MODELLING A QUANTIZER — PROBLEMS AND POSSIBLE APPROACHES

T. DOBROWIECKI

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

Summary

Since the quantizer became a permanent part of modern information processing schemes it is usually of great importance to know how the quantizer error can be taken into consideration. The description of the quantizer depends greatly on the description of its input and output signals, thus stochastic or alternatively, deterministic methods can be chosen. Sometimes mixed description methods turn out to be very fruitful.

In the following a short survey is given on the different possibilities of describing the quantizer effect.

Introduction

Amplitude quantization has recently become a permanent link in the chain of operations performed on information carrier and leading from the system under study to the expected measurement results. The term "quantizer" refers to many different but essentially all steplike system characteristics which transform the continuous signal amplitude into the finite amount of the possible values.

Considering the place the quantizer takes in data processing schemes we differentiate between input quantization (input signal amplitude quantization), coefficient quantization (finite precision in the machine realization of the theoretical algorithm coefficients) and arithmetical quantization (rounding etc. due to the finite precision arithmetic) [1], [2].

In the majority of data processing algorithms all these error-sources occur, as a rule, together. Input quantization nevertheless is considered to be the most interesting case, since the remaining two can be successfully modelled with it.

From the signal flow point of view, a quantizer can occur in the close-loop or in the open-loop as well. The close-loop case refers mainly to DM, DPCM data conversion and transmission and their adaptive versions ([3], [4], [5], Fig. 1), the open-loop case corresponds usually to the normal AD data conversion ([6], [7], Fig. 2).

Regarding the quantizer characteristics we make a distinction between a uniform and a nonuniform one. The uniform quantizer is in some way equivalent to the linear transducer and usually used in all-purpose AD conversion. Apart from this in special purpose design nonlinear quantizer characteristics are used, with the nonlinearity being optimally fitted to the special requirements of the measurement. Such quantizer characteristics are

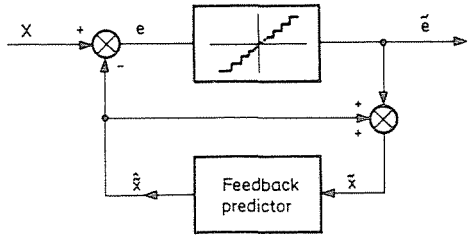

Fig. 1

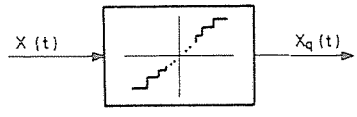

Fig. 2

used in processing signals possessing some specific features, e.g. speech processing ([8], [7], [4]).

The open-loop quantizer schemes were extended from the common 1-dimensional quantizers (operating on one amplitude sample at the same time) to the N-dimensional, so-called block quantizers (operating on N amplitude samples at the same time) ([9], [10]) but in the following we will leave this case out of consideration.

Modelling a quantizer—general considerations

The quantizer transforms signal characteristics and as such must be subject to the appropriate system modelling. In its most simplified version a fine (small step) quantizer can be regarded as the unit transfer, if, of course, the maximal $\pm q/2$ error entering the signal processing can be neglected. If the quantization error must be taken into consideration it has to be mentioned that the quantized signals can be described in the stochastic or deterministic way and this will, of course, influence our quantizer model.

If the quantizer input signal $x(t)$ is a stochastic one, so is the quantizer error signal $\varepsilon(t)$ (Fig. 3) and although each sample function of $\varepsilon(t)$ depends directly on the corresponding sample function of $x(t)$, the average properties of the error may turn out to be more independent from the properties of the input signal. If $x(t)$ is described in a deterministic way, so is $\varepsilon(t)$ and their dependence

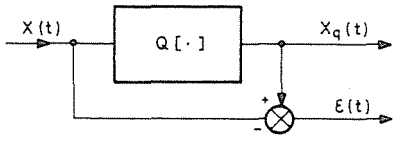

Fig. 3

is much more noticeable (although the connection between $x(t)$ and $x_q(t)$ is of course far from a unique one-to-one, but many-to-one — the input x determines the error ε and the output x_q completely, but the knowledge of x_q still does not permit the full determination of x).

The stochastic approach

Close-loop schemes

If we consider the description of the quantizer and its influence on the signals we must stress the difference between the close-loop and the open-loop applications. In close-loop schemes (DM, DPCM and their adaptive versions) the error due to the intermediate quantization has two components: a granular and a slope overload one (Fig. 4, [5]). Both components depend greatly on the input signal characteristics, the quantizer itself and the properties of the feedback circuitry used.

Fortunately the exact characteristics of the noise components are usually not demanded, it is enough to deal with the signal-to-noise ratios (which are very complicated formulas). For example [11] gives noise power values for the single integration DM encoder in the case of Gaussian signals. With the assumption of perfect integration and the separability of the granular and slope overload noise components the formulas for average noise power (beside the usual quantizer step size) involve also the variance of $\dot{x}(t)$ and of $\ddot{x}(t)$, correlation of $x(t)$, bandwidth of $x(t)$, system sampling frequency and the double infinite summations.

In the case of DPCM the situation is still more difficult to handle.

As far as the frequency spectra are concerned there is no unambiguous solution for the general properties of the DPCM (DM) error spectra, because of the determining role of the feedback circuitry. The error signal is by no means white and the frequency distribution of its power can be relatively simply manipulated. In [3] a reduction of the error power and the signal distortion is considered by shifting the error signal (with a proper choice of feedback) to the

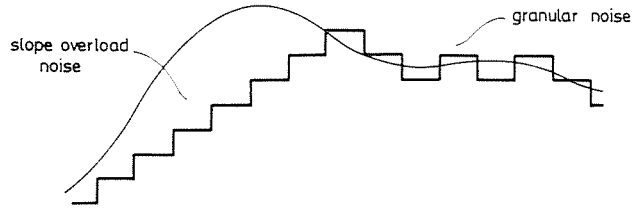

Fig. 4

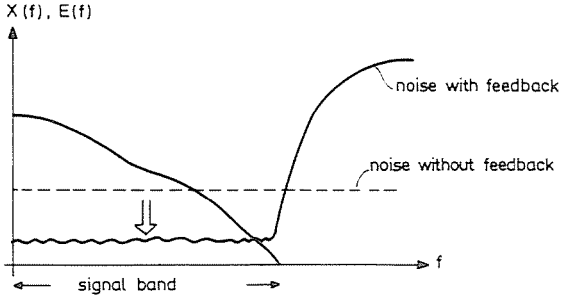

Fig. 5

frequencies higher than the input signal band and partly removing it by a low-pass filtering operation (Fig. 5).

The adaptive case is still more difficult because despite say, a stochastic description of the input signal, the choice of quantizer steps depends on the concrete sample function of the input. The desired results can be gained only by means of simulations as for example in [12].

Before switching over to the open-loop structures we must note that the close-loop scheme (because of the feedback) exhibits a peculiar feature of the so-called idle channel noise, which in the open-loop case is not present ($x_q(t) \neq 0$ with $x(t) = 0$).

As reported in [13], the spectral characteristics of the idle channel noise depend greatly on the quantizer step imbalance and eventual threshold hysteresis.

In the case of the open-loop quantizer (Fig. 2), the situation is much more simple. First there is no slope overload noise component, only the granular-like one (due to the finite quantizer steps and the possible amplitude saturation). Another difference is that open-loop quantizers usually have a finer (smaller) step size, which compensates the lack of feedback.

Open-loop schemes — additive noise model

In the description of open-loop quantizers the most successful attempt was the stochastic notion leading to the uniform density, input independent white noise model ([14], [15], Fig. 6).

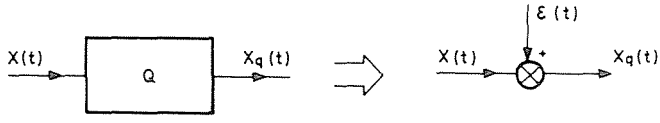

Fig. 6

It can be noticed that this feature may be introduced, somehow intuitively ([14]) if one takes into consideration that with ever finer quantization steps the signal variation in the successive amplitude windows will be essentially linear and the saw-tooth-like error will thus have uniform amplitude density.

The additive noise model was successfully derived (with a full theoretical background) for the case of the uniform, nonsaturating quantizer only, in the other cases the arising analytical difficulties did not allow a similar level description. For the nonlinear quantizers the most widely used quantity became the signal-to-noise ratio (e.g. [8]) which is usually minimized in some way.

Ideal uniform quantizer — a summary

In the following we will give a short summary of the analytical tractable uniform quantizer. The quantizer description is based on the characteristic function approach, treated in detail in [14], [16], [15]. We will recall here the most significant and essential relations only.

The density function for the quantizer error is:

$$f_\varepsilon(\varepsilon)=\begin{cases}\dfrac{1}{q}+\dfrac{1}{q}\displaystyle\sum_{n\neq 0} W_x\left(\frac{2\pi n}{q}\right)\exp\left(-j\frac{2\pi n\varepsilon}{q}\right) & |\varepsilon|\leq\dfrac{q}{2}\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

the second order density is:

$$f_{\varepsilon_1\varepsilon_2}(\varepsilon_1,\varepsilon_2)=\begin{cases}\dfrac{1}{q^2}+\dfrac{1}{q^2}\sum\limits_{l\neq 0}\sum\limits_{k\neq 0} W_{x_1x_2}\left(\dfrac{2\pi l}{q},\dfrac{2\pi k}{q}\right)\exp\left[-j\dfrac{2\pi}{q}(l\varepsilon_1+k\varepsilon_2)\right] & \\ \qquad\qquad \text{for } |\varepsilon_1|,|\varepsilon_2|\leq\dfrac{q}{2}; & \\ 0 \qquad\qquad \text{otherwise}. & \end{cases} \tag{2}$$

and the error-input correlation is as follows:

$$E\{x\cdot\varepsilon\}=\frac{q}{2\pi}\sum_{k\neq 0}\frac{(-1)^{(k-1)}}{k}\cdot\dot{W}_x\left(\frac{2\pi k}{q}\right) \tag{3}$$

it can be noticed that if:

$$W_x(\alpha)=0 \qquad \text{for all} \qquad |\alpha|\geq\frac{2\pi}{q} \tag{4}$$

then:

$$f_\varepsilon(\varepsilon)=\frac{1}{q} \qquad E\{x\cdot\varepsilon\}=0 \tag{5}$$

and the quantizer error has a uniform density and is uncorrelated with the input.

If additionally:

$$W_{x_1x_2}(\alpha_1,\alpha_2)=0 \quad \text{for all} \quad |\alpha_1|\geq\frac{2\pi}{q},\ |\alpha_2|\geq\frac{2\pi}{q} \tag{6}$$

then

$$f_{\varepsilon_1\varepsilon_2}(\varepsilon_1,\varepsilon_2)=f_{\varepsilon_1}(\varepsilon_1)\cdot f_{\varepsilon_2}(\varepsilon_2)=\frac{1}{q^2} \tag{7}$$

and the quantization error is white [14], [16].

The (4) and (6) conditions of bandlimitedness are unfortunately only the sufficient ones, and also they are not realizable physically. The (4) and (6) conditions (quantization theorems) can be thus fulfilled only approximately [14]. Nevertheless there are physically realizable signals which do not fulfill (4) or (6), their quantization errors are still uniformly distributed and white (for example a uniformly distributed input signal), so the suitable conditions must be further generalized. We notice, moreover, that the fulfilment of the quantization theorem leads to so-called Sheppard-correction terms ([14]), which express the connection between the moments of the quantizer input and output signals.

In [16] the generalization of the quantization theorems is given. The new sufficient, but also necessary conditions for the quantization noise being

uniform and white are:

$$W_x\left(\frac{2\pi k}{q}\right)=0 \quad \text{for all} \quad k\neq 0 \tag{8}$$

$$W_{x_1 x_2}\left(\frac{2\pi l}{q},\frac{2\pi k}{q}\right) \quad \text{for all} \quad l\neq 0,\ k\neq 0$$

The uncorrelatedness of the quantization error with the input is by (8) unfortunately not granted. It requires an additional condition in the form of:

$$\dot{W}_x\left(\frac{2\pi k}{q}\right)=0 \qquad \text{for all} \quad k\neq 0 \tag{9}$$

In the case of an input signal not fulfilling the conditions in (4), (6) or (8), as for example Gaussian signals (the characteristic function of the Gaussian signal is not bandlimited and possesses no zeros), the quantizer error is naturally

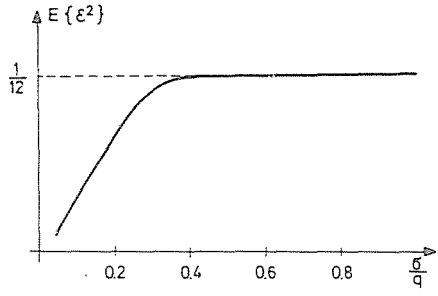

Fig. 7

nonuniform, colored and correlated with the input. The measure of deviation from the ideal model depends strongly on the ratio σ/q [16].

The variance of error is shown in Fig. 7. It is obvious that if $\sigma/q \geq 1$, the additive noise model is valid as far as the first-order statistics are concerned (if $\sigma/q \gg 1$, the model is also valid in the sense of second-order statistics).

Dithering and the additive noise model

We will consider the problem of dithering separately. If $x(t)$ does not fulfill the quantization theorem, the Sheppart-corrections are not valid, in particular:

$$E\{x_q\} \neq E\{x\} \tag{10}$$

A remedy is an auxiliary, so-called dither signal, added to the input (Fig. 8), which by itself fulfills the generalized quantization theorem (8), [6]. In this case:

$$E\{x_q\} = E\{x\} + E\{\mathrm{d}\} \tag{11}$$

and

$$E\{x_q^2\} = E\{x^2\} + E\{\mathrm{d}^2\} + \frac{q^2}{12} + q \sum_{k \neq 0} \frac{(-1)^{k-1}}{\pi k} W_x\left(\frac{2\pi k}{q}\right) \dot{W}_d\left(\frac{2\pi k}{q}\right) \tag{12}$$

thus the measurement of the mean can be corrected but the measurement of the variance is still biased, with the bias depending on the input signal characteristics.

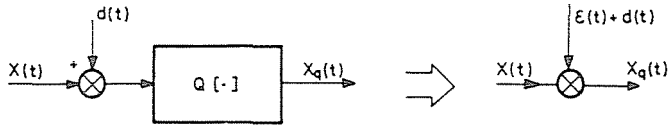

Fig. 8

Generally, if the characteristic function $W_d(\mathrm{d})$ of the dither signal fulfills the generalized quantization theorem up to its Nth derivative, inclusively:

$$W_d\left(\frac{2\pi k}{q}\right) = \dot{W}_d\left(\frac{2\pi k}{q}\right) = \ldots = W_d^{(N)}\left(\frac{2\pi k}{q}\right) = 0 \qquad \text{for all} \quad k \neq 0 \tag{13}$$

the moment correction terms are independent from the input signal up to the $(N-1)$-th moment, inclusively.

Considering that the dither signal will be intrinsically part of the measurement device, it would be more sensible from the input-output point of view, to reduce the two noise sources to one and keep the additive noise model untouched (Fig. 8). This, however, automatically rules out the usual uniform noise model (since the sum of the uniformly distributed error and the dither will have a triangular-like or more complicated density function) and makes the problem of modelling the uniform quantizer more specific.

The case of a general (nonlinear) quantizer characteristic

The usefullness of the additive error model for the uniform quantizer depends strongly on the fulfillment of the quantization theorems because if so, the error is uniform, white and uncorrelated. On the other hand, however, if the theorems are not fulfilled, there is no general approximate additive model, the

degree of approximation depends heavily on the input signal involved and must be carefully checked in every particular case.

The analytical description of the quantizer for the arbitrary input signal and for the nonuniform quantizer characteristic leads to great difficulties, whether a deterministic or stochastic kind of description. In case of a general quantizer characteristic the study of the time-domain or the frequency-domain behaviour of the quantized data may be made somewhat easier by the mathematical formalism, but the crucial data is supplied by way of simulations.

Of course some very general properties of the quantized signal can still be deduced. In the frequency-domain, for example, the quantization will increase the signal band beyond any limit, regardless its definition or the kind of signal description (deterministic, stochastic) [17]. Furthermore, since the quantized signal will have infinite slopes, its spectrum will be bounded by a hyperbole, the least upper bound may, however, be relatively simply determinated ([21]). Thus in the case of an arbitrary signal:

$$E\{|X_q(f)|\} \leq \frac{\text{const}}{f} \tag{14}$$

the constant depends on the quantizer threshold and step values, on the signal density and its average level-crossing slopes at the quantizer thresholds.

Mixed methods

One possible way for overcoming the difficulties connected with a general description is to mix the stochastic and the deterministic methods. In [18], a successful attempt was made to compute quantized spectra for different kinds of (deterministically described) signals, with the assumption that the quantizer error is strictly saw-tooth-like (ideal uniform, nonsaturating quantizer characteristic). The error signal of such an idealized quantizer can then be written as:

$$\varepsilon(t) = \sum_{k=1}^{\infty} (-1)^k \frac{\sin 2\pi k x(t)}{\pi k} = \sum_{k=1}^{\infty} \varepsilon_k(t) \tag{15}$$

each $\varepsilon_k(t)$ being a phase-modulated signal with a phase deviation $2\pi k$ and amplitude $1/\pi k$. Therefore the spectrum of the quantization error is the sum of the spectra of the phase-modulated signals. For a further computation, the following assumptions must be made:

— the power spectral density of a phase-modulated signal is proportional to the amplitude distribution of the derivative of the modulating

signal, so:

$$S_k(\omega) = \frac{1}{2\pi^2 k^3} \cdot f_x(\omega/2\pi k) \tag{16}$$

— the spectral components of various $\varepsilon_k(t)$ add on a power basis, so the power spectral density of the quantization noise is:

$$S_\varepsilon(\omega) = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} f_x(\omega/2\pi k)/k^3. \tag{17}$$

For example the error spectrum for the sinusoidal input signal is shown in Fig. 9 ([18]). One can observe that the spectrum is flat at the lower frequencies, so if

Fig. 9

we restrict ourselves to this band, the white additive noise model would still be approximately valid. The band of the flat error spectrum could be extended, if the quantizer step q were decreased ([18]).

The pure deterministic approach

If the quantizer is not a uniform one, if the input signals do not permit proper stochastic description or are strictly deterministic ones, the additive noise model is usually out of question. Furthermore if (because of the necessary assumptions) we do not consider the mixed methods to be reliable, the purely deterministic approach remains.

In the deterministic notion an open-loop quantizer is the extreme case of a general, memoriless, nonlinear transfer characteristic. Since the fine quantizer can be modelled after all with the additive noise model, the subject of approach will be mainly the coarse quantizer characteristic.

The nonlinear quantizer characteristic in the deterministic approach can be approximated with some "better behaving" functions or at least we can make attempts to treat it analytically ([19], [20]).

The approximation can be performed in different ways, usually in the form of:

$$x_q(t) = Q[x(t)] = \text{l.i.m.} \sum_k a_k \varphi[x(t)] \tag{18}$$

where the choice of the function system, $\varphi_k[\cdot]$, depends on the properties of the input signals and the limit operation is usually the mean-square limit. (The above-mentioned approximation is naturally valid for an arbitrary quantizer characteristic.)

The approximating series in (18) leads to different kinds of power series in the time-domain in terms of $x(t)$, or if we are interested in the frequency domain analysis, to the multiple convolutions in terms of $X(f)$.

In the case of arbitrary input signals there is nothing to do but to perform simulations if, however, the multiple convolutions can be evaluated, the quantized signal spectra will be computable. Such is the case of input signals $x(t)$ with non-negative Fourier-transforms. Their multiple convolutions tend to the limit possessing a Gauss-like shape and the Fourier-transform of the deterministic quantized signal $x_q(t)$ at the frequency of f_0 can thus be computed as:

$$X_q(f_0) \simeq \sum_{k=M}^{N} \xi_k \cdot G(f_0/\sigma_k)\,[1 + E_k(f_0)] \tag{19}$$

where ξ_k are the coefficients of the quantizer characteristic expansion into the power series, $G(\cdot)$ is the Gauss-density function and $E_k(\cdot)$ are the error terms corresponding to the k-th multiple convolution. The M, N summing limits depend on the accuracy demands and on the relationship between the $x(t)$ signal band and the desired frequency f_0.

The performed simulations show good agreement with the higher frequency values [19], Fig. 10.

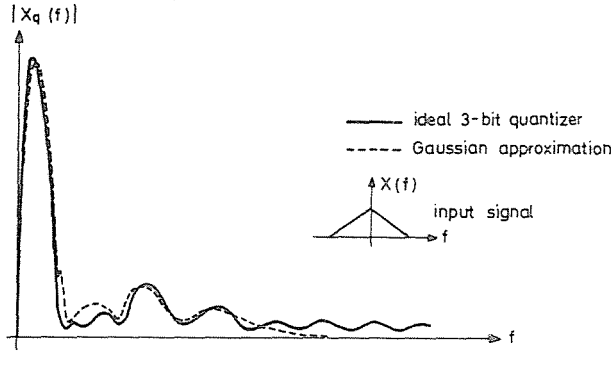

Fig. 10

The explicit, analytical approach works only in the case of slightly distorted sinusoidal signals. The 1-bit quantizer (hard limiter), and with it an arbitrary, but symmetrical quantizer characteristic can be expressed as:

$$x_q(t) = Q[x(t)] = \frac{2}{\pi} \int_0^\infty \sin[\xi x(t)] \cdot H(\xi) \frac{d\xi}{\xi} \tag{20}$$

where

$$H(\xi) = 2 \sum_{k=1}^{N} \beta_k \cdot \cos \xi a_k$$

and β_k, a_k are the quantizer step and threshold values.

If the input signal is of the form:

$$x(t) = A \sin \omega_0 t + y(t) \qquad |y(t)| \ll A \tag{21}$$

the (20) integral can be evaluated in terms of Bessel functions of the first kind.

Conclusions

In the above survey we gave a short summary of the possible approaches to the problem of modelling the quantized signals. The best model would of course be the one that is independent from the input signals, generally, however, this is not the case. In the close-loop schemes no possibilities for the unified treatment of the quantization error arise. In the adaptive cases only simulations give useful results, the properties of the error in the nonadaptive close-loop schemes are greatly variable and depend not only on the quantizer and the input signal, but on the feedback system characteristics as well. The usual description method is the signal-to-noise ratio. The general nonlinear open-loop quantizer gives as yet too few opportunities for successfull modelling attempts.

In the stochastic case only the signal-to-noise ratio remains a powerful description tool. Up to the present only the uniform, nonsaturating quantizer characteristic turned out fruitful enough and resulted in a very well-known additive model. But problems remain, the additive noise model requires a fine, ideally uniform characteristic and in case of any deviations the validity of the model is highly questionable and must be established in all particular cases.

The dithering method, if performed ideally, still results in an additive noise model, the distribution of the noise varies however. The rough quantization (1-bit, 2-bit, etc.) without dithering is another problem, the additive noise model is not valid and the descriptional difficulties are

overwhelming. As a possible way the mixed or the purely deterministic approach can be suggested. The deterministic approach — to investigate the quantizer output in time- or frequency- domain in the case of arbitrary quantizer characteristics and signals is just as hopeless as its stochastic counterpart.

Some more possibilities remain if the quantizer is coarse (few steps only) and the input signals have some well-defined properties, useful from the deterministic point of view.

References

1. HORVÁTH, G.: FFT-based Spectrum Analysis from the View-point of Hardware realization, Period. Polytechn. El. Eng. *28.*, 2—3, p. 201—213 (1984).
2. PÉCELI, G.: Finite Wordlength Effects in Digital Filters, Period. Polytechn. El. Eng. *28., 2—3, pp. 191—200 (1984).*
3. SPANG, H. A., SCHULTHEISS, P. M.: Reduction of Quantizing Noise by Use of Feedback, IRE, vol. CS-10, Dec. 1962, pp. 373—380.
4. GERSHO, A., GOODMAN, D. J.: Theory of an Adaptive Quantizer, IEEE, vol. COM-22, Aug. 1977, pp. 1037—1045.
5. JAYANT, N. S.: Digital Coding of Speech Waveforms: PCM, DPCM and DM Quantizers, Proc. IEEE, vol. 62, May 1974, pp. 611—632.
6. SCHUCHMAN, L.: Dither Signals and Their Effect on Quantization Noise, IEEE, vol. COM-22, Dec. 1964, pp. 162—165.
7. MAX, J.: Quantization for Minimum Distortion, IRE, vol. IT-6, March 1960, pp. 7—12.
8. PANTER, P.F., DITE, W.: Quantization Distortion in Pulse-Count Modulation with Nonuniform Spacing of Levels, IRE, vol. 39, Jan. 1951, pp. 44—48.
9. GERSHO, A.: Assymptoticaly Optimal Block Quantization, IEEE, vol. IT-25, July 1979, pp. 373—380.
10. HUANG, J. J. Y., SCHULTHEISS, P. M.: Block Quantization of Correlated Gaussian Random Variables, IEEE, vol. CS-11, Sept. 1963, pp. 289—296.
11. O'NEAL, J. B.: Delta Modulation of Data Signals, IEEE, vol. COM-22, March 1974, pp. 334—339.
12. LIU, B., GOLDSTEIN, L. H.: Power Spectra of ADPCM, IEEE, vol. ASSP-25, Feb. 1977, pp. 56—62.
13. CHING, Y. C., GOTZ, B., BALDWIN, G. L.: Idle Channel Noise Characteristics of a Delta Modulator with Step Imbalance and Quantizer Hysteresis, Conf. Rec., 1977 IEEE Int. Conf. on Commun., June 17—19, 1974, pp. 13A-1—13A-6.
14. KOLLÁR, I.: Statistical Theory of Quantization: Results and Limits, Period. Polytechn. *28., 2—3, pp. 173—189 (1984).*
15. WIDROW, B.: Statistical Analysis of Amplitude Quantized Sampled Data Systems, Trans. AIEE, Pt. II, Applications and Industry, vol. 79, Jan. 1960, pp. 555—568.
16. SRIPAD, A. B., SNYDER, D. L.: A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White, IEEE, vol. ASSP-25, Oct. 1977, pp. 442—448.
17. WISE, G. L., TRAGANITIS, A. P., THOMAS, J. B.: The Effect of a Memoryless Nonlinearity on the Spectrum of a Random Process, IEEE, vol. IT-22, Jun. 1977, pp. 84—88.

18. CLAASEN, T. A. C. M., JONGEPIER, A.: Model for the Power Spectral Density of Quantization Noise. IEEE, vol. ASSP-29, Aug. 1981, pp. 914—917.
19. DOBROWIECKI, T.: Quantized Error Spectra at High Frequencies for a Certain Class of Signals, Proc. of the IMEKO symp. on Application of Statistical Methods in Measurement, Leningrade, 1978.
20. DOBROWIECKI, T.: Amplitude Quantization — the Deterministic Case, Proc. of the IMEKO Symp. on Problems and Trends in Measurement and Instrumentation Education — Microprocessors and Allied Techniques, Budapest, 1980.
21. DOBROWIECKI, T.: Description of the Measurement Process by Means of Functional Analysis, C. Sc. Thesis Budapest, 1980.

Tadeusz DOBROWIECKI H-1521 Budapest