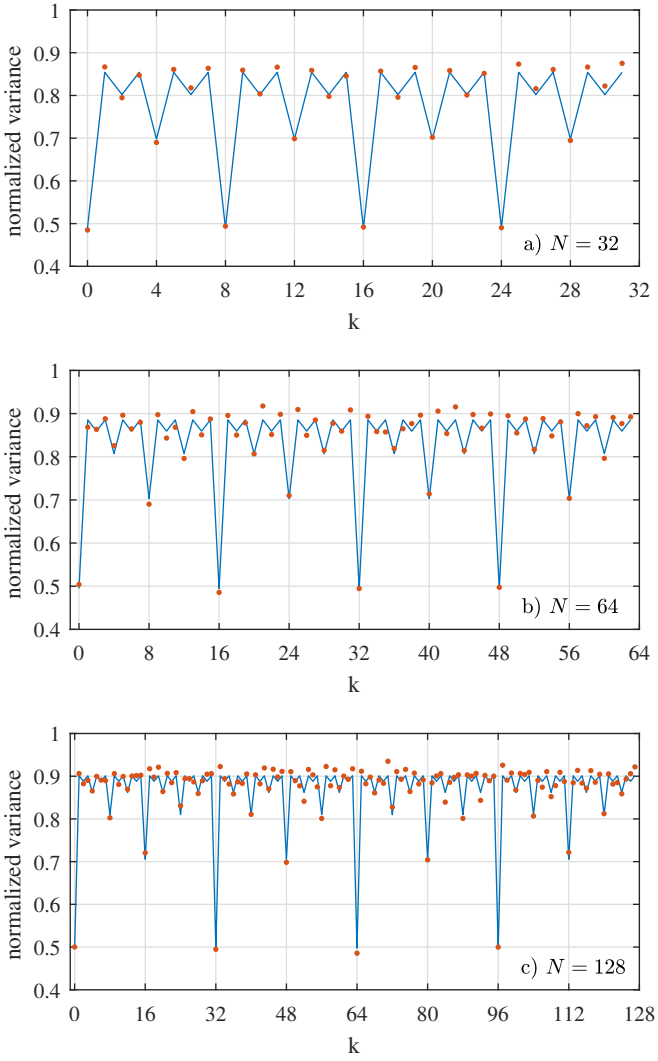

Fig. 2. Fixed point FFT normalized discretization variance for a) $N = 32$, b) $N = 64$ and c) $N = 128$. Rounding toward zero is used. Theoretical results are presented with line and statistical results obtained by averaging over 5000 realizations by dots. The variance is normalized with Δ^2.

B. Example 2

Here we will consider situation when division by 2 is performed with random tie-breaking, i.e., the situation when the discretization error, caused by division by 2, is discrete with possible values $-\Delta/2$, 0, and $\Delta/2$ that occurs with probabilities 0.25, 0.5, and 0.25 respectively. In this case the discretization variance σ_h^2 will be

$$\sigma_h^2 = \frac{1}{8}\Delta^2 = 2^{-2b-3}. \tag{22}$$

Note that this rounding is unbiased.

Variance σ_{cs}^2 is same as in the previous example.

By replacing these values into (11) and (12) we obtain

$$\sigma_p^2(k) = \frac{1}{4}\sigma_{p-1}^2(k) + \frac{1}{4}\sigma_{p-1}^2(2^{p-1}+k) + \delta_{k,p} \tag{23}$$

$$\delta_{k,p} = \begin{cases} \frac{1}{2}\Delta^2 & \text{for } k = 0 \text{ or } k = 2^{p-2} \\ \frac{7}{12}\Delta^2 & \text{otherwise,} \end{cases} \quad (24)$$

where $p = 1, 2, \ldots, r$ and $k = 0, 1, \ldots, 2^{p-1} - 1$.

The lowest variance is obtained as

$$\sigma^2_{min} = 2^{-r}\frac{\Delta^2}{6} + \frac{\Delta^2}{2}\left(1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{r-1}}\right) = \Delta^2\left(1 - \frac{5}{6}2^{-r}\right). \quad (25)$$

This variance is obtained for frequency index $k = 0$, $k = N/4$, $k = N/2$, and $k = 3N/4$.

The highest variance is

$$\sigma^2_{max} = 2^{-r}\frac{\Delta^2}{6} + \frac{7}{12}\Delta^2\left(1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{r-3}}\right) + \frac{1}{2}\Delta^2\left(\frac{1}{2^{r-2}} + \frac{1}{2^{r-1}}\right) = \Delta^2\left(\frac{7}{6} + \frac{23}{8}2^{-r}\right). \quad (26)$$

This variance is obtained for every odd frequency index k.

For large $N = 2^r$ we can approximate output variance bounds as

$$\Delta^2 \leq \sigma^2 \leq \frac{7}{6}\Delta^2. \quad (27)$$

Signal to quantization noise ratio (SQNR) bounds are $10.79\,\text{dB} < SQNR < 11.46\,\text{dB}$.

The results are presented in Fig. 3 for $N = 32, 64, 128$. The results are compared with statistically estimated error variances. It can be concluded that theoretically derived variances are very close to the statistically obtained results.

IV. Conclusion

In this paper, we derive recursive formula for discretization error variance for common decimation-in-time FFT algorithm. It is shown that discretization error depends on frequency index k. Upper and lower bounds are derived as well. The formula is statistically checked for two rounding methods and high accuracy is obtained.

Future work on this topic could include other FFT algorithms and various rounding procedures.

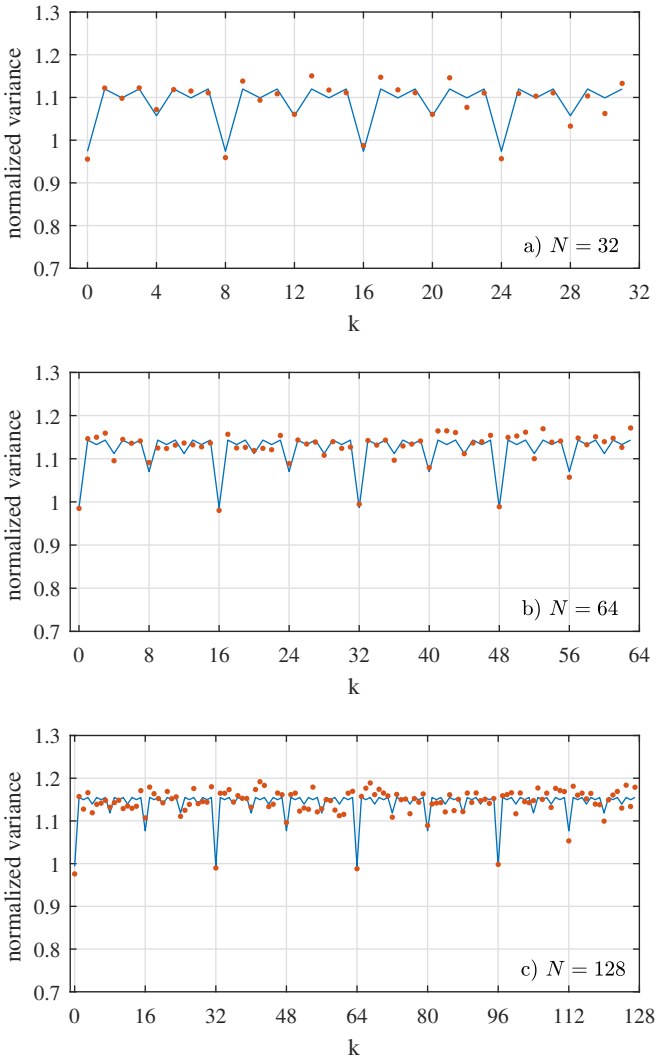

Fig. 3. Fixed point FFT normalized discretization variance for a) $N = 32$, b) $N = 64$ and c) $N = 128$. Random tie-break rounding is used. Theoretical results are presented with line and statistical results obtained by averaging over 5000 realizations by dots. The variance is normalized with Δ^2.

References

[1] J. W. Cooley, J. Tukey, "An algorithm for the machine calculation of complex Fourier series," Math. Comput., vol. 19, 1965, pp. 297—301

[2] LJ. Stanković, *Digital Signal Processing with Selected Topics*, CreateSpace Independent Publishing Platform, An Amazon.com Company, November 4, 2015

[3] A. V. Oppenheim, R. W. Schafer, J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Englewood Cliffs, New York: Prentice-Hall, 1998

[4] S. K. Mitra, Y. Kuo, *Digital signal processing: a computer-based approach*, New York: McGraw-Hill, 2006

[5] J. G. Proakis, D. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, 4th ed, Pearson, 2006

[6] H. J. Nussbaumer, *Fast Fourier transform and convolution algorithms*, Springer Science & Business Media, 2012

[7] P. Welch, "A fixed-point fast Fourier transform error analysis," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 2, pp. 151–157, 1969. doi: 10.1109/TAU.1969.1162035

[8] W. Knight and R. Kaiser, "A simple fixed-point error bound for the fast Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 6, 1979, pp. 615-620. doi: 10.1109/TASSP.1979.1163314

[9] W. H. Chang and T. Q. Nguyen, "On the Fixed-Point Accuracy Analysis of FFT Algorithms," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4673-4682, Oct. 2008. doi: 10.1109/TSP.2008.924637

[10] O. Sarbishei and K. Radecka, "Analysis of Mean-Square-Error (MSE) for fixed-point FFT units," *IEEE International Symposium of Circuits and Systems (ISCAS)*, Rio de Janeiro, 2011, pp. 1732-1735, doi: 10.1109/ISCAS.2011.5937917

[11] P. Gupta, "Accurate performance analysis of a fixed point FFT," *IEEE Twenty Second National Conference on Communication (NCC)*, Guwahati, India, 2016, Mar 4, pp. 1-6, doi: 10.1109/NCC.2016.7561147

[12] R. Meyer, "Error analysis and comparison of FFT implementation structures," *International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, 1989, pp. 888-891, doi: 10.1109/ICASSP.1989.266571