

# Effects of Level Quantization and Threshold Clipping of the Signal and Basis Functions of Discrete Fourier Transform

Gamlet S. Khanyan, *Member, IEEE*

**Abstract**—The paper studies the influence of signal quantization levels number on accuracy of the results of spectral analysis. The overflow effect (signal threshold clipping due to shortage of the quantizing device bits) is also considered. A formula is derived for transforming a real number to its nearest quantization level. Numerical modeling of the quantized realizations of harmonic signal (pure one and mixed with noise) as well as its Fourier transform's basis functions is performed to construct characteristics – dependencies between program-assigned signal parameters and those measured in the course of digital processing under various quantization and clipping conditions.

**Index Terms**—Analog-to-digital conversion, discrete and fast Fourier transform, level quantization, threshold clipping.

## I. INTRODUCTION

LEVEL quantization is an inherent procedure of digital signal processing (DSP). Besides, discrete variability is the basis of most physical and biological world, and of various mathematical constructions (energy levels in quantum mechanics, DNA encoding, Boolean algebra, Walsh functions, etc.). Yet, wherever the reality idealization is based on the description of continually changing values, level quantization is considered as a distorting phenomenon. This is also true of spectral analysis based on discrete Fourier transform (DFT) method, where the transformed series of numbers, which are supposed to be continual in theory, are level quantized in practice.

Level quantization takes place not only during signal acquisition – when it undergoes analog-to-digital conversion (ADC), but also during its further processing by a computing device with finite number of memory cells' and processor registers' bits. In the latter case it is convenient to study, as an example, the level quantization of basis functions (sines and cosines) of DFT calculated in a direct (but slow) way, and by a rather sophisticated in mathematical description method of fast Fourier transform (FFT), having multiple versions of its

implementation algorithms [1]. These versions (data decimation in time or frequency domain with their prior or post processing by the “butterfly” scheme, etc.) “shuffle” differently the interim results and are not commutative in their ultimate calculation with the level quantization operation.

Level quantization is often accompanied by clipping the signal on thresholds established, e.g., for protection against overload.

Questions about the effects of these and other ways of signal restriction on the accuracy of the Fourier transform are relevant to the metrology of spectral analysis as a means of measuring the amplitude-phase-frequency characteristics of oscillatory processes of different nature. Answers to those questions will allow manufacturers and customers of microelectronic devices to coordinate more precisely the actual performance of the industry's products with features envisaged during their design stage.

Level quantization of the signal has been subject of a large number of works – from the earliest to the present time of DSP development history [2]–[7]. However, effects related to it cannot be considered to have been fully studied. For instance, well-known manuals on theory and practice of signal processing [2]–[4] confine themselves to describing the quantization phenomenon and rationale of probability distributions of the rounding noise (error), in particular the assumption of its uniform distribution. Among the quoted (by no means complete) list of the literature the papers [5]–[7] can be highlighted where quantization effects are studied as applied to one of the vast areas of DSP – digital filtering.

As for another major DSP area – discrete spectral analysis, one can feel a noticeable lack of references here, and the present work is aimed, if not to fill this gap, but to demonstrate a possible research direction of the problem, which occupies a prominent place among other problems of measurements and signal processing.

## I. MATHEMATICAL BACKGROUND

Let us consider an analog signal  $s(t)$  describing a physical process, and its digital realization  $s_n$  of duration  $T$  s and length of  $N$  samples, obtained (starting at a time moment  $t_0$ ) with sampling frequency  $F$  Hz. The signal is assumed to be clipped on constant thresholds  $A$  and  $B$ :

Manuscript received December 14, 2010.

G. S. Khanyan is with the Central Institute of Aviation Motors named after P. I. Baranov, 2, Aviamotornaya st., Moscow, 111116, Russian Federation (phone: +7 906 099 9958; fax: +7 495 552 4847; e-mail: dep007@rtc.ciam.ru, khanyan@mail.ru).

$$\left\{ \begin{array}{l} s_n = \begin{cases} A, & s(t_n) \leq A \\ s(t_n); & A < s(t_n) < B \\ B, & s(t_n) \geq B \end{cases} \\ t_n = t_0 + n/F; \quad n = 0, 1, \dots, N-1; \quad N = TF. \end{array} \right. \quad (1)$$

Before digital processing, such a signal undergoes level quantization, which can be represented as a chain of operators affecting  $s_n$  and transforming it into a quantized number

$$\bar{s}_n = \mathbf{P}^{-1} \mathbf{Q} \mathbf{P} s_n. \quad (2)$$

At first, the sample  $s_n$ , possessing a physical dimension of process  $s(t)$ , is affected by scaling (calibration) operator  $\mathbf{P}$ , which transforms it into a dimensionless number contained in the range with fixed integer boundaries  $L_A = \mathbf{P}A$  and  $L_B = \mathbf{P}B$  corresponding to the thresholds  $A$  and  $B$ . The operation of this is linear

$$s'_n = \mathbf{P} s_n = C s_n + D, \quad (3)$$

with the coefficient  $C$  and the displacement  $D$ , equal

$$C = (L_B - L_A)/(B - A); \quad D = (L_A B - L_B A)/(B - A). \quad (4)$$

Then quantization operator  $\mathbf{Q}$  itself comes into effect and rounds up the scaling result to the nearest integer number:

$$s''_n = \mathbf{Q} s'_n = [s'_n + 1/2]. \quad (5)$$

Finally, the obtained integer (5) is descaled, i.e., returned to the original physical dimension by applying the inverse operator  $\mathbf{P}$ :

$$\bar{s}_n = \mathbf{P}^{-1} s''_n = (s''_n - D)/C. \quad (6)$$

Thus, we arrive at the function, depending on the single variable  $s_n$  and four parameters  $A, B, L_A, L_B$ :

$$\bar{s}_n = \left[ \frac{(s_n - A)L_B - (s_n - B)L_A}{B - A} + \frac{1}{2} \right] \frac{B - A}{L_B - L_A} + \frac{L_B A - L_A B}{L_B - L_A}. \quad (7)$$

Real ADC devices contain an even number of quantization levels  $L=2^l$ , where  $l$  is the number of binary digits. The levels are numbered from  $L_A = -L/2$  to  $L_B = L/2 - 1$ . Of interest is an odd number of levels  $L$  with the boundaries  $L_A = -(L-1)/2$ ,  $L_B = (L-1)/2$ . Both cases of  $L$  are easy to combine:

$$L_A = -[L/2], \quad L_B = [(L-1)/2]; \quad L \geq 2. \quad (8)$$

It turns out that formula (7) is invariant under the displacement by an integer constant  $L'$ , i.e., it does not change when replacing  $L_A$  and  $L_B$  by  $L_A + L'$  and  $L_B + L'$ . In particular, we can set  $L_A = 0$ ,  $L_B = L - 1$  in (7) and simplify it:

$$\bar{s}_n = \left[ \frac{(s_n - A)(L-1)}{B - A} + \frac{1}{2} \right] \frac{B - A}{L - 1} + A. \quad (9)$$

Considering now an important case of threshold symmetry  $B = -A > 0$  and normalizing the signal per that threshold ( $\sigma_n = s_n/B$ ), we obtain a formula of quantization and clipping

$$\bar{\sigma}_n = \begin{cases} q(\sigma_n) = \frac{2}{L-1} \left[ \frac{(L-1)\sigma_n + L}{2} \right] - 1, & |\sigma_n| < 1 \\ \text{sgn } \sigma_n, & |\sigma_n| \geq 1, \end{cases} \quad (10)$$

which depends on a single variable  $\sigma_n$  and contains a single parameter – the number of quantization levels  $L$ .

Formula (10) looks most simply for the cases of two-level ( $L=2$ ) and three-level ( $L=3$ ) quantization

$$\bar{\sigma}_n = \begin{cases} +1, & \sigma_n \geq 0 \\ -1, & \sigma_n < 0 \end{cases}; \quad \bar{\sigma}_n = \begin{cases} +1, & \sigma_n \geq +1/2 \\ 0, & -1/2 \leq \sigma_n < +1/2 \\ -1, & \sigma_n < -1/2 \end{cases} \quad (11)$$

Note that the quantization function  $q$  is properly defined on clipping thresholds (equal to  $\pm 1$ ): putting  $\sigma_n = \pm 1$  in (10) we get  $\bar{\sigma}_n = \pm 1 = q(\pm 1)$ . But for an even  $L$  the zero value of  $\sigma_n$  shifts:

$$\bar{\sigma}_n(0) = 1/(L-1), \quad L \bmod 2 = 0; \quad \bar{\sigma}_n(0) = 0, \quad L \bmod 2 = 1. \quad (12)$$

An important property of the  $q(\sigma)$  function lies in the fact that an integer constant  $K$  can be moved beyond the brackets:

$$q(\sigma + K) = q(\sigma - J) - Jq(0) + K, \quad J = (KL - K) \bmod 2. \quad (13)$$

If  $K$  is even or  $L$  is odd, then  $J=0$  and  $q(\sigma)$  becomes periodical.

Difficulties of further research lie in the fact that it is impossible to obtain an analytical expression for the Fourier transform of any type of quantized signal, except for the trivial  $s(t) = \text{const}$ .

The exact formulas for the spectra of the amplitudes and phases are derived for time-limited pure harmonic signal under the assumption of continuity of its samples [8]. They are the basis of special methods of spectral analysis that increase accuracy of estimating harmonic oscillation's parameters, and are applied in the present paper. Here, despite the fact that the quantized signal can be described by means of piecewise constant function (10), to define the coordinates of those constants' boundaries on the horizontal axis seems to be not possible. Hence the impossibility of the DFT calculated sum fragmentation. Therefore, the only way to study the problem is numerical simulation. A shortcoming of this approach is the need to consider a lot of special cases, whose classification is unlikely to be carried out comprehensively.

## II. DESCRIPTION OF NUMERICAL EXPERIMENTS

Numerical experiments to determine the influence of level quantization and threshold clipping on the accuracy of spectral analysis were performed using the dynamic measurements digital signal processing program *quatrix.exe*<sup>®</sup> [9], in which the procedures described by formulas (1)–(10) were included.

As a model of physical process (1) the analog harmonic signal of frequency  $f_0$ , amplitude  $a_0$ , and initial phase  $\varphi_0$  mixed with noise and observed in a window with initial time  $t_0 = -T/2$  was taken:

$$s(t) = a_0 \cos(2\pi f_0 t + \varphi_0) + r(t); -T/2 \leq t < T/2. \quad (14)$$

Harmonic signal, by virtue of the superposition principle (the linearity of Fourier transform), serves as a structural component of various types of dynamic processes (vibration, pulsation, etc.). Therefore, the results obtained in a study of this model can be rightfully extended to the complex physical processes that have polyharmonic nature and occur in reality.

The noise in model (14) was formed by the software generator of uniformly distributed pseudorandom numbers,  $r' = \text{random}$ , so that the noise component's samples

$$r_n = r(t_n) = b_0(2r' - 1); \quad 0 < r' < 1 \quad (15)$$

varied within limits  $\pm b_0$ , whereas the harmonic component was between  $\pm a_0$ . Hence, samples  $s_n$  were limited and normalized per threshold  $B = 1 + b_0$  so that for  $a_0 > 1$  signal clipping was imitated.

The objective of each experiment was to build characteristics, that is a plot of a harmonic signal's parameter under measurement depending on the threshold or the number of quantization levels of either the signal or the basis functions of DFT or FFT when that parameter is fixed as a given constant in the model (14). Effect of quantization or clipping was assessed by comparing the experimentally measured signal parameters – frequency  $f$ , amplitude  $a$ , and phase  $\varphi$  with the program-assigned values  $f_0$ ,  $a_0$ ,  $\varphi_0$ .

Signal parameters were estimated on the basis of processing the results of discrete Fourier transform

$$S_m = \frac{1}{N} \sum_{n=0}^{N-1} \sigma_n e^{-i2\pi mn/N}; \quad m = 0, 1, \dots, N-1 \quad (16)$$

that calculates the spectral function  $S_m$  of the integer frequency variable (bin)  $m$ .

The first half of samples  $S_m$  (hermitic conjugate with the second one) is used (assuming  $N$  to be an even number) to establish both the spectra of amplitudes  $A_m$  and phases  $\Phi_m$

$$A_m = 2 |S_m|, \quad \Phi_m = \arg S_m; \quad m = 0, 1, \dots, m_0, \dots, N/2 \quad (17)$$

of the signal's digital realization  $s_n$  on the frequencies  $f_m = m/T$ . Here the bin  $m_0$  is singled out, which is the address of the maximum peak in the amplitude spectrum searched by sorting out and comparing components of the array of numbers  $A_m$ .

#### A. Frequency measurement

It is clear that  $m_0$  is a rough estimate of the process (14) harmonic component's dimensionless frequency  $f_0 T = m_0 + \mu_0$ .

Fractional adjustment  $\mu_0$  is estimated by formula

$$\mu_0 = (A_{m_0+1} - A_{m_0-1}) / (A_{m_0+1} + A_{m_0-1}), \quad (18)$$

possessing, as was shown in [8], good accuracy for moderate noise level  $b_0$  and for  $m_0$  location sufficiently distanced from spectrum edges  $m=0$  and  $m=N/2$ .

This formula, in which the nearest (left and right) neighbors

of the amplitude spectrum maximum peak are presented, was used to measure frequencies with accuracy exceeding the spectral resolution.

#### B. Amplitude measurement

The main problem that arises when evaluating the amplitude is its lowering at non-integer value of the dimensionless frequency  $f_0 T$  (maximum peak of amplitude spectrum turns out to be  $a_0 \sin(\pi \mu_0) / (\pi \mu_0)$  instead of  $a_0$ ). This phenomenon accompanied by the appearance of false sidelobe components, is known in the literature as a leakage effect (see, e.g., [2], [4]). There is no leakage at  $\mu_0=0$ , and the leakage is maximum at  $\mu_0=1/2$  (in the latter case, the peak amplitude is ~64% of the harmonic oscillation's amplitude's true value).

To smooth the leakage effect, focusing method proposed in [8] was applied. The essence of the method consists in summing the square of amplitude spectrum maximum peak with squares of the related sidelobe components, and taking the square root  $a$  of that sum for evaluation of the oscillation's amplitude  $a_0$ .

#### C. Phase measurement

To assess phase  $\varphi_0$ , alternating method [8] was applied, which eliminates phase distortion – its false shift in the  $\pi \mu_0$  taking place in the traditional spectral analysis (provided in time window  $0 \leq t < T$  alleged “by default”). The essence of the alternating method (provided in time window  $-T/2 \leq t < T/2$ ) is to swap the first ( $0 \leq n < N/2$ ) and second ( $N/2 \leq n < N$ ) halves of the signal's digital realization  $s_n$  before performing DFT. And then the phase  $\varphi_0$  is estimated with accuracy  $\pi \mu_0 / N$  as the value of phase  $\Phi_m$  at the address  $m_0$  of the amplitude spectrum peak.

The results of phase measurements were outputted in degrees (within  $\pm 180^\circ$ ).

### III. SIGNAL QUANTIZATION AND CLIPPING RESEARCH RESULTS

Digital realizations of process (14) with duration of  $T=1$  s and length of  $N$  samples were formed with sampling frequency  $F$  Hz, so that the spectral resolution  $F/N=1/T$  was 1 Hz and the dimensionless frequency  $f_0 T$  coincided with the dimensioned one  $f_0$ .

#### A. Appearance of the oscillograms and spectrograms

Prior to discussing the results, it is of interest to regard the external look of examined signals and their amplitude spectra.

Fig. 1a indicates the first half of the oscillogram drawn with points and the amplitude spectrum of pure harmonic signal ( $a_0=1$  V,  $b_0=0$ ) of a semi-whole dimensionless frequency 256.5 drawn with solid line. The realization length is  $N=2048$  samples, and the number of quantization levels is  $L=4096$ , which is typical of 12-digit ADC-board often used in practice.

With this number of levels accepted to be “infinite”, as is shown in further analysis, the samples of signal can be considered as continuum numbers and both direct and fast Fourier transform methods – as identical in terms of their precision.

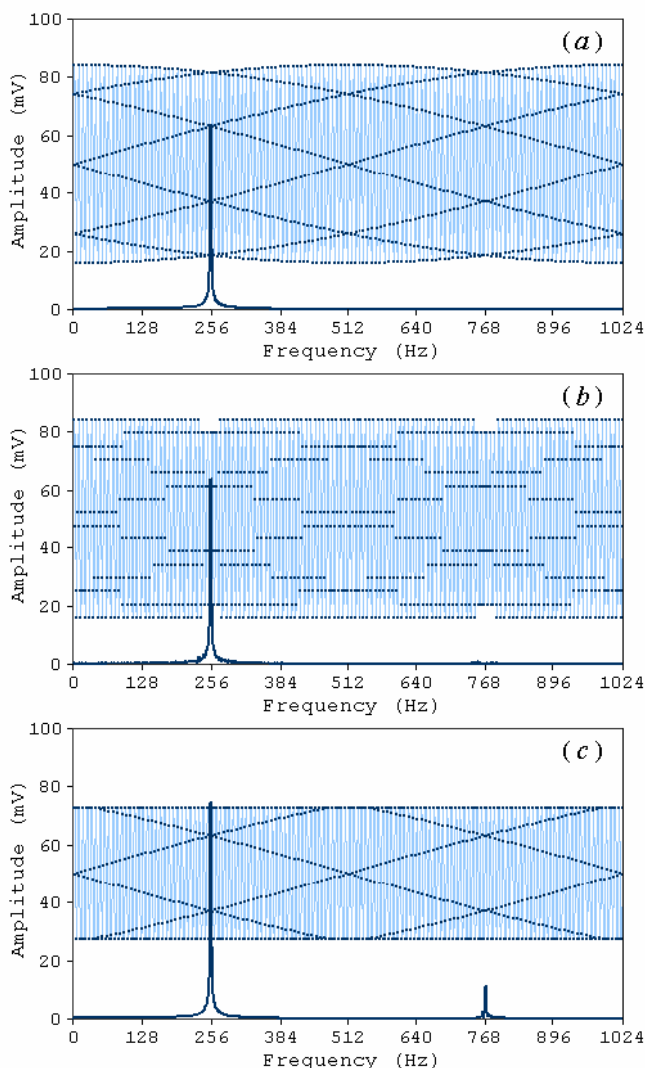


Fig. 1. Waveform and amplitude spectrum of a sinusoid under various conditions of quantization and clipping: *a* – large, *b* – small number of signal quantization levels without clipping; *c* – signal clipping with a large number of its quantization levels

The same signal quantized with a small number of levels  $L=16$  is shown in Fig. 1*b*, and in Fig. 1*c* – with  $L=4096$  levels, but clipped on amplitude ( $a_0=1.5$ ).

A comparison of these illustrations shows how little the amplitude spectrum view is affected by the strong oscillogram view change. Spectrum distortion is more noticeable in case of the signal clipping, which is proved by the harmonic occurring at frequency 767.5 Hz.

Such a weak effect of level quantization on the amplitude spectrum (inadequate degree of waveform distortion) is certainly a positive fact, and at the same times an unexpected paradox.

### B. Amplitude-frequency characteristics

Fig. 2 shows amplitude-frequency characteristics obtained for a short length  $N=32$  of the signal's digital realization. Measurements were made at a constant phase  $\varphi_0=90^\circ$ .

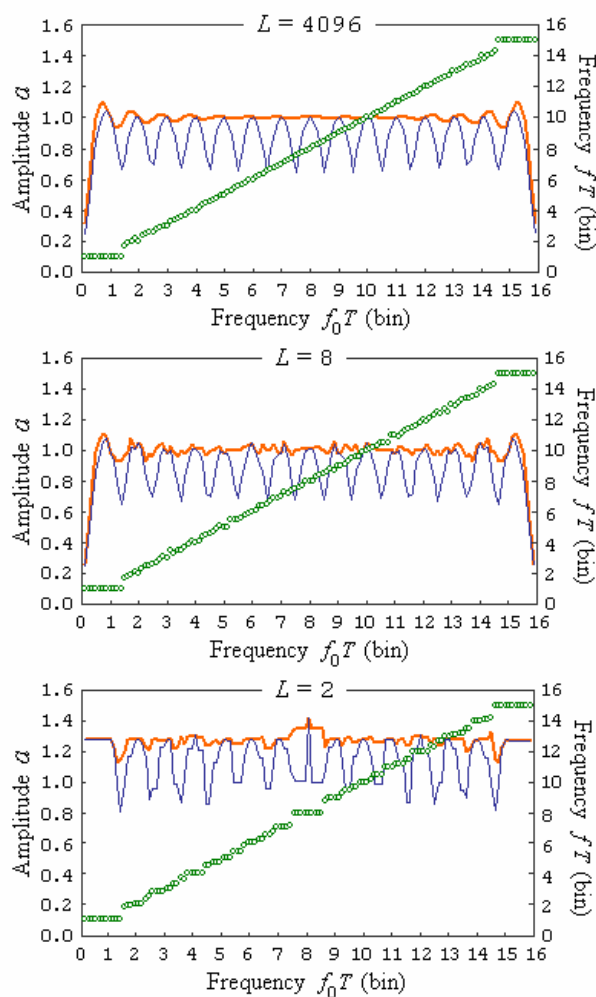


Fig. 2. Amplitude-frequency characteristics of pure harmonic signal for different number  $L$  of its quantization levels

Variable for each of the specified quantization levels  $L$  was the dimensionless signal frequency  $f_0T=m_0+\mu_0$  incrementing by step  $1/8$  from spectrum origin to Nyquist frequency  $N/2$ . Amplitude was built depending on it: the orange line – using focusing method, the blue line – without adjustment (peak value of amplitude spectrum at the address  $m_0$ ). Green circles show measured signal frequencies adjusted according to (18).

Amplitude measurement result close to the ideal one  $a=1$  is ensured for  $L \geq 8$  through focusing application. Amplitude oscillations calculated without adjustments are explained by leakage lowering it at semi-whole frequencies to  $\sim 64\%$  of  $a_0$ .

The frequency measurement plot practically coincides with the program-assigned line, except in the spectrum central zone for two-level quantization and at the spectrum edges for any number of quantization levels. In the latter case, the so-called edge effect holds for the amplitude, being smoothed only if  $L=2$ .

An extremely small number of levels ( $L < 8$ ) raises the amplitude measurement result (by 20–30%), practically not touching the nature of genuine spectral deficiencies (leakage and edge effects).

### C. Phase characteristics of quantization

Fig. 3 shows phase characteristics of harmonic signal – results of phase measurement depending on the number of quantization levels  $L$ .

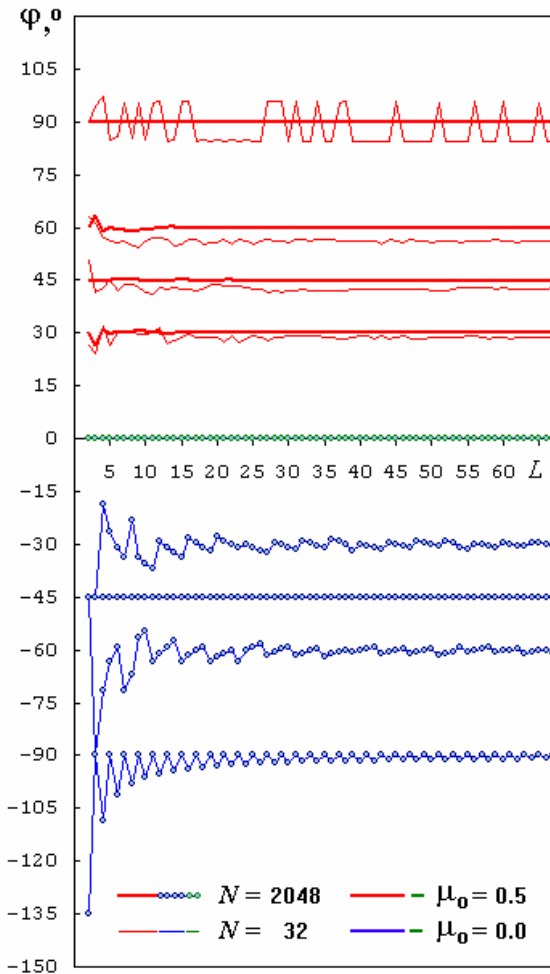


Fig. 3. Phase characteristics of pure harmonic signal with  $N=32$  and 2048 realization length measured in the absence and with the utmost leakage effect

Characteristics were taken in the spectrum center (on bin  $m_0=N/4$ ) with constant phase  $\varphi_0$  taken from the row  $0^\circ, \pm 30^\circ, \pm 45^\circ, \pm 60^\circ, \pm 90^\circ$ . Using this sign symmetry to analyze the leakage effect for  $\varphi_0 \geq 0$   $\mu_0=1/2$  was set, and for  $\varphi_0 \leq 0$   $\mu_0=0$ . The signal's realization length was assigned to be large ( $N=2048$ , thick line and points) and small ( $N=32$ , thin line).

Results of phase measurement for long and short realizations are the same when there is no leakage (as it is evident for the curves in the lower part of the figure). It is surprising that distortion is either totally absent (for phases  $0^\circ, -45^\circ$ ), or for small  $L$  it is large and can have either “sawtooth” behavior (for phases  $-30^\circ, -60^\circ$ ) or appear only when the number of quantization levels is even (for phase  $-90^\circ$ ).

If leakage is present (see the upper part of the figure), phase distortion depends on realization's length. It is minor for  $N=2048$  and proportional to the phase itself for  $N=32$ .

In any case the phase measurement result can be considered to be approaching the asymptotic limit value for  $L > 64$ .

### D. Noise influence

To assess the effect of noise the harmonic signal was mixed with uniformly distributed noise of the same intensity ( $b_0=a_0=1$ ), so that the ideal result of amplitude measurement was  $a=0.5$ . In the characteristics in Fig. 4  $N=512$ ,  $\varphi_0=0$ ,  $f_0T=128$ .

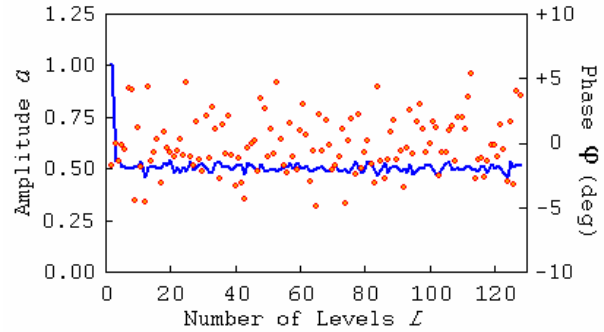


Fig. 4. Amplitude  $a$  (solid curve) and the phase  $\varphi$  (points) of harmonic signal mixed with noise depending both on the number of quantization levels  $L$

A strong (twofold) overestimation of the amplitude (focused) is observed only for two-level quantization. For  $L > 2$  the noise-caused scattering of both phase (by several degrees) and amplitude (by some percentage) is not sensitive to quantization.

It can be argued that the noise is a distorting factor which is almost independent of the quantization (as well as of leakage, anyway).

### E. Signal clipping

To research the signal threshold clipping effect, characteristics given in Fig. 5 were taken.

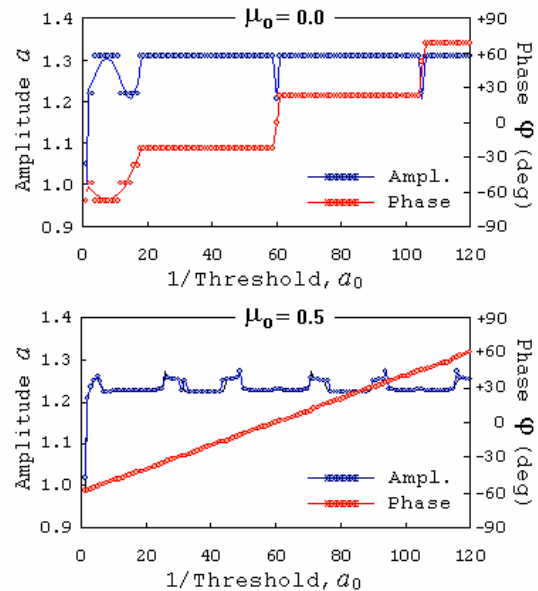


Fig. 5. Amplitude  $a$  and phase  $\varphi$  of a pure harmonic signal depending on the signal clipping threshold

The variable on  $X$ -axis is “overload factor” – parameter  $a_0$  of model (14) showing how many times the signal amplitude is larger than unity threshold in (10).  $Y$ -axes feature amplitude

$a$  and phase  $\varphi$  of the pure harmonic signal with length of  $N=512$  samples and of frequency  $128+\mu_0$ , where  $\mu_0$  adopted values 0 and  $\frac{1}{2}$  corresponding to the minimum and maximum leakage effects. In each case a large ( $L=4096$ , thin line) and a small ( $L=4$ , points) number of quantization levels were taken. Phase  $\varphi_0$  was set to vary linearly: from  $-60^\circ$  to  $+60^\circ$ .

The experimental results radically differ for different cases of  $\mu_0$ . Paradoxically, for the maximum leakage phase measurements fit perfectly on program-assigned line, whereas in the absence of leakage, they are “ragged,” piecewise constant, differing for small values of  $a_0$  by cases of  $L$  (this difference is also clearly seen for amplitude measurements).

As for the amplitude, clipping leads to its overestimation of about 30%. The clipped signal with unlimited growth of  $a_0$  approximates the signal with 2–3 quantization levels.

#### IV. FOURIER TRANSFORM’S BASIS FUNCTIONS QUANTIZATION RESEARCH RESULTS

We simulated and compared three situations related to the Fourier transform’s basis functions level quantization.

In the first of them samples of signal (14) with amplitude  $a_0=1$ , equal to threshold  $B=1$  (which ensures absence of signal clipping), were quantized. The second situation was created by quantization of the DFT (16) basis functions’ samples during the cycles passing on  $m$  and  $n$  in the course of their direct calculation, and in the third situation the FFT sines and cosines table’s entries during their preliminary calculation were quantized.

As trigonometric functions vary within limits  $\pm 1$ , quantization in the two latter cases, according to (10), was also carried out without threshold clipping.

##### A. Appearance of the spectrograms

In Fig. 6 signal is almost continual – it has  $L=4096$  quantization levels, but both the samples of DFT and FFT basis functions have only  $L=8$  levels.

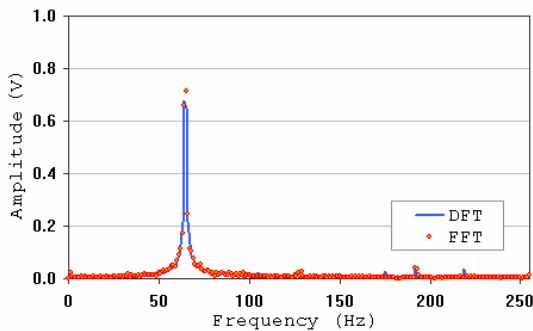


Fig. 6. Amplitude spectrum of a sinusoid with  $L=8$  quantization levels of both DFT and FFT basis functions

We see a generally weak, but noticeable difference in the peak amplitude of the DFT and FFT spectra from each other, which already indicates the impact of quantization on the metrological properties of fast computational algorithms.

##### B. Frequency characteristics of the quantization

Fig. 7 shows the frequency characteristics plotted in the absence of leakage and when it has the maximum effect.

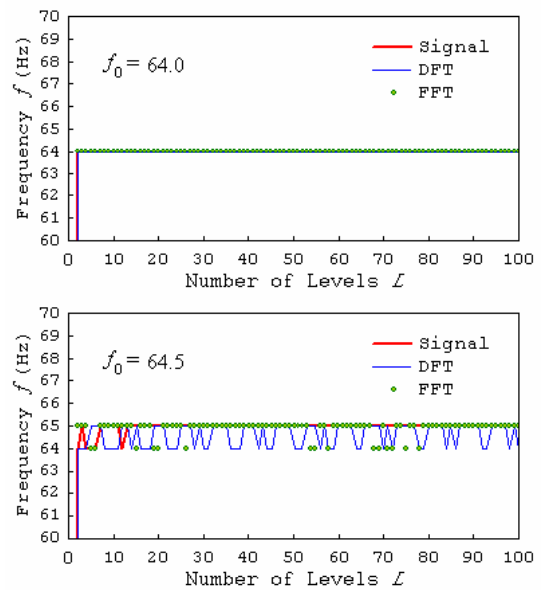


Fig. 7. The results of measuring an integer and a half-integer frequency of sine wave depending on the number of quantization levels of signal and basis functions of DFT and FFT

In the first case ( $f_0T=64$ ) the ideal outcome measure  $f=f_0$  is obtained at the two phase values ( $\varphi_0=0^\circ$ ,  $\varphi_0=45^\circ$ ) beginning from number of quantization levels  $L=3$  (both for signal and for basis functions of the Fourier transform).

In the second case ( $f_0T=64.5$ ,  $\varphi_0=0^\circ$ ) the value of  $fT$  varies between the nearest bins 64 and 65 giving them “equal preference” for the direct method and calculating the spectral function (16), which is a more plausible outcome compared to the fast method.

##### C. Amplitude characteristics of the quantization

Fig. 8 shows the amplitude characteristics obtained for  $\varphi_0=0^\circ$  on the integer and half-integer signal frequency.

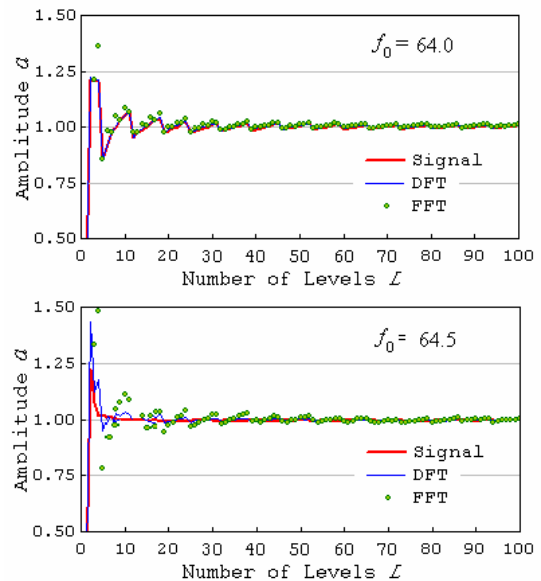


Fig. 8. The amplitude of the integer and half-integer frequency sine wave depending on the number of quantization levels of signal and basis functions of DFT and FFT

It is evident that the behavior of curves in all the three quantization modes depends strongly on the leakage effect. Closer to each other are results given by quantization of signal and DFT basis functions. However, the quantization of the FFT basis functions increases the existing error of the amplitude  $a$ .

#### D. Phase characteristics of the quantization

Fig. 9 shows the results of measuring the zero initial phase  $\varphi_0=0$  of a harmonic signal obtained in the absence of leakage and when it has the maximum effect.

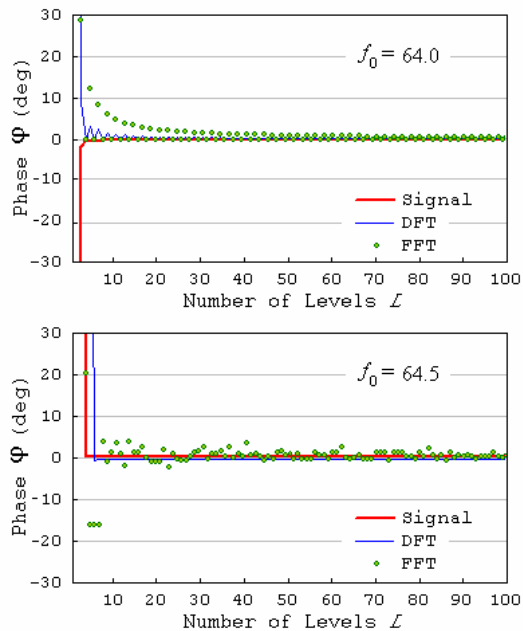


Fig. 9. Measurements of zero-phase of an integer and a half-integer frequency sine wave depending on the number of quantization levels of signal and basis functions of DFT and FFT

Just as in the measurement of the amplitude, closer results are obtained by quantization of signal and DFT basis functions.

However, the quantization of the FFT basis functions significantly increases the error in phase  $\varphi$ , that indicates the downside of fast algorithm application for small  $L$ .

#### V. CONCLUSION

Level quantization, which dramatically changes the signal appearance, has little effect on the precision of spectral analysis results. Frequencies of process harmonics are determined with a sufficient accuracy for  $L > 2$  signal quantization levels, amplitudes – for  $L > 8$  levels, and phases – for  $L > 64$  levels. The amplitude and phase spectra distortions are more noticeable in case of the signal clipping.

Level quantization of discrete Fourier transform's basis functions produces by means of the fast calculation method a markedly greater error of spectral estimates of the parameters of the harmonic components of the varying process

as compared to the direct calculation method. Therefore, FFT and other fast algorithms of digital signal processing, implemented in microelectronic devices with a small number of memory cells' bits, should undergo metrological certification as a means of measuring the parameters of processes.

In those DSP areas where spectral analysis methods can be applied, processing a large amounts of data from many sources (sensors) of data acquisition (aviation engines stand testing, meteorology, monitoring of environment and engineering facilities of great length: railways, oil and gas pipelines, etc.) can be carried out with multiple data contraction by selecting a small number of quantization levels for the signals. This justifies demand for designing, producing and purchasing electronic and computing devices with a small number of bits of data provided.

The proposed method of investigating the effects of quantization and clipping can be used in various fields of digital signal processing and related disciplines (wavelet analysis, sequential analysis, etc.).

#### REFERENCES

- [1] R. E. Blahut, *Fast algorithms for digital signal processing*. MA: Addison-Wesley, 1984.
- [2] L. R. Rabiner, and B. Gold, *Theory and application of digital signal processing*. NJ: Prentice-Hall, 1975.
- [3] I. S. Gonorovsky, *Radio-engineering circuits and signals*. Moscow: Sovetskoye Radio, 1977.
- [4] A. V. Oppenheim, and R. W. Schaffer, *Discrete-time signal processing*. NJ: Prentice-Hall, 1989.
- [5] A. Yurdakul, and G. Dünder, "Statistical Methods for the Estimation of Quantization Effects in FIR-Based Multirate Systems," *IEEE Trans. Signal Process.*, vol. 47, no. 6, pp. 1749–1753, Jun. 1999.
- [6] Yu. A. Bryuhanov, "Quantization and overflowing effects in digital recursive first-order filters with round-off under constant external influence," in *Proc. of 8th Int. Conf. "Digital Signal Processing and its Applications"*, Moscow, 2006, vol. 1, pp. 191–194.
- [7] J. Paduart, J. Schoukens, and Y. Rolain, "Fast measurement of quantization distortions in DSP Algorithms," *IEEE Trans. Instrum. Meas.*, vol. 56, no. 5, pp. 1917–1923, Oct. 2007.
- [8] G. S. Khanyan, "Analytical investigation and estimation of errors involved in the problem of measuring the parameters of a harmonic signal using the Fourier transform method," *Measurement Techniques*, NY: Springer, Aug. 2003, vol. 46, pp. 723–735 [*Izmeritel'naya Tekhnika*, Russia, 2003, no. 8, pp. 3–10].
- [9] G. S. Khanyan, and N. V. Sheina, "A digital signal processing system for data support of GTE stand tests," in "*Scientific Contribution to the Creation of Aviation Motors*", vol. 2, V. A. Skibin, and V. I. Solonin (ed.), Moscow: Mashinostroenie, 2000, pp. 534–536.

**Gamlet S. Khanyan** (M'10) – Was born in 1950. Graduated from the Moscow Physical-Technical Institute in 1974. Works at the Central Institute of Aviation Motors in the Department of Dynamic Measurements and Signal Processing as a Senior Research Fellow. In 2004, defended his thesis for a Ph.D. degree in Technical Sciences on the subject "Development of the Discrete Spectral Analysis of Fast Varying Dynamic Processes as Applied to the Information Support of Gas Turbine Engines Tests". Member of the Russian A. S. Popov Society for Radioengineering, Electronics and Communications since 2009.

