FFT-BASED SPECTRUM ANALYSIS FROM THE POINT OF VIEW OF HARDWARE REALIZATION

G. HORVÁTH

Department of Measurement and Instrumentation Engineering,
Technical University, H-1521 Budapest

Summary

This paper deals with the FFT-based methods which can be used to estimate the power density spectrum of stationary stochastic signals. It reviews some errors of the digital spectrum estimation, especially those due to finite word length representation, and some suggestions are presented to reduce the effects of limited word length in hardware realizations.

Introduction

In analysing stationary stochastic signals one of the most essential and most frequently applied methods is the power density spectrum analysis.

Many methods have been developed to analyze the spectrum, but all the digital methods can be regarded as different versions only of two basically important spectrum estimation procedures.

— In the Blackman—Tukey (B—T) method first the autocorrelation function is calculated using the input data points, then the power density spectrum estimator can be computed as the discrete Fourier transform (DFT) of the autocorrelation function.

$$\hat{S}_x(k) = \sum_l \hat{R}_x(l) \cdot e^{-j2\pi lk/N} \tag{1}$$

where $\hat{R}_x(l)$ is some estimator of the autocorrelation function. The so-called sample autocorrelation function is a possible estimator of $R_x(l)$. Using the input data points $\{x_n\}_{n=0}^{N-1}$

$$\hat{R}_x(l) = \frac{1}{N} \sum_{n=0}^{N-l-1} x_n \cdot x_{n+l} \tag{2}$$

— Using the direct Fourier transform method the first step is to compute the DFT of a given time-limited data record, then the spectrum can be estimated by

$$\hat{S}_x(k) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x_n \cdot e^{-j2\pi nk/N} \right|^2 \tag{3}$$

where $\{x_n\}_{n=0}^{N-1}$ represents again the input data points.

This latter estimator is often called a periodogram. The importance of the direct method has increased since the development of the fast Fourier transform (FFT) algorithm. (The application of the FFT algorithm greatly reduces the calculation time, so in the B—T method it is often applied for the estimation of the correlation function as well. In this case the estimator of the correlation function can be determined calculating the inverse FFT of a preliminary spectrum estimator. The preliminary estimator can be calculated using periodograms. The lag-windowed correlation function is then transformed back to the frequency domain [1]).

The various methods differ concerning the speed, the amount of calculations or the properties of the result. The direct method needs less computation but its drawback is that using one periodogram as an estimator, the variance is very large. As an example, if the data points come from a Gaussian process it can be shown [2] that var $\{\hat{S}_x(k)\} \cong S_x^2(k)$, where $S_x(k)$ is the true spectrum. To improve the properties of the estimators various modifications of the two essential methods have been developed. Different windowing techniques in the time-, lag-, or frequency domain or averaging over the periodograms are the most important possibilities [3], [4]. Whichever method is used, the estimator will never be error-free. The errors are partly due to the limited length data record and due partly to the finite word length number representation. In the following a short review of the effects of the finite word length is given.

The effects of the finite word length

In digital signal processing whether we use software or hardware means, the data are represented by a finite number of bits. The software implementation mostly uses floating-point numbers whereas in hardware solutions usually fixed-point numbers are preferred. To increase the computing speed or to reduce the hardware complexity needs as short word length as possible. But the limited word length, especially when fixed-point numbers are used introduces accuracy problems.

To compute the spectrum estimator the tasks to be performed are:

— data collection,
— preprocessing (e.g. applying some data window)
— computing the Fourier coefficients, $X_N(k)$ using one of the FFT algorithms,
— smoothing in the frequency domain (optional),
— squaring and averaging.

Therefore, the major error sources are:

— A/D conversion (input data quantization),
— applying finite word length data window,
— accumulation of errors caused by rounding the arithmetic during the FFT,
— representation of the sin/cos coefficients by finite number of bits.

The analysis of the error sources is important because on the basis of the results the following questions may be answered:

— How can one take into consideration the effects of the various error sources?
— How can the appropriate A/D converter be chosen?
— Can an optimal sin/cos coefficient set be found?
— Which FFT algorithm (radix2, radix4, DIT, DIF, etc.) is the less sensitive to quantization?
— Is there any way (e.g. modification of the processing algorithm) to decrease the effects of quantization?
— etc.

The effects of the quantization of the input signal

The input data points, which are samples of a stochastic signal, can be regarded as the realizations of random variables. Quantization obviously changes the statistical parameters (expected value, variance) of these random variables. The influence of quantization can be determined by applying the quantizing theorem [5]. Using a uniform quantizer (Fig. 1) and if the input signal is a Gaussian one the expected value and the variance after quantization are approximately as follows [6]:

$$m_y \cong m_x + \frac{q}{\pi} e^{-2\pi^2\left(\frac{\sigma_x}{q}\right)^2} \sin\frac{2\pi m_x}{q} \tag{4}$$

and

$$\sigma_y^2 \cong \sigma_x^2 + \frac{q^2}{12} + \frac{2m_x q}{\pi} e^{-2\pi^2\left(\frac{\sigma_x}{q}\right)^2} \sin\frac{2\pi m_x}{q} + e^{-2\pi^2\left(\frac{\sigma_x}{q}\right)^2} \cdot \left(4\sigma_x^2 + \frac{q^2}{\pi^2}\right) \cos\frac{2\pi m_x}{q} \tag{5}$$

where

— m_x is the expected value
— σ_x is the variance
— q is the quantum size.

} of the input data before quantization

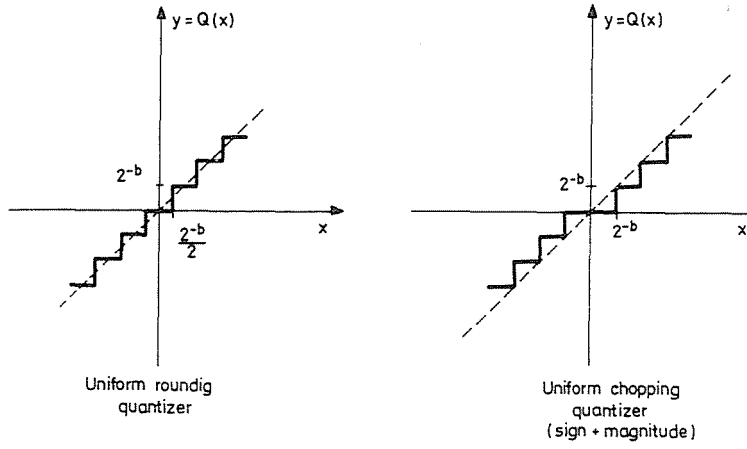

Fig. 1

Assuming that the input signal has a zero mean value, the parameters in the quantized case are:

$$m_x = m_y = 0 \tag{6}$$

$$\sigma_y^2 \cong \sigma_x^2 + \frac{q^2}{12} + \left(4\sigma_x^2 + \frac{q^2}{\pi^2}\right) e^{-2\pi^2\left(\frac{\sigma_x}{q}\right)^2} \tag{7}$$

In this case only the variance is changed. The first term in (7) is the variance of the unquantized signal. The second term $\frac{q^2}{12}$ can be regarded as the variance of an additive noise with uniform probability density function over $\left[-\frac{q}{2}, +\frac{q}{2}\right]$ which is uncorrelated with the signal. The third term is the largest term of the approximation error (other terms are neglected). This error term can also be neglected if the σ/q ratio is larger than a given value. Assuming that $\sigma/q = 1$, the sum of all error terms (including the third term in (7)) is approximately $10^{-8}\,\sigma_x^2$.

If the expected value of the input signal is not zero and if $\sigma/q = 1$, the error term in (4) is about $10^{-9}\,q$. The quantum size can therefore be choosen in such a way that the effect of quantization may be modelled by an additive, uniformly distributed noise which has no correlation with the signal. (Fig. 2) Because of the presence of this noise the spectrum estimator will be biased and its variance will be increased.

The bias and the increase of the variance can easily be determined if the following assumptions are valid:

— The input signal is a stationary Gaussian white noise process, therefore the input data points $\{x_n\}$ are mutually uncorrelated random variables with zero mean and σ_x^2 variance,
— the quantization noise is statistically independent of the signal.

Fig. 2

It can be seen, that the propagation of the signal and that of the noise through the whole FFT calculation are similar. If the input data is Gaussian, the $X_N(k)$, the DFT of the input data is also Gaussian with zero mean and with a variance proportional to $\sigma_x^2 + \frac{q^2}{12}$.

Having squared the $X_N(k)$ values, the points of the periodogram are obtained. It can be seen that

$$E\{|X_N(k)|^2\} = \text{var}\{X_N(k)\} \sim \sigma_x^2 + \frac{q^2}{12} \tag{8}$$

which means, that the bias is proportional to $\frac{q^2}{12}$ and so it can be corrected.

According to our assumptions the points of the periodogram are χ^2 distributed random variables with a variance proportional to $\sigma_x^2 + \frac{q^2}{12}$. If $\sigma_x \geq q$, the increase of the variance, because of the input quantization, is less than 17%. If the input data are not Gaussian but its probability density function is known, the quantizing theorem can be applied and the effects of quantization can be determined. But the calculation is usually rather cumbersome.

The errors of fixed-point FFTs

In the spectrum estimation the most complex operation is the FFT computation.

The effects of quantization, the propagation of the error depends on the implemented FFT algorithm, so it is necessary to study different FFT algorithms.

The subject of roundoff error in the FFT has been studied in many papers [7]—[10]. Next a review of the main point is given.

The various error analyses are based on different error models. In the most often applied model the effect of limited word length is considered as the presence of an additive, uncorrelated noise with known statistical properties. The propagation of the noise of the different error sources is taken into consideration and the mean-squared value of the output error is determined.

Welch [7] studies the most often applied radix 2 DIT FFT algorithm. In this case the FFT is an ordered set of $M=\log_2 N$ stages of $N/2$ computations termed as butterfly calculations, where N is an integer power of two, the number of the input data points. Each butterfly calculation has two points $X(i)$ and $X(j)$ as inputs. The output points are obtained through complex multiplications and additions using the complex sin/cos coefficients W^k which are the appropriate integer power of $W=e^{-j2\pi/N}$. A butterfly operation at the stage m is as follows:

$$\begin{aligned} X_{m+1}(i) &= X_m(i) + W^k \cdot X_m(j) \\ X_{m+1}(j) &= X_m(i) - W^k \cdot X_m(j) \end{aligned} \tag{9}$$

The results of the intermediate stages of the calculations $X_{m+1}(i)$ and $X_{m+1}(j)$ are used as the inputs to the next stage and the results obtained in the last stage are the transformed values. A general butterfly operation given by (9) can be expressed in terms of real arithmetics.

$$\begin{aligned} \operatorname{Re} X_{m+1}(i) &= \operatorname{Re} X_m(i) + \operatorname{Re} W^k \cdot \operatorname{Re} X_m(j) - \operatorname{Im} W^k \cdot \operatorname{Im} X_m(j) \\ \text{and} \qquad & \\ \operatorname{Im} X_{m+1}(i) &= \operatorname{Im} X_m(i) + \operatorname{Re} W^k \cdot \operatorname{Im} X_m(j) + \operatorname{Im} W^k \cdot \operatorname{Re} X_m(j) \end{aligned} \tag{10}$$

There are two errors generated during a butterfly calculation:

— the error due to rounding after every real multiplication, and

— the error caused by the right shifts which are necessary to prevent overflow at every real addition.

Using a b-bit plus sign data format the rounding of the multiplications to a b-bit number gives a uniformly distributed noise in the interval $[-2^{-b}/2, +2^{-b}/2]$. This noise has zero mean and $\Delta_1^2=2^{-2b}/12$ variance.

As a result of the right shift, a zero mean noise is generated with $\Delta_2^2=2^{-2b}/2$ variance.

Welch assumes the uncorrelatedness of the noise and one overflow occurring in every stage which needs stage-by-stage right shifts.

Recognizing that in the first two stages there are only error-free multiplications, Welch determines the variance of the noise at the output.

The increase of the variance from the m-th stage to the $(m+1)$-th stage is:

$$\text{var}\,(m+1)=2\text{var}\,(m)+4^{m+1}\Delta_1^2+4^{m+1}\Delta_2^2 \tag{11}$$

Using this relation, the mean square value of the output noise is:

$$\text{var}\,(M)\approx 8N^2\Delta_1^2=8N^2\frac{2^{-2b}}{12} \tag{12}$$

From (12), and if we calculate the propagation of the signal through the FFT, the ratio of the rms noise output to the rms signal output can be determined.

$$\frac{\text{rms (error)}}{\text{rms (signal)}}\approx\frac{\sqrt{N}2^{-b}\cdot 0.3\cdot\sqrt{8}}{\text{rms (input)}} \tag{13}$$

The error to signal ratio increases as $\sqrt{N}$ which means a 1/2 bit/stage increase.

Thong and Liu [8] extended Welch's results. They give expressions of the output error for different radix 2 algorithms using rounded or chopped arithmetic. This gives the possibility to compare the effects of rounding and chopping and to compare the DIT and the DIF algorithms.

The results show that:

— the quantization noise in the rounded case is significantly smaller than in the chopped case,

— from the point of view of roundoff errors there is no significant difference between the DIT and the DIF algorithms. When no rescaling is necessary, the DIT algorithm is more appropriate than the DIF algorithm. However, using the stage-by-stage shift method the DIF becomes superior.

This can be explained by the differing influence of error-free multiplications in the two algorithms. The error-free multiplications are at the first two stages of the DIT and at the last two stages of the DIF algorithm. For both the DIT and DIF algorithms the properties of the error propagation show that the sooner an error is introduced the more output points are affected by it. When no shift is done the lack of an error term in the first steps has a larger effect at the output than if the case when the last steps are error-free. When stage-by-stage shifts are done the errors introduced in the first steps are smaller than those in the last steps so the DIF is better than the DIT.

Remarks

1. The comparing of the two types of errors arising during the FFT calculations shows that $\Delta_2^2=6\Delta_1^2$ i.e. the error caused by the right shift is larger by almost an order of magnitude than the error due to rounding. The previous

results [7], [8] assume that either shifts are needed at every stages, or there is no shift at all. The former case gives a somewhat pessimistic result and in the latter case the result is overly optimistic.

For a more realistic error analysis we ought to estimate the total number of required shifts. Kaiser and Knight [9] give an expression which approximates the number of shifts s_M:

$$s_M \cong \frac{M}{2} + \log_2 \frac{\rho_n}{\rho_0} \leq M + 1 \tag{14}$$

where ρ_0 and ρ_n are the peak/rms ratios of the input and output signals, respectively. It is obvious that ρ_n is known a posteriori only, but sufficient information about the spectrum will often be available to permit a reasonable a priori estimate.

2. The previous investigations assumed that the different errors can be modelled as uncorrelated additive noise. The validity of this assumption can be proved using the quantizing theorem.

If the signal is Gaussian and $\sigma/q \geqq 1$ the additive noise model is true with great accuracy. However, if we study the roundoff error after the multiplications in a butterfly, we can see that this assumption is not always valid. If e.g. $\mathrm{Re}\, W \ll 1$ the variance of the multiplication $\mathrm{Re}\, W \cdot \mathrm{Re}\, X_m(i)$ will be $(\mathrm{Re}\, W)^2 \cdot \sigma_x^2$. Its effect is as if the quantum size were $1/\mathrm{Re}\, W$ times as large as the original one. In the case of a 256 point spectrum analysis the minimal value of W is approximately 0.025, which means that the relative quantum size at the output of the butterfly can be 40 times the input. In this case the additive noise model is not true.

3. Estimating the power density spectrum we are interested not only in the errors at the output of an FFT, but the errors of the spectrum points, too.

Assuming that the signal and the noise at the output of the FFT are uncorrelated and if the output data points are Gaussians, the bias and the variance of the power spectrum can be calculated similarly to the case of the input quantization. Since the peak-to-rms ratio of the output is known a posteriori, the bias can be corrected.

Effects of coefficient quantization

Error is also caused by the finite word length representation of the sin/cos coefficients. There are rather few results dealing with this error [9], [10]. The nature of coefficient quantization is inherently nonstatistical. The same inexact values are used repeatedly in the computation of one spectrum point, so

accurate results from a stochastic analysis cannot be obtained. In spite of this fact, Oppenheim and Weinstein [10] have obtained some useful results by means of rough statistical analysis. The error due to the coefficient quantization is considered again as an uncorrelated noise with uniform distribution between plus and minus $2^{-b}/2$.

They determine the ratio of the mean-square output noise to mean-square output signal

$$\frac{\text{ms (coeff error)}}{\text{ms (signal)}} \cong \frac{2^{-b}}{12} \cdot M \tag{15}$$

The result, which was tested by simulation is useful to compare the effects of the various error sources. From this rough result it can be concluded that this error is negligible compared to roundoff errors of arithmetic. Kaiser and Knight [9] give a worst-case analysis. They survey the maximum absolute quantizing error in the sin/cos values. The measure of the ms output noise to the ms output signal they obtained is

$$\frac{\text{ms (coeff error)}}{\text{ms (signal)}} \leq 4^2 \varepsilon^2 M^2 \cdot 2 \tag{16}$$

Here ε is the max error of the coefficients. This result shows that the error-to-signal ratio is to increase as M^2 compared to (15) where the rate of increase is only proportional to M. However, even if the latter expression is a good approximation of the error it is less by more than an order than the error caused by the arithmetic rounded, assuming that the number of input data points $N \leq 4096$.

Effects of windowing

In digital spectrum analysis only a time-limited segment of the input signal can be processed. The result of this truncation in the time domain is a modification of the spectrum. The time-limited signals can be obtained by multiplying the input signal with a rectangular window function:

$$x_N(t) = x(t) \cdot w_N(t)$$

$$\text{where} \qquad w_N(t) = \begin{cases} 1 & 0 \leq t \leq T \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

The effect of this multiplication is that the calculated Fourier transform is the convolution of the transform of the original signal and the transform of the

window function. The magnitude spectrum of the rectangular window function can be seen in Fig. 3. Its effects (widening the main lobe and the appearance of spurious sidelobes) can be modified by choosing different window functions. In digital processing the samples of the window function are used. The window samples, however, are represented by limited number of bits, so the parameters of the quantized window will be different from the parameters of the original

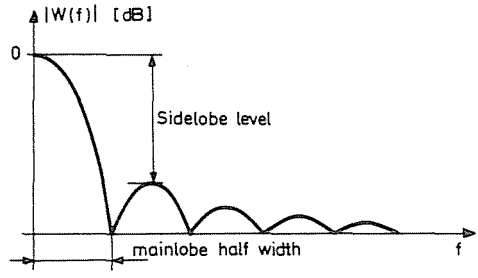

Fig. 3

one. The most important parameters of a window in the frequency domain are the mainlobe half width and the sidelobe level (Fig. 3). For spectrum analysis purposes window functions are needed which have low sidelobes combined with a not-too-wide mainlobe.

The effects of quantization of the window samples are studied by Prabhu and Agrawal [11]. They determined the parameters of the most often applied windows using various word length samples. They concluded that using more then 8 bits is sufficient to closely approximate the unquantized case. It is well known that data (time) windows can be implemented either in the time domain or the frequency domain. The windows which can be represented by the general form

$$w_N(n) = \sum_{i=0}^{j} a_i \cos \frac{2\pi i n}{N} \tag{18}$$

where a_i $i=0, 1, 2, \ldots j$ are constants and j is usually less than 4, can be implemented easily in the frequency domain. Their effect in the frequency domain is

$$X_w(k) = \sum_{i=0}^{j} a_i [X(k+i) + X(k-i)] \tag{19}$$

Because of the quantization errors, windowing in the time domain affects all the FFT computations, since these errors get embedded into the input data to FFT. The frequency domain implementation avoids these errors, since

windowing is realized after FFT computations. But frequency domain windowing needs extra multiplications with a_i constants. These constants are represented also with finite word length. Temes and Babic [13] study this problem and give some results. They found window functions where the a_i coefficients can be quantized without destroying the performance.

Hardware design considerations

Developing a hardware spectrum analyser the effects of quantization must be taken into consideration. The natural way to decrease these effects would be to use more-bit length data. This leads to higher cost and reduced operational speed.

In the following, some simple suggestions are given to form a hardware implementation which may be less sensitive to quantization errors.

1. The errors in a butterfly calculation can be reduced if instead of rounding after every multiplication, only the results of the whole butterfly are rounded. This means that in expression (11) the last term disappears. In this case the disadvantageous effect of the multiplication by $W \ll 1$ is also cancelled. The cost of this modification is that the hardware will be a bit more complicated. The results of the b-bit multiplications must be summed without previous rounding; instead of a b-bit adder a $2b$-bit adder must be used. Further, to avoid owerflows inside a butterfly calculation, the adder must be two-bit wider. The number of bits of the multiplier and the capacity of the data store, however, are not increased.

2. In an FFT calculation the fewer stages of butterflies are needed the more the errors can be reduced. Applying higher radix FFT the number of butterfly iterations can be decreased without shortening the original data record. The application of higher radix algorithms decreases not only the number of iterations but also reduces the total of the required computations. However, as the base of the algorithm increases, the algorithm becomes involved which needs more sophisticated hardware. Thus radix 4 seems to be a good compromise.

3. It is well known that the effect of quantization can be reduced by the application of dither [6]. Dithering means that a noise with proper parameters is added to the signal before quantization. Applying an independent noise which is uniformly distributed between $-k\frac{q}{2}$ and $+k\frac{q}{2}$ where q is the quantum size and k is an integer, the erroneous effects of the quantization can be cancelled. But if the amplitude interval of the dither does not exactly equal to $k \cdot q$ the

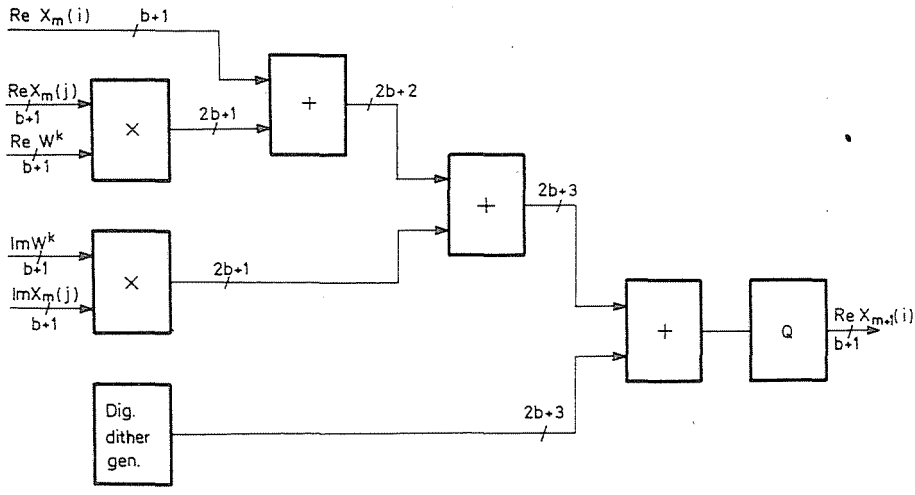

Fig. 4

error reducing effect is worsening quickly. Thus, instead of a uniformly distributed dither it is common to use other noises (e.g. a Gaussian) which can be implemented easier. But these dithers do not cancel only reduce the errors.

The studies mentioned before show that the most significant quantization error originates in the arithmetic rounding. To reduce the intermediate roundoff errors using the dithering technique, a simple hardware modification is suggested. It is easy to generate uniformly distributed random (or pseudo random) data sequences with an amplitude range which corresponds to the quantum size. Adding the subsequent samples of this digital dither signal to the result of a butterfly calculation before rounding, the quantization error can be reduced. But this dither also increases estimate variances. In Fig. 4 the hardware block scheme of the suggested butterfly calculator can be seen.

References

1. NUTTAL, A. H.—CARTER, G. C.: A Generalized Framework for Power Spectral Estimation. IEEE. Trans. on ASSP. vol.-ASSP-28. no. 3. pp. 334—335.
2. JENKINS, G. M.—WATTS, D. G.: Spectral Analysis and its Applications. San Francisco CA: Holden Day, 1968.
3. HARRIS, F. C.: On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. Proc. IEEE. vol. 66. no. 1. pp. 51—83.
4. NUTTAL, A. H.: Some Windows with Very Good Sidelobe Behaviour IEEE. Trans. on ASSP. vol. ASSP-29. no. 1. pp. 84—91.

5. WIDROW, B.: A Study of Rough Amplitude Quantization Thesis, M.I.T. 1956.
6. WATTS, D. G.: A General Theory of Amplitude Quantization with Application to Correlation Determination. Proc. IEEE. vol. 109 Part C. pp. 209—218.
7. WELCH, P. D.: A Fixed Point Fast Fourier Error Analysis. IEEE. Trans. on AU. vol. AU-17 pp. 151—157.
8. TRAN-THONG—LIU, B.: Fixed Point Fast Fourier Transform Error Analysis. IEEE. Trans. on ASSP. vol. ASSP-24 no. 6. pp. 563—573.
9. KAISER, R.—KNIGHT, W. R.: A Simple Fixed Point Error Bound for the Fast Fourier Transform. IEEE. Trans. on ASSP. vol. ASSP-27. no. 6. pp. 615—620.
10. OPPENHEIM, A. V.—WEINSTEIN, C. J.: Effects of Finite Register Length in Digital Filtering and the Fast Fourier Transform Proc. IEEE. vol. 60. no. 8. pp. 957—976.
11. PRABHU, K. M. M.—AGRAWAL, J. P.: Selection of Data Windows for Digital Signal Processing. Proc. of the IEEE. Conf. on ASSP 1978. pp. 79—92.
12. BABIC, H.—TEMES, G. C.: Windows for Digital Spectrum Analysis.

Gábor HORVÁTH H-1521 Budapest