REFERENCE PROPERTIES OF UNIFORM QUANTIZERS — WIDROW'S VS. DIRECT APPROACHES

Andrzej Pacut, Konrad Hejn

Faculty of Electronics and Information Technology, Warsaw University of Technology
Nowowiejska 15/19, 00–665 Warsaw, Poland
e-mail: A.Pacut@ia.pw.edu.pl, K.Hejn@ise.pw.edu.pl
web: http://www.ia.pw.edu.pl/∼pacut, http://whisky.ise.pw.edu.pl/staff/konrad

Abstract - *We present an analysis of the quantization error for generalized (nonuniform and/or nonmonotonic) quantizers. The mean value and mean squared value of the quantization error and the quantized output are derived in terms of the observed signals. The results can be used to identify basic parameters of ADC converters. The results can be utilized in the reference models used in the ADC test methods recommended by the IEEE.*

Keywords - Analog-digital conversion, Error analysis, Identification, IEEE standards, Quantization, Statistical analysis.

1. MOTIVATIONS

Widrow's model of the quantization error [1] treats the quantization error as an additive random variable. The classical Quantization Theorem [2] became the basic analytical tool for this model and led to practical results for a sinusoidal test signal without and with dither [3, 4], or for a sinusoidal signal with an offset [5]. Unfortunately, while Widrow's *model structure* can be applied to a broad class of quantizers, the classical Quantization Theorem along with its later generalizations and refinements (see e.g. [8]) can be applied only to *uniform quantizers*, i.e. the quantizers with equidistant *transition levels* (*switching points*) and equidistant *digital codes*. Since quantizers' quality is related, among other, to a precision and stability of setting the switching points, the classical quantization theory cannot be applied to quality analysis of 'real life' quantizers.

One of the basic determinants of quantizers' quality is based on the notion of the effective resolution (efr) or related concepts. Popular definition of the effective resolution, like the one in IEEE standards [9, 10], hides several simplifying assumptions. These assumptions are not fulfilled for real-life quantizers. The effective resolution calculated on the base of IEEE standards *for physically the same quantizer* varies from a measurement to a measurement, not because the quantizer's effective resolution varies but because the simplifying assumptions are better or worse fulfilled. In other words, the definition which was to measure a quality undergoes uncontrolled variations whose extend depends on the quality itself. In order to derive a more stable definition it is necessary to analyze the quantization error in more detail in order to make the efr definition dependent only on quantizer's parameters rather then on the testing signal.

There is also a second reason for the approach other than the one based on classical Quantization Theorem. The formulas which express basic quantization error characteristics, like mean and the mean-square quantization error, *look* compact yet require infinite summations. For instance, for the most popular sinusoidal test signal, they require infinite summation of Bessel's functions. Number of terms needed to obtain a sufficient accuracy is quite high and use of finite summation leads to a numerical error. This error combines with the error resulting form a finite sample length. Consequently, the quantization error characteristics are biased. We found [13, 14] alternative versions of the quantization error characteristics which call only for *finite-term summations*. The alternative versions are not based on 'characteristic functions' approach of classical theorem but rather on exact analytical approach. This approach can be applied not only to 'ideal' quantizers, but can be extended to general quantizers. In this order we developed a novel direct approach to ADC analysis [15], verified the theory in extensive simulations [12], and then applied the theory to actual experimental data [17]. The theory leads to a corrected definition of the effective resolution [7] which is stable in measurements. It is well suited to applications and omits most of the simplifying assumptions applied to IEEE standards [9, 10]. Our approach basically follows the ideas of Gray [11]. For the application of the presented theory to efr measurement see [6, 7]. Our approach is valid for *any static quantizers*, including *non-uniform* and/or *non-monotonic* ones, and for both deterministic and stochastic input signals.

A comprehensive mathematical basis of this theory will be presented elsewhere [16]. Here we present the results obtained with the use of the proposed method for uniform quantizers, for which an alternative approach of Widrow exists. We compare usefulness of both approaches. We

ISBN 972-98115-5-5

also present examples of general results obtained for sine wave signals, which belong to the most popular test signals.

2. A REVIEW OF THE RESULTS FOR GENERAL QUANTIZERS

2.1. Definitions and notation

We first introduce — for a better reference – a notation and a terminology related to quantization. Quantization can be understood as a mapping of the real axis $\mathbb{R}$ (the analog axis) into a finite set of codes (the digital axis), Fig. 1, the upper part. Such a quantization *does not* involve any dynamics and may thus be called the *static quantization*. Any quantization which *does involve* dynamics, through a hysteresis or another dynamic effects, can be in this context called the *dynamic quantization*.

The points of the analog axis at which the mapping switches from one digital value to another one will be called the *transition levels*. To keep as close as possible to the real world, we assume that the *number of transition levels is finite*. To simplify the derivation we assign index 1 to the smallest transition level and continue to the biggest one, thus indexing the transition levels by consecutive integers

$$u_1 < u_2 < \ldots < u_n \tag{1}$$

The transition levels divide the analog axis into $n+1$ *bins* U_k, $k = 0, 1, \ldots, n$, namely

$$\begin{aligned} U_0 &= (-\infty, u_1), \\ U_k &= [u_k, u_{k+1}) \quad k = 1, \ldots, n-1 \\ U_n &= [u_n, \infty) \end{aligned} \tag{2}$$

Each bin U_k is mapped to a code which has certain numerical value y_k (Fig. 1, upper graph)

$$U_k \longmapsto y_k, \quad k = 0, \ldots, n \tag{3}$$

Formally, the general quantizer $\mathcal{Q}$ can thus be defined by a sequence of n transition levels and $n+1$ code values

$$\mathcal{Q} = (u_1, \ldots, u_n, y_0, y_1, \ldots, y_n) \tag{4}$$

The static quantization is uniquely defined by its *static characteristic* g (Fig. 1, lower graph) which in general is a step function which switches from one level (code value) to another at the transition levels. Assumption on finite number of transition levels makes g saturated both from above and from below.

We assume that quantizer's input signal $\mathbf{u}$ is stationary and denote by $f_{\mathbf{u}}$ and $F_{\mathbf{u}}$ the density and the cumulative distribution function of $\mathbf{u}$, resp., namely

$$F_{\mathbf{u}}(u) = \int_{-\infty}^{u} f_{\mathbf{u}}(x)\, dx \tag{5}$$

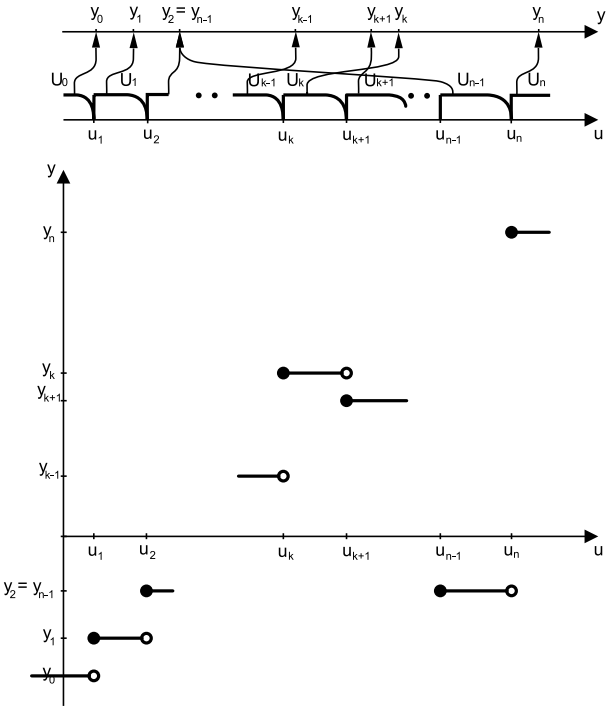

Figure 1. *Upper part:* Mapping the bins into the code values for a general static quantizer. Due to non-monotonicity, the arrows which characterize the mapping may cross, and the code values may not be unique, namely the arrows may point to same code value (here $y_2 = y_{n-1}$). *Lower part:* The static characteristic $y = g(u)$ of a general quantizer characterized by this mapping.

The complement of $F_{\mathbf{u}}$ to its limit value at ∞ (equal to 1) will be denoted by $\overline{F}_{\mathbf{u}}$, namely

$$\overline{F}_{\mathbf{u}}(u) = 1 - F_{\mathbf{u}}(u) = \int_{u}^{\infty} f_{\mathbf{u}}(x)\, dx \tag{6}$$

Moreover, the difference between F and $\overline{F}$ will be denoted by $\widetilde{F}$, namely

$$\widetilde{F}_{\mathbf{u}}(u) = F_{\mathbf{u}}(u) - \overline{F}_{\mathbf{u}}(u) = 2\, F_{\mathbf{u}}(u) - 1 \tag{7}$$

We assume that the first two moments of $\mathbf{u}(t)$ (the mean value $\mathcal{E}\mathbf{u}$ and the variance $\mathcal{V}\mathbf{u}$) exist and are finite. This guarantees the existence of the *incomplete mean* function

$$H_{\mathbf{u}}(u) = \int_{-\infty}^{u} x\, f_{\mathbf{u}}(x)\, dx, \tag{8}$$

its complement to its limit value $H_u(\infty) = \mathcal{E}\mathbf{u}$, namely

$$\overline{H}_{\mathbf{u}}(u) = \mathcal{E}\mathbf{u} - H_{\mathbf{u}}(u) = \int_{u}^{\infty} x\, f_{\mathbf{u}}(x)\, dx \tag{9}$$

and the difference of the two, namely

$$\widetilde{H}_{\mathbf{u}}(u) = H_{\mathbf{u}}(u) - \overline{H}_{\mathbf{u}}(u) = 2\, H_{\mathbf{u}}(u) - \mathcal{E}\mathbf{u} \tag{10}$$

We also use the difference operator δ, defined as

$$\delta z_k = z_k - z_{k-1} \tag{11}$$

for any sequence $\{z_k\}$.

2.2. Basic results for general quantizers

The most important characteristic of the quantizer is the quantization error defined as

$$\mathbf{e} = g(\mathbf{u}) - \mathbf{u} \tag{12}$$

This notion has a practical meaning only if the digital values are to be "close" to the analog values so it is important to measure this closeness. We will be interested here in the *mean quantization error*

$$\mathcal{E}\mathbf{e}(t) = \mathcal{E}g(\mathbf{u}(t)) - \mathcal{E}\mathbf{u}(t) \tag{13}$$

and the *mean square quantization error*

$$\mathcal{E}\mathbf{e}^2(t) = \mathcal{E}(g(\mathbf{u}(t)) - \mathcal{E}\mathbf{u}(t))^2 \tag{14}$$

Since we assumed that the analog signal $\mathbf{u}$ is stationary, the above error characteristics do not depend on time and depend only on the distribution of $\mathbf{u}(t)$ which is also time invariant. Further results presented here are based on the following result proved in [16].

Proposition 1 (General Quantizer) *For any static quantizer* (3) *with stationary random input* $\mathbf{u}$*, the mean quantization error is given by*

$$\mathcal{E}\mathbf{e} = -\mathcal{E}\mathbf{u} + \frac{y_0 + y_n}{2} - \sum_{k=1}^{n} \frac{\delta y_k}{2}\, \widetilde{F}_{\mathbf{u}}(u_k) \tag{15}$$

and the mean square quantization error is given by

$$\begin{aligned}\mathcal{E}\mathbf{e}^2 = {} & \mathcal{V}\mathbf{u} + 0.5\,(y_0 - \mathcal{E}\mathbf{u})^2 + 0.5\,(y_n - \mathcal{E}\mathbf{u})^2 \\ & + \sum_{k=1}^{n} \delta y_k \left(\widetilde{H}_{\mathbf{u}}(u_k) - \overline{y}_k\, \widetilde{F}_{\mathbf{u}}(u_k)\right)\end{aligned} \tag{16}$$

where $\mathcal{V}$ *is the variance and* $\overline{y}_k$ *denote the averages of two consecutive code values, namely*

$$\overline{y}_k = 0.5\,(y_k + y_{k-1}), \quad k = 1, \ldots, n \tag{17}$$

2.3. Bounded signals

Formulas (15, 16) are valid even for analog signals which extend beyond the quantizer range, including "infinite" amplitude signals. In the usual practice, the input signal is yet kept within the quantizer range and is bounded to a certain interval $U = [\underline{u}, \overline{u}]$. Consequently,

$$\begin{aligned} F_{\mathbf{u}}(u) &= 0 \quad \text{for } u \le \underline{u}, \\ \overline{F}_{\mathbf{u}}(u) &= 0 \quad \text{for } u \ge \overline{u} \end{aligned}$$

In other words, only the transition levels from the inside of U, i.e. from $(\underline{u}, \overline{u})$ contribute to the overall value of $\mathcal{E}\mathbf{e}$ and $\mathcal{E}\mathbf{e}^2$. Consequently, the transition levels $u_k \le \underline{u}$ and $u_k \ge \overline{u}$ *may* be removed from the error formulas. Let k_1 and k_n be the indexes of the smallest and the largest transition levels u_k still *inside* $(\underline{u}, \overline{u})$, and k_0 be the index of the transition level just preceding k_1, namely

$$u_{k_0} \le \underline{u} < u_{k_1} < \ldots < u_{k_n} < \overline{u} \ge u_{k_n+1} \tag{18}$$

Therefore, (15, 16) may be stripped of the excessive summation terms by replacing y_0 by y_{k_0}, y_n by y_{k_n}, and restricting the summation to $k_0 + 1, \ldots, k_n$. The following Proposition is proved in [16].

Proposition 2 (Bounded signals) *If an analog signals bounded to an interval* U*, the mean and the mean-square quantization errors of the general quantizer* (3) *are given by*

$$\mathcal{E}\mathbf{e} = -\mathcal{E}\mathbf{u} + \frac{y_{k_0} + y_{k_n}}{2} - \sum_{k=k_0+1}^{k_n} \frac{\delta y_k}{2}\, \widetilde{F}_{\mathbf{u}}(u_k) \tag{19}$$

$$\begin{aligned}\mathcal{E}\mathbf{e}^2 = {} & \mathcal{V}\mathbf{u} + .5\,(y_{k_0} - \mathcal{E}\mathbf{u})^2 + .5\,(y_{k_n} - \mathcal{E}\mathbf{u})^2 \\ & + \sum_{k=k_0+1}^{k_n} \delta y_k\, \widetilde{H}_{\mathbf{u}}(u_k) - \overline{y}_k\, \widetilde{F}_{\mathbf{u}}(u_k)\end{aligned} \tag{20}$$

where $k_1 = k_0 + 1$ *and* k_n *denote the indexes of the smallest and the largest transition levels* u_k *still* inside U (18).

Let us stress that if any end point of U is *equal* to a transition level then it may be removed from the summations. In fact, we can as well leave such points, as any other points that do not lay inside U, since they just do not contribute to the overall values of $\mathcal{E}\mathbf{e}$ and $\mathcal{E}\mathbf{e}^2$. Note yet that leaving a point that does not contribute must be matched by a parallel modification of y_{k_0} and/or y_{k_n}, i.e. always k_1 and k_n must be the indexes of the lowest and the biggest transition levels included in the summation.

2.4. Offset signals

It is often useful to separate an offset of the quantized signal in the error formulas. Suppose then that the input signal to the general quantizers is modified by a deterministic shift C, i.e. the input is equal to $\mathbf{u}(t) + C$ rather than to $\mathbf{u}(t)$. It is proven in [16] that for offset signals the mean and the mean-square quantization errors are given by

$$\mathcal{E}\mathbf{e} = -C - \mathcal{E}\mathbf{u} + \frac{y_0 + y_n}{2} - \sum_{k=1}^{n} \frac{\delta y_k}{2}\, \widetilde{F}_{\mathbf{u}}(u_k - C) \tag{21}$$

$$\begin{aligned}\mathcal{E}\mathbf{e}^2 = {} & \mathcal{V}\mathbf{u} + .5\,(y_0 - C - \mathcal{E}\mathbf{u})^2 + .5\,(y_n - C - \mathcal{E}\mathbf{u})^2 \\ & + \sum_{k=1}^{n} \delta y_k \left(\widetilde{H}_{\mathbf{u}}(u_k - C) - (\overline{y}_k - C)\, \widetilde{F}_{\mathbf{u}}(u_k - C)\right)\end{aligned} \tag{22}$$

3. UNIFORM QUANTIZERS

A quantizer will be called *uniform* if all the finite bin widths are identical, i.e.

$$u_k - u_{k-1} = Q, \quad k = 2, \ldots, n. \tag{23}$$

and the code values are equidistant

$$\delta y_k = y_k - y_{k-1} = \Delta, \quad k = 1, \ldots, n. \tag{24}$$

The static characteristic of uniform quantizers has a staircase graph, with constant "raiser heights" equal to code steps Δ, constant "tread depths" equal to bin widths Q, and two "landings" of infinite lengths. The above properties do not uniquely define the quantizer, i.e. there is a whole family of uniform quantizers. Since Q and Δ determine only the "scale" properties of the quantizer, it is still necessary to know any point *on* the characteristics to establish its position with respect to the origin. While the error characterization of uniform quantizers are well known, they are often derived for different quantizers.

To specify a uniform quantizer we may safely assume that there are some nonnegative code values. Consider the smallest nonnegative code value Y^* and the corresponding transition level U^*, Fig. 2. According to our definitions, U^* is equal to the analog value at which the static characteristic jumps from $Y^* - \Delta$ to Y^*. We will call the index k^* of both U^* and Y^* the *reference index*, and the point (U^*, Y^*) the *reference point*. The coordinates of the reference point will be called the *analog offset* U^* and the *digital offset* Y^*, resp., of the uniform quantizer. In other words, the digital offset Y^* is the smallest nonnegative code value, and the analog offset U^* is the transition level corresponding to this code value. The *reference point* fixes the position of the staircase characteristic with respect to the origin, and the reference index k^* fixes the positions of the lower and upper "landings". It is easy to notice that the "stair edges" (u_k, y_k) belong to the line

$$\frac{y - Y^*}{\Delta} = \frac{u - U^*}{Q} \tag{25}$$

and the graph of the static characteristic (except for its 'landings' parts) belongs to a ribbon (see Fig. 2)

$$-1 < \frac{y - Y^*}{\Delta} - \frac{u - U^*}{Q} \leq 0 \tag{26}$$

In other words, since $y = Y^* + j\,\Delta$ if $U^* + j\,Q \leq u < U^* + (j+1)\,Q$ then the static characteristics is for $y_0 \leq y \leq y_n$ formally given by

$$y = Y^* + \Delta \left\lfloor \frac{u - U^*}{Q} \right\rfloor \tag{27}$$

where $\lfloor x \rfloor$ denotes the integer part of x (i.e. x rounded down). It is convenient to express the offsets as relatives values, namely $u^* = \frac{U^*}{Q}$ and $y^* = \frac{Y^*}{\Delta}$. We see that the relative offsets (u^*, y^*) together with the scaling parameters Q, Δ and the reference index k^* uniquely determine

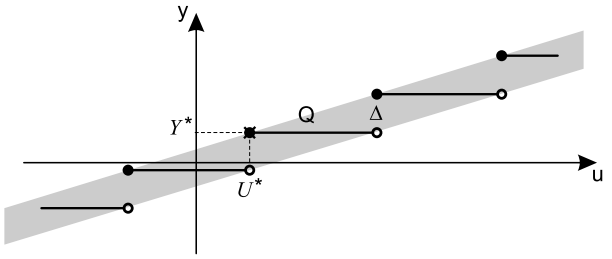

Figure 2. Uniform quantizer's characteristic inside the characteristic's ribbon. The position of the staircase plot within the ribbon is determined by the reference point (denoted by a star). The analog and digital offsets are equal to the coordinates of the reference point.

the uniform quantizer. Obviously, $0 \leq y^* < 1$, while u^* may take any value, both positive and negative. Various typical uniform quantizers can be differentiate by their offset u^*, y^*, for instance (Fig. 3)

$$\begin{aligned} (0, 0) &\Rightarrow \text{rounding-down quantizers} \\ (-1, 0) &\Rightarrow \text{rounding-up quantizers} \\ (-0.5, 0) &\Rightarrow \text{rounding (mid-tread) quantizers} \\ (0, 0.5) &\Rightarrow \text{"mid-riser" quantizers} \end{aligned} \tag{28}$$

where for the last two quantizers the transition levels and the code values, resp., are proportional to odd integers. In fact, use of different offsets by different authors is responsible for apparent differences in their results obtained for the error characteristics.

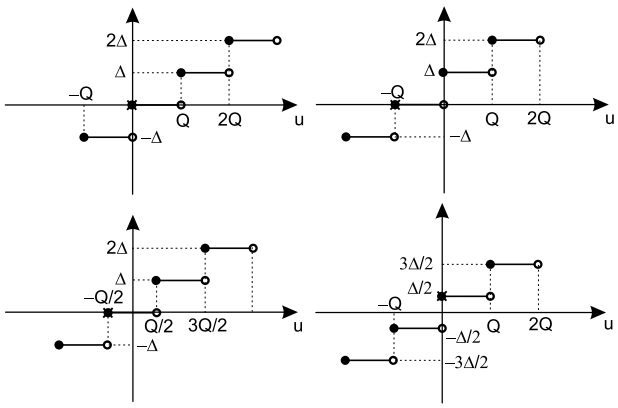

Figure 3. Typical uniform quantizers, the stars denote the reference points. *Top left:* rounding down quantizer, *Top right:* rounding up quantizer, *Bottom left:* mid-tread rounding quantizer, *Bottom right:* mid-riser quantizer

It is convenient to scale the analog values in Q units, and the code values in Δ units, which is equivalent to taking $Q = \Delta = 1$. We can formulate the following proposition for bounded signals:

Proposition 3 (Uniform quantizers) *For the uniform quantizer with unitary Δ and Q, quantization error char-*

acteristics (19, 20) simplify to

$$\mathcal{E}\mathbf{e} = \frac{1}{2}\left(\gamma_0 + \gamma_n\right) - \frac{1}{2}\sum_{i=k_0+1}^{k_n} \widetilde{F}_{\mathbf{u}}(i+u^*) \tag{29}$$

$$\mathcal{E}\mathbf{e}^2 = \mathcal{V}\mathbf{u} + \frac{1}{2}\left({\gamma_0}^2 + {\gamma_n}^2\right) + \sum_{i=k_0+1}^{k_n} \widetilde{H}_{\mathbf{u}}(i+u^*) - (i+y^*-0.5)\,\widetilde{F}_{\mathbf{u}}(i+u^*) \tag{30}$$

where

$$\begin{aligned} \gamma_0 &= y^* - \mathcal{E}\mathbf{u} \\ \gamma_n &= n + y^* - \mathcal{E}\mathbf{u} \end{aligned} \tag{31}$$

and

$$\begin{aligned} k_0 &= 0 \\ k_n &= n \end{aligned} \tag{32}$$

If the analog signal is bounded to $[\underline{u}, \overline{u}]$ *then* (32) *may be replaced by*

$$\begin{aligned} k_0 &= \lfloor \underline{u} - u^* \rfloor \\ k_n &= \lceil \overline{u} - u^* \rceil - 1 \end{aligned} \tag{33}$$

where $\lceil x \rceil$ *denotes rounding up to the nearest integer.*

3.1. Widrow's formulas

We must stress that equivalent error formulas can be in this case derived directly from Widrow's quantization theorem. The results derived in [5] for the offsets $(u^*, y^*) = (-0.5, 0)$

$$\begin{aligned} \mathcal{E}\mathbf{e} &= \sum_{n>0} \frac{(-1)^n}{\pi n} J_0(2\pi n V)\sin(2\pi n C) \\ \mathcal{E}\mathbf{e}^2 &= \frac{1}{12} + \sum_{n>0} \frac{(-1)^n}{\pi^2 n^2} J_0(2\pi n V)\cos(2\pi n C) \end{aligned} \tag{34}$$

can be generalized to any u^*, y^*. Denote by g^* and $e*$ the quantizer characteristic and the quantization error obtained for the offsets $(-0.5, 0)$, resp. We thus have for any u, and any offsets (u^*, y^*)

$$g(u) = y^* + g^*(u - u^* - 0.5) \tag{35}$$

therefore the quantization error $e(u) = g(u) - u$ for any u^*, y^* is related to the quantization error $e^*(u)$ by

$$\begin{aligned} e(u) &= y^* + g^*(u-u^*-0.5) - (u-u^*-0.5) - u^* - 0.5 \\ &= e^*(z)\Big|_{z=u-u^*-0.5} + s^* \end{aligned} \tag{36}$$

where

$$s^* = y^* - u^* - 0.5 \tag{37}$$

Note that $s^* = 0$ on the line which passes through the point $(-0.5, 0)$ and whose slope is equal to 1. In fact, it may be proven that $s^* = 0$ makes the condition for the uniform quantizer to be optimal in the mean-square sense, see [16]. Application of (34) in (36) finally leads to

$$\mathcal{E}\mathbf{e} = s^* + \sum_{n=1}^{\infty} \frac{(-1)^n}{\pi j n}\,\phi^o_{\mathbf{u}}(2\pi n) \tag{38}$$

$$\mathcal{E}\mathbf{e}^2 = \frac{1}{12} + {s^*}^2 + 2\,s^*\,\mathcal{E}\mathbf{e} + \sum_{n=1}^{\infty} \frac{(-1)^n}{\pi^2 n^2}\,\phi^e_{\mathbf{u}}(2\pi n) \tag{39}$$

Here

$$\begin{aligned} \phi^e_{\mathbf{u}}(x) &= \frac{\widetilde{\phi}_{\mathbf{u}}(x) + \widetilde{\phi}_{\mathbf{u}}(-x)}{2} \\ \phi^o_{\mathbf{u}}(x) &= \frac{\widetilde{\phi}_{\mathbf{u}}(x) - \widetilde{\phi}_{\mathbf{u}}(-x)}{2} \end{aligned} \tag{40}$$

are the "even" and "odd" parts of a function $\widetilde{\phi}$,

$$\widetilde{\phi}(x) = e^{-jx(0.5+u^*)}\,\phi_{\mathbf{u}}(x) \tag{41}$$

and $\phi_{\mathbf{u}}$ denotes the characteristic function of $\mathbf{u}(t)$. A serious drawback Widrow's formulas (38, 39) is that they call for infinite summations. On the contrary, the summation in the direct formulas (29, 30) is over finite number of terms. Consequently, the latter lead to more effective and more accurate implementations.

The next Section illustrates our results by specializing to a sine wave input.

4. UNIFORM QUANTIZERS FOR SINE WAVE INPUT

Consider the uniform quantizers with unitary Q and Δ and offsets (u^*, y^*), tested with the use of the offset sine wave input, namely

$$\mathbf{u}(m) = V\cos(Fm + \mathbf{P}) + C \tag{42}$$

where C is the sine wave offset, and $\mathbf{P}$ is distributed uniformly over $[0, 2\pi]$. The error characteristics (29, 30) take form

$$\mathcal{E}\mathbf{e} = \frac{1}{2}\left(\gamma_0 + \gamma_n\right) - \frac{1}{\pi}\sum_{k=k_0+1}^{k_n} \arcsin\kappa_k \tag{43}$$

$$\mathcal{E}\mathbf{e}^2 = \frac{1}{2}\left(V^2 + \gamma_0^2 + \gamma_n^2\right) - \frac{2V}{\pi}\sum_{k=k_0+1}^{k_n} \sqrt{1-\kappa_k^2} + (\kappa_k + \beta)\,\arcsin\kappa_k \tag{44}$$

where we denoted

$$\begin{aligned} \gamma_0 &= k_0 + y^* - C \\ \gamma_n &= k_n + y^* - C \\ \kappa_k &= \frac{k + u^* - C}{V} \\ \beta &= \frac{y^* - u^* - 0.5}{V} = \frac{s^*}{V} \end{aligned} \tag{45}$$

and

$$\begin{aligned} k_0 &= \lfloor C - V - u^* \rfloor \\ k_n &= \lceil C + V - u^* - 1 \rceil \end{aligned} \tag{46}$$

Note that the role of the signal offset C is not identical to the role of the analogue offset U^* of the quantization characteristic.

4.1. Relation to Widrow's formulas

A specialization of Widrow's formulas (38, 39) to the sine-wave input (42) leads to relations equivalent to (43, 44). Since for the considered sine-wave we have

$$\phi_{\mathbf{u}}(x) = J_0(xV)\exp(jxC) \tag{47}$$

then by (38, 39) we obtain

$$\mathcal{E}\mathbf{e} = s^* + \sum_{n>0} \frac{(-1)^n}{\pi n} J_0(2\pi n V)\sin(2\pi n \widetilde{C}) \tag{48}$$

$$\mathcal{E}\mathbf{e}^2 = \frac{1}{12} + s^{*2} + 2s^*\mathcal{E}\mathbf{e} + \sum_{n>0} \frac{(-1)^n}{\pi^2 n^2} J_0(2\pi n V)\cos(2\pi n \widetilde{C}) \tag{49}$$

where s^* is given by (37) and

$$\widetilde{C} = C - u^* - \frac{1}{2} \tag{50}$$

While (48, 49) *look* more compact then their direct equivalents (43, 44), they call for infinite summations. Moreover, each term requires that a value of the Bessel function J_0 is calculated. Consequently, finite-term approximations are necessary and the resulting accuracy is inferior to (43, 44). Accuracy problems can be expected when the infinite sums in (48, 49) are replaced with their finite n-term counterparts. Define the approximation error

$$\Delta\mathcal{E}\mathbf{e}(n; C, V) = \mathcal{E}\mathbf{e}_n(C, V) - \mathcal{E}\mathbf{e}(C, V) \tag{51}$$

where $\mathcal{E}\mathbf{e}(C, V)$ denotes the exact value calculated with the use of our formula, and $\mathcal{E}\mathbf{e}_n(C, V)$ denotes the n-term approximation of Widrow's formula. A plot of this error for chosen values of C and V is shown in Fig. 4. The biggest errors, very slowly decreasing with n, correspond to such values of C and V at which $\mathcal{E}\mathbf{e}(C, V)$ is not differentiable. Since the exact values of $\mathcal{E}\mathbf{e}(C, V)$ are of order of 10^{-2}, the approximation error at those points cannot be accepted in applications. Note that non-differentiability is not visible at all in Widrow's formulas and only our approach shows this phenomenon.

4.2. Example: Mid-tread quantizer, sine-wave input

The error characterization of the mid-tread quantizers can be obtained directly from the last example, by taking the offsets as $(u^*, y^*) = (-0.5, 0)$. We thus obtain the

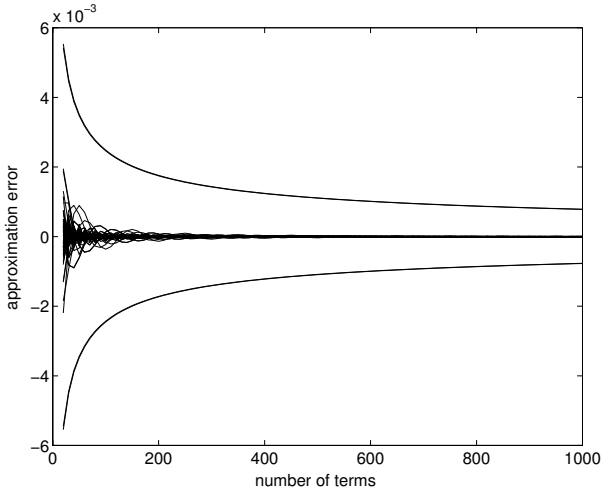

Figure 4. Finite-terms approximation error for Widrow's $\mathcal{E}\mathbf{e}$ vs. the number of summation terms, for chosen values of C and V. Please note that the external curves *are not* the asymptotes: they correspond to the extremal error characteristics obtained for integer even values of $2C + 2V$ and $2C-2V$. The error, otherwise negligible for about 200 summation terms, is for these special values of C and V unacceptably high even for thousand summation terms. Consequently, the use of Widrow's formulas for C and V close to these values is highly not recommended.

result derived in [14]

$$\begin{aligned}
\gamma_0 &= k_0 - C \\
\gamma_n &= k_n - C \\
\kappa_k &= \frac{k - C - 0.5}{V} \\
\beta &= 0 \\
k_0 &= \left\lfloor C - V + \frac{1}{2} \right\rfloor \\
k_n &= \left\lceil C + V - \frac{1}{2} \right\rceil
\end{aligned} \tag{52}$$

Actually, in [14] we used $k_n = \lfloor C+V+\frac{1}{2} \rfloor$ which involves an additional term for integer $C + V - \frac{1}{2}$. As discussed earlier, this does not influence the overall sum. For $C = 0$ and the mid-tread quantizer, similar result was obtained in [11].

The mean quantization error $\mathcal{E}\mathbf{e}$ and the mean-square quantization error $\mathcal{E}\mathbf{e}^2$ are both functions of C and V. The mean quantization error, Fig. 5, is periodic as a function of C, with the basic period equal to 1. The dependence on V is more complicated, namely the graph shows 'damped periodicity', in a sense that

$$|\mathbf{e}(C, V)| \geq |\mathbf{e}(C, V + 1)| \geq \ldots \tag{53}$$

for every C, V. One can prove that $\mathcal{E}\mathbf{e}$ is not differentiable at certain C, V, see Fig. 5. In fact, it is easy to show that $\mathcal{E}\mathbf{e}$ is not differentiable if either of quantities $C+V+0.5$,

$C - V + 0.5$ are integer. Consequently, the entire domain is divided into squares by lines

$$\begin{aligned} C + V &= n + 0.5 \\ C - V &= m + 0.5 \end{aligned} \tag{54}$$

where m and n are integers. The borders of these squares form 'edges' along which $\mathcal{E}\mathbf{e}$ is not differentiable, Fig. 7.

Similar analysis can be performed for the mean-square quantization error. Similarly to the mean quantization error, it is periodic in C, with the basic period equal to 1, and 'damped periodic' in V. Contrary to $\mathcal{E}\mathbf{e}$, the mean-square error is everywhere 'smooth', and show various shapes, depending on particular parameters, see Fig. 6. It attains minima and maxima at the same lines at which the mean error is not differentiable, Fig. 8.

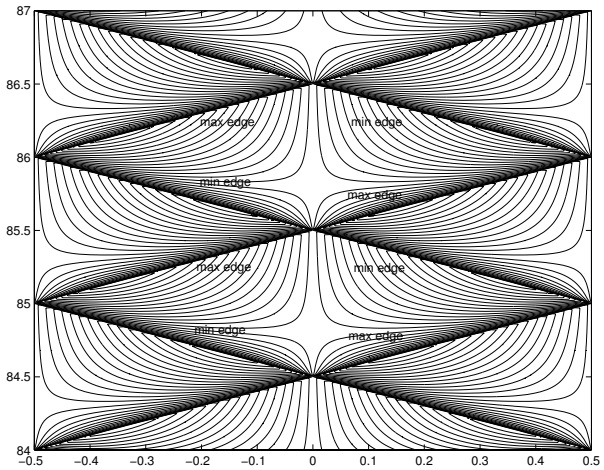

Figure 7. Contour plot of the mean quantization error as a function of the DC offset C (horizontal axis) and the sine wave amplitude V (vertical axis). One period of C and three periods of V are shown. Note the 'edges' along which the function is not differentiable. Maximum of the function is attained at points at the edges.

For comparison, Widrow's formulas (48, 49) simplify in this case to the well known relations (34). Again, they only *look simpler* than our direct formulas, yet they involve infinite term summations and may lead to large numerical errors.

5. CONCLUSIONS

While a direct approach to quantization error can be applied to any static quantizer, e.g. non-uniform or non-monotonic, it is instructive to use this approach in the case for which another approach exists, namely for uniform quantizers. Here we compare the results obtained with both approaches. In the general case, our direct approach enables to avoid certain limitations of the classical approach, like the assumption of equal bin length. Our approach shows yet its merits even for the uniform case, since it leads to finite summations in error formulas, unlike the infinite summations involved in classical

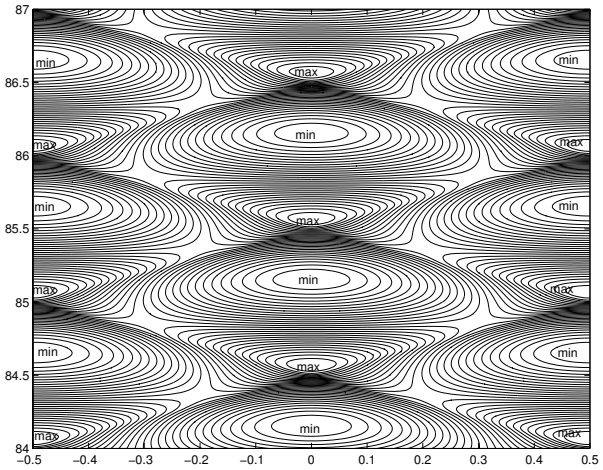

Figure 8. Contour plot of the mean-square quantization error as a function of the DC offset C (horizontal axis) and the sine wave amplitude V (vertical axis). One period of C and three periods of V are shown. The minima are attained along upper left-lower down diagonals, and the minima are along the other diagonal.

Widrow's formulas. Consequently, the direct approach is very well fitted to applications. In particular, it can be applied to effective resolution analysis to make the efr measurements depend only on the quantizer's parameters rather than the test signal parameters. Although the last revision of IEEE Standard 1057 provides a few new matrix algorithms for the sine-wave test, it does not support with an adequate theory of the reference quantizer. This paper is aimed into making the uniform quantizer a reference one for real-life experimental conditions.

REFERENCES

[1] B. Widrow, "Statistical analysis of amplitude-quantized sampled-data system," *Trans. AIEE*, Pt. II, Applications and Industry, Vol. 79, pp. 555–568, Jan. 1961.

[2] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-25, pp. 442–448, Oct.1977.

[3] M. F. Wagdy and Wai-Man Ng, "Validity of Uniform Quantization Error Model for Sinusoidal Signals Without and With Dither," *IEEE Trans. Instrumentation and Measurements*, Vol. 38, no. 3, pp. 718–722, June 1989.

[4] M. Wagdy and S.S. Awad, "Determining adc effective number of bits via histogram testing," *IEEE Trans. Instr. Measurement*, vol. 40, pp. 770-772, 1991.

[5] K. Hejn and A. Pacut, "Generalized model of the quantization error — A unified approach," *IEEE Trans. Instrumentation and Measurements*, Vol. 45, no. 1, pp. 41-44, Feb. 1996.

[6] K. Hejn and A. Pacut, "Error correction in effective resolution techniques," *IEE Conf. Publication on Advanced A/D and D/A Conversion Techniques and Their Applications, Glassgow, UK, July 1999*, No. 466, pp. 70-73, 1999.

[7] K. Hejn and A. Pacut, "Improved definition of adc effective resolution," *Computer Standards and Interfaces*, vol. 23, No. 2, 2001.

[8] D. T. Sherwood, "Some theorems on quantization and an example using dither", *Conf. Record, XIX Asilomar Conference*

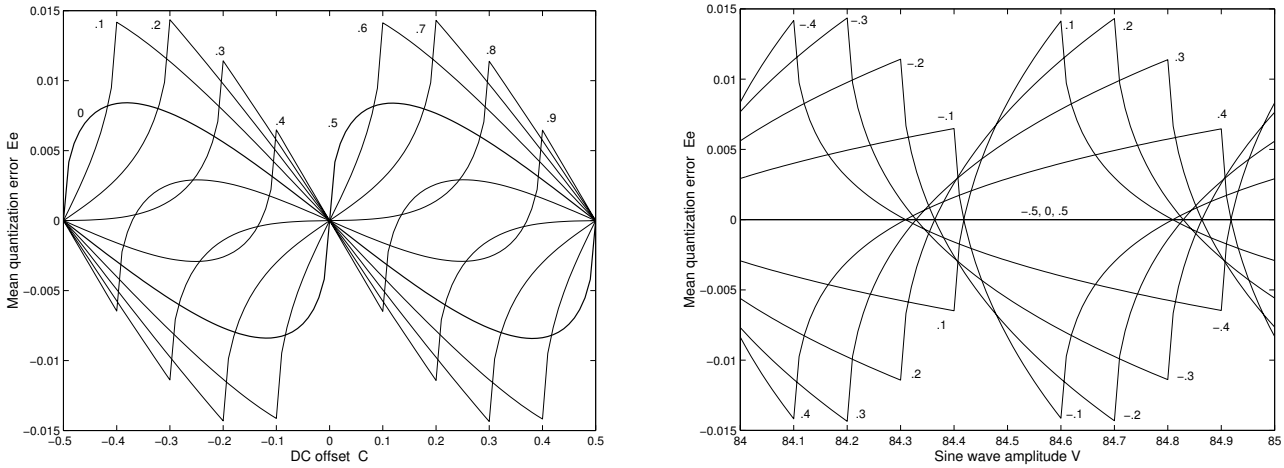

Figure 5. The mean quantization error as a function of DC offset C, with the sine wave amplitude V as a parameter (left; actual V are equal to 84 plus the marked values) and as a function of the sine wave amplitude V, with the DC offset C as a parameter (right).

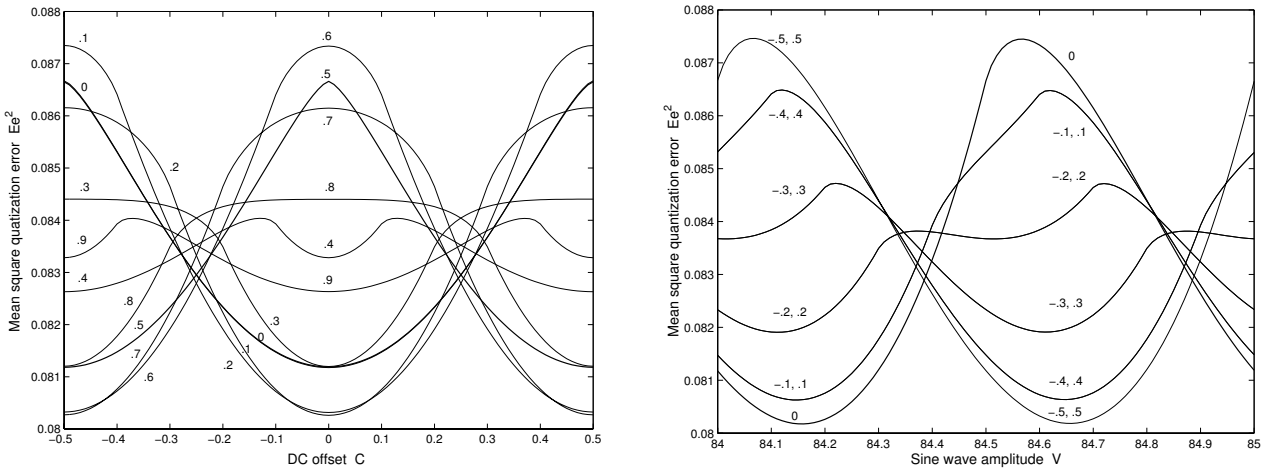

Figure 6. The mean-square quantization error as a function of the DC offset C, with the sine wave amplitude V as a parameter (left; actual V are equal to 84 plus the marked values) and as a function of the sine wave amplitude V, with the DC offset C as a parameter (right).

on Circuits, Systems and Computers, Pacific Grove, California USA, November 6–8, 1985.

[9] IEEE Std. 1057-1994 (Revision of IEEE Std. 1057-1989), "IEEE Standard for Digitizing Waveform Recorders," Published by IEEE Inc. in December 30, 1994, SH94245.

[10] IEEE Std. 1241, "Standard for Terminology and Test Methods for Analog-to-Digital Converters," Draft printed May 12, 1997.

[11] R. M. Gray, "Quantization noise spectra," *Trans. Information Theory*, Vol.IT-36, pp. 1220–1244, November 1990.

[12] K. Hejn, A. Pacut, and L. Kramarski, "The effective resolution measurement in scope of sine-fit test," *IEEE Trans. Instrumentation and Measurements*, vol. 47, pp. 45–50, 1998.

[13] A. Pacut, K. Hejn, and L. Kramarski, "Generalized model of the quantization error — A numerical approach," *2nd International Workshop on ADC Modeling and Testing*, Tampere, Finland, 1-6 June 1997, published on CD-ROM by Finnish Society of Automation, see also http://www.tut.fi/mit/imeko.

[14] A. Pacut and K. Hejn, "Equivalence of Widrow's and Gray's approaches to uniform quantizers", *Computer Standards & Interfaces*, vol. 19, No. 3–4, Sep. 1998, pp. 205–212.

[15] A. Pacut and K. Hejn, "Analog-to-digital converters: Towards a generalization of Widrow's Theorem," *Proc. of IEEE IMTC/98, St. Paul, Minnesota, May 1998*, vol. 2, pp. 1190–1197, 1998.

[16] A. Pacut and K. Hejn, "Analog-to-digital converters: direct approach," *submitted to IEEE Trans. on Instrumentation and Measurements*, 2000.

[17] K. Hejn, J. Jdrachowicz, and A. Leniewski "Experimental verification of a new method for effective resolution valuation," *Proc. of XVI IMEKO World Congress IMEKO 2000*, Wien, Austria, vol. X, pp. 159–164, Sep. 2000.