A STATISTICAL CHARACTERIZATION OF THE UNIFORM QUANTIZATION PROCESS

Andrei Cubitchi, Alexandru Isar, Electronics and Telecommunications Faculty Timisoara, 2 Bd. V. Parvan, 1900 Timisoara, Romania, isar@hermes.ee.utt.ro

Abstract. Any digital signal processing of a continuous in time signal can be done after the sampling and the quantization of this signal. The aim of this paper is a statistical analysis of the quantization process. The results of this analysis will stay at the basis of the design of analog to digital converter (ADC) used in digital speech processing. First a statistical analysis of first degree is realized. The relation between the characteristic functions of the ADC input and output signals is established. The Widrow’s quantization theorem is analyzed. This theorem gives the conditions to be satisfied by the signal to be quantified for its perfect reconstruction after the quantization process. The disadvantages of this theorem are envisaged and some solutions to decrease the effects of these disadvantages, in the case of different classes of signals used in practice, are presented. Then a statistical analysis of second order is performed. The statistical correlation function of the random variables from the output of the ADC is computed. Some important input signal classes are considered. The advantage of use of dither signals is highlighted. Some conclusions useful for the processing of speech signals are presented.

1. Introduction

The expression of the probability density of the random variable \( \hat{S} \), at the output of an uniform ADC is:

\[
p_{\hat{S}}(x) = \sum_{k=-K}^{K} \left[ p\left(\hat{S} = \hat{x}_k\right) \delta(x - \hat{x}_k) \right]
\]

\[+ \sum_{k=-\infty}^{-K-1} 0 \cdot \delta(x - \hat{x}_k) + \sum_{k=K+1}^{\infty} 0 \cdot \delta(x - \hat{x}_k)\]  

(1)

[1], [2], where \( S \) represents the input random variable, because \( p\left(\hat{S} = \hat{x}_k\right) = 0 \) for \( k < -K \) and for \( k > K \), (the device has two saturation zones). This is the reason why it can be written:

\[
p_{\hat{S}}(x) = \sum_{k=-\infty}^{\infty} p\left(\hat{S} = \hat{x}_k\right) \delta(x - \hat{x}_k)
\]

(2)

The characteristic function of the output random variable is:

\[
\Phi_{\hat{S}}(u) = \sum_{k=-\infty}^{\infty} \text{sinc} \frac{q}{2} \left( u + k \frac{2\pi}{q} \right) \Phi_S \left( u - k \frac{2\pi}{q} \right)
\]

(3)

[3], where \( \Phi_S(u) \) represents the characteristic function of the input random variable. This is the reason why the quantization system has the model presented in figure 1.

Because an ideal sampling is used, if the hypotheses of the WKS theorem are respected, [4], [5], the signal \( A(x) \) can be reconstructed from the signal \( p_{\hat{S}}(x) \). To do this task is necessary that the support of the function \( \Phi_S(u) \) to be compact.

The characteristic function of the probability density \( A(x) \) has the expression:

\[
\Phi_V(u) = \Phi_S(u) \cdot \text{sinc} \frac{q}{2} u
\]

Using the notation:

\[
\Phi_U(u) = \text{sinc} \frac{q}{2} u
\]

it can be written:

\[
\Phi_V(u) = \Phi_S(u) \cdot \Phi_U(u)
\]

Because the characteristic function of the random variable \( V \) is the product of the characteristic functions of the random variables \( U \) and \( S \), the random variable \( V \) represents the sum of the random variables \( U \) and \( S \) and these random variables are independent, [6],

![Figure 1. The model of the quantization system.](image-url)
\[ V = U + S \]  

If \( \Phi_S(u) \) has compact support and if a small enough value for the quantization step \( q \) is selected, then using \( \Phi_S(u) \) the function \( \Phi_V(u) \) can be reconstructed. Subtracting the random variable \( U \) from the random variable \( V \), the random variable \( S \) can be reconstructed. So, if the mentioned hypotheses are satisfied, then from the output random variable can be reconstructed the input random variable. Hence the quantization process can be inverted. This is the aim of the Widrow’s quantization theorem, proposed in 1960, [7].

2. The quantization theorem

The enunciation of the quantization theorem, already proved (in the previous paragraph), is the following:

**The necessary and sufficient condition that \( p_S(x) \) be perfectly reconstructed from \( p_S(x) \) is that \( \sup \{ \Phi_S(u) \} \) have a length of \( 2u_M \) and to work with a quantization step \( q \):

\[ \frac{2\pi}{q} > 2u_M \]  

If the hypotheses of this theorem are satisfied then filtering with an ideal low-pass filter the signal with the spectrum \( \Phi_S(u) \), [8], we can obtain the characteristic function of the **reconstructed random variable** \( R \):

\[
\Phi_R(u) = \Phi_V(u) = \Phi_S(u) \cdot \Phi_U(u)
\]

from where we can obtain the probability density function of the input random variable.

The compact support hypotheses for the functions \( p_S(x) \) and \( \Phi_S(u) \) are in contradiction because the second function represents the Fourier transform of the first one. The support of \( p_S(x) \) is compact because in the signal processing chain there are other operations, before the quantization, implemented with systems with saturated input-output characteristics. So, the support of \( \Phi_S(u) \) can not be compact. Hence in practice the quantization system can not be perfectly inverted. In the following are presented some supplementary conditions necessary for the reconstruction of the probability density of the input random variable starting from the probability density of the output random variable.

3. The computation of the moments of the random variable \( R \)

For the beginning it must be observed (on the base of its characteristic function) that the random variable \( U \) is uniformly distributed in the interval \( \left[ -\frac{q}{2}, \frac{q}{2} \right] \). In the following the k-th order moment of the random variable \( R \) is computed:

\[
M[R^k] = \frac{1}{j^{k} \, du} \left( \Phi_R(u) \right)_{u=0} = \sum_{p=0}^{k} C_p^k \left( \frac{1}{j^p} \Phi_S(u) \right) \Phi_{U^{(k-p)}}(o) = \sum_{p=0}^{k} C_p^k M[S^p] M[U^{(k-p)}]
\]

For \( k=1 \), the last relation becomes:

\[
\]

or:

\[
M[R] = M[S]
\]

So the random variable \( R \) has the same average like the input random variable. For \( k=2 \), it can be written:

\[
\]

or:

\[
M[R^2] = M[S^2] + \frac{q^2}{12}
\]

Hence the power of the random variable \( R \) can be computed summing the powers of the random variables \( S \) and \( U \). Because these random variables are independent (hence not correlated) the random variable \( U \) can be regarded like a noise proper for the quantization system. This is the reason why using the last relation we can compute the signal to noise ratio at the output of the quantization system:

\[
RSZ = \frac{\sigma_S^2}{q^2} = \frac{12\sigma_S^2}{q^2}
\]

For \( k=3 \), we obtain:

\[
\]

or:

\[
M[R^3] = M[S^3] + \frac{M[S]^2 q^2}{4}
\]
4. The general case

In the following, the hypothesis of compact support for the characteristic function $\Phi_S(u)$ is rejected and the moments of the random variable $\hat{S}$ are computed. Using the relation (3) it can be written:

$$M[\hat{S}^k] = \sum_{p=0}^{k} C_k^p M[S]^p M[U^{(k-p)}] +$$

$$+ \frac{1}{j^k} \sum_{l=-\infty}^{l=\infty} \sum_{p=0}^{k} C_k^p \Phi_S^{(p)}(\frac{2\pi}{q}) \Phi_S^{(k-p)}(\frac{2\pi}{q})$$

(10)

Comparing the relation (6) and (10) the error due to the rejection of the hypothesis of compact support for $\Phi_S(u)$ can be obtained:

$$\varepsilon_k = \frac{1}{j^k} \sum_{l=-\infty}^{l=\infty} \sum_{p=0}^{k} C_k^p \Phi_S^{(p)}(\frac{2\pi}{q}) \Phi_S^{(k-p)}(\frac{2\pi}{q})$$

Tacking into account the expressions of the derivatives of the characteristic function of the random variable $U$ we can write:

$$\varepsilon_1 = \frac{q}{2\pi j} \sum_{l=-\infty}^{l=\infty} (-1)^l \Phi_S(\frac{2\pi}{q})$$

So, the relation between the averages of the random variables $S$ and $\hat{S}$ is:

$$M[S] = M[\hat{S}] = \frac{q}{2\pi j} \sum_{l=-\infty}^{l=\infty} (-1)^l \Phi_S(\frac{2\pi}{q})$$

(11)

For $k=2$ we obtain the following expression of the error:

$$\varepsilon_2 = -\frac{q}{2\pi} \sum_{l=-\infty}^{l=\infty} \frac{q}{2\pi} \Phi_S(\frac{2\pi}{q}) (-1)^{l+1} +$$

$$+ 2\Phi_S(\frac{q}{2\pi}) (-1)^l$$

So, the second order moment of the output random variable is:

$$M[\hat{S}^2] = M[S^2] + \frac{q^2}{12} - \varepsilon_2$$

(12)
where \( U_1, U_2 \) are independent random variables, uniformly distributed in the interval \([-q/2, q/2]\).

**E2.** The random variable with the characteristic function:

\[
\Phi_S(u) = \sin \frac{qu}{2} \frac{(2K-1)qu}{2}
\]

can be generated using the system in figure 2, but the random variable \( U_2 \) must be, this time, uniform distributed in the interval \([- (2K-1)q/2, (2K-1)q/2]\).

The other input random variable, \( U_1 \), is the same like in the example E1.

**E3.** The random variable with the characteristic function:

\[
\Phi_S(u) = \left( \frac{2}{2K-1}qu \right)^2
\]

can be generated using the system in figure 2, but both the input random variables must be, this time, uniform distributed in the interval \([- (2K-1)q/2, (2K-1)q/2]\).

The probability density of this random variable is represented in figure 3.

**E4.**

\[
\Phi_S(u) = \prod_{m=1}^{M} \sin \frac{A_{m}u}{2} = \sin \frac{A_{1}u}{2} \prod_{m=3}^{M} \sin \frac{A_{m}u}{2}
\]

where \( A_{m} \) is a multiple of \( q \). The system for the generation of this random variable is presented in the next figure, where \( U_{um} \) are random variables uniform distributed in the intervals \([- A_{m}, A_{m}]\).

This example includes a lot of random variables. In conformity with the central limit theorem, when \( M \to \infty \) the random variable \( S \) becomes a normal distributed (gaussian) random variable. This is the reason why we can affirm that in the case of a gaussian random variable the following relations are satisfied in an asymptotic manner:

\[
M[S] = M[S]
\]

\[
M[S^2] = M[S^2] + \frac{q^2}{12}
\]

It must be observed that a gaussian random variable do not can be connected at the input of a quantizer because this random variable takes very high values and these values are not acknowledged by the circuits connected at the input of the ADC.

All the elements of the input classes presented in this paragraph generate output random variables for the ADC satisfying the relations (14) and (15), specific for the Widrow’s quantization theorem without satisfying the hypotheses of this theorem. So the relations (14) and (15) are satisfied for a large class of input signals. This is the reason why the expression of the output signal to noise ratio in (8’) is a good estimation of this parameter.

**5.1. The use of dither**

In the following we consider another class of input random variables for an ADC. These random variables have a characteristic function of the form:

\[
\Phi_S(u) = \sin \frac{q}{2} \cdot \sin \frac{qu}{2}
\]

Let:

\[
\Phi_{U_1}(u) = \sin \frac{q}{2} \cdot \sin \frac{qu}{2}
\]

The conditions for the verification of the following relation:

\[
\Phi_{S_1}(l \frac{2\pi}{q}) = 0, \quad l \neq 0
\]

are investigated.

A solution of this problem is for example:

\[
\Phi_{S_1}(u) = \sin \left( \frac{u}{q} \right) \Phi_{S_1}(u)
\]

Using the notation:

\[
\Phi_{S_1}(u) = \sin \left( \frac{u}{q} \right) \Phi_{S_1}(u)
\]
\[ \Phi_{U_1}(u) = \sin \left( q \frac{u}{2} \right) \]

the second derivative of the characteristic function of the input random variable can be computed:

\[ \Phi''(l \frac{2\pi}{q}) = 2 \left( \sin \left( q \frac{u}{2} \right) \right)^2 \left| \Phi_{S_3}(l \frac{2\pi}{q}) \right| (18) \]

To vanish this expression for all the values of the integer \( l \) with the exception of 0 (see the relation (17)), is necessary that \( \Phi_{S_3}(l \frac{2\pi}{q}) \) to be equal with zero for every value of \( l \) with the exception of 0. This new condition is accomplished if, for example:

\[ \Phi_{S_3}(u) = \sin \left( q \frac{u}{2} \right) \Phi_{S_1}(u) \]

Using the notation:

\[ \Phi_{U_1}(u) = \sin \left( q \frac{u}{2} \right) \]

the expression of the characteristic function of the input random variable with the first two derivatives equal with 0 for \( l \) non-null becomes:

\[ \Phi_S(u) = \prod_{l=1}^{3} \Phi_{U_1}(u) \Phi_{S_3}(u) \] (19)

Let \( D \) be the random variable obtained summing the independent random variables \( U_1, U_2, U_3 \):

\[ D = U_1 + U_2 + U_3 \]

Because:

\[ \Phi_D = \prod_{l=1}^{3} \Phi_{U_l} \]

the relation (19) becomes:

\[ \Phi_S(u) = \Phi_D(u) \Phi_{S_3}(u) \]

So if to an input random variable, \( S_3 \), is added the dither \( D = U_1 + U_2 + U_3 \) and the random variable \( S \) is connected at the input of an ADC then at the output of this device is obtained the random variable \( \hat{S} \) with the moments:


and:

\[ M[S^2] = M[S^2] + \frac{q^2}{12} = M[S_3^2] + \frac{q^2}{3} \] (21)

So if is realized the quantization of the sum of the signal \( S_3 \) and of a dither and then is performed the computation of the average of the result of the quantization, the average of the signal \( S_3 \) can be obtained without any error. If the component \( S_3 \) of the input signal is a deterministic signal it can be written:

\[ S_3 = M[S_3] \]

So if is added to this signal a dither and if an averager is used after the quantization process then at the output of the averager the signal \( S_3 \) is obtained without any distortion. So the operation of adding a dither conducts (due to the use of the averager at the output of the quantization system) to the possibility of the perfect reconstruction of a signal from its quantized version. This procedure for the artificial increasing of the resolution of an ADC was used at Hewlett-Packard to construct some waveform recorders.

6. The second order statistical analysis

The expression of the second order probability density of the two-dimensional random variable, at the output of the ADC is:

\[ p_{\hat{S}}(x_1, x_2) = \sum_{k=\text{integer}} A(\hat{x}_1, \hat{x}_2) \delta(x_1 - \hat{x}_1, x_2 - \hat{x}_2) \] (22)
where:

\[
A(\hat{x}_{1k}, \hat{x}_{2m}) = F_S\left(\hat{x}_{1k} + \frac{q}{2}, \hat{x}_{2m} + \frac{q}{2}\right) - F_S\left(\hat{x}_{1k} - \frac{q}{2}, \hat{x}_{2m} + \frac{q}{2}\right) - F_S\left(\hat{x}_{1k} + \frac{q}{2}, \hat{x}_{2m} - \frac{q}{2}\right) + F_S\left(\hat{x}_{1k} - \frac{q}{2}, \hat{x}_{2m} - \frac{q}{2}\right)
\]

and \(F_S(x_1, x_2)\) represents the mutual repartition function of the two-dimensional input random variable \(S\).

So the probability density of the two-dimensional random variable at the output of the ADC is obtained by ideal sampling of the function \(A(x_1, x_2)\). The expression of the characteristic function of the two-dimensional output random variable is:

\[
\Phi_S(u_1, u_2) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \sin(c) \left(\frac{qu_1}{2} + k\pi\right) \cdot \sin(c) \left(\frac{qu_2}{2} + l\pi\right) \cdot \Phi_S\left(u_1 + k\frac{2\pi}{q}, u_2 + l\frac{2\pi}{q}\right)
\]

(23)

The model of the quantization system inspired by the last relation is presented in the following figure.

![Figure 6. The model of the quantization system inspired by the second order statistical analysis.](image)

Using this system the Widrow’s two-dimensional quantization theorem can be proved. The enunciation of this theorem is the following:

**The necessary and sufficient condition for the perfect reconstruction of the probability density of the input random variable \(p_S(x_1, x_2)\) starting from the output probability density \(p_S(x_1, x_2)\) is that the length of the greatest dimension of sup \(p(\Phi_S(u_1, u_2))\), \(2u_M\), to satisfy the condition:**

\[
\frac{2\pi}{q} > 2u_M
\]

This theorem can be proved using the same method like in the case of the Widrow’s one-dimensional quantization theorem (see the paragraph 1 of this paper).

Unfortunately all the two-dimensional input random variables used in practice have probability density with compact support. So their characteristic functions don’t have compact support. Hence this theorem can not be applied in practice, the reconstruction of the input signal being affected by the quantization error.

In the following are computed the moments of the two-dimensional random variable \(\hat{S}\). The symetry properties of the function \(\Phi_S(u_1, u_2)\):

\[
\Phi_S(-u_1, u_2) = \Phi_S^*(u_1, u_2)
\]

\[
\frac{\partial}{\partial u_1} \Phi_S(0, -u_2) = -\frac{\partial}{\partial u_1} \Phi_S^*(0, u_2)
\]

\[
\frac{\partial}{\partial u_2} \Phi_S(-u_1, 0) = -\frac{\partial}{\partial u_2} \Phi_S^*(u_1, 0)
\]

are used to accomplish this task.

Using the relation (23) the correlation of the output random variable \(\hat{S}\) can be computed:

\[
R_S(x_1, x_2) = R_S(x_1, x_2) - \frac{4}{\pi} \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} q^2 \left(\frac{1}{k^2} + \frac{1}{l^2}\right) \cos\left(\frac{2\pi}{q} k u_1 + \frac{2\pi}{q} l u_2\right)
\]

(24)

This relation was obtained in [9], too. If the hypotheses of the two-dimensional quantization theorem are satisfied then the last relation becomes:

\[
R_S(x_1, x_2) = R_S(x_1, x_2)
\]

(25)

So in this case the signal \(S\) pass trough the quantization system without any supplementary correlation. In other cases, more frequently found in practice, the quantization system introduces a supplementary correlation of the input signal. The relation (24) proves that the correlation of the signal at the output of the quantization system is bigger than the correlation of the signal at the input of the quantization system. The correlation of the samples of a signal proves that every sample contains two kind of information, the information specific for that sample and an information specific for the neighborhood samples. This is why the representation with a signal with correlated samples is redundant. The reduction of this redundancy is a way to realize the compression of the considered signal. This is the reason why the reduction of the relation (24) to the relation (25) is very important.

### 6.1 Remarkable classes of input signals

The relation (24) can be reduced to the relation (25) without satisfying the hypotheses of the two-dimensional quantization theorem if the following conditions are satisfied:
In this case \( S \) is the sum of the two independent random variables \( S = (S_1, S_2) \), satisfying the conditions (26).

**E1.** \( S_1 \) and \( S_2 \) are independent random variables uniform distributed in the interval \( \left[-\frac{q}{2}, \frac{q}{2}\right] \).

**E2.** \( S \) is a two-dimensional random variable with the probability density:

\[
p_S(x_1, x_2) = \begin{cases} \frac{1}{ABq} & x_1 \in \left[\frac{-Aq}{2}, \frac{-Aq}{2}\right] \times \left[\frac{-Bq}{2}, \frac{-Bq}{2}\right] \\ 0 & \text{in rest} \end{cases}
\]

\( A, B \in \mathbb{N} \)

If \( A = B = (2K - 1) \) then \( S \) is a two-dimensional uniform random variable.

**E3.** \( S \) is a two-dimensional random variable with the characteristic function:

\[
\Phi_S(u_1, u_2) = \prod_{m_1=1}^{M_1} \sin c \frac{A_{m_1} u_1 q}{2} \prod_{m_2=1}^{M_2} \sin c \frac{B_{m_2} u_2 q}{2}
\]

\( A_{m_1} \in \mathbb{N}, \forall m_1 \in \{1, ..., M_1\} \)

\( B_{m_2} \in \mathbb{N}, \forall m_2 \in \{1, ..., M_2\} \)

It must be observed that for \( M_1 \to \infty \) and \( M_2 \to \infty \) the two-dimensional random variable \( S \) is a couple of two gaussian random variables (due to the central limit theorem).

**E4.** The two-dimensional random variable \( S \) with the characteristic function:

\[
\Phi_S(u_1, u_2) = \Phi_U(u_1, u_2) \Phi_{S_1}(u_1, u_2)
\]

where \( U \) is a two-dimensional random variable uniformly distributed in the domain \( \left[-\frac{q}{2}, \frac{q}{2}\right] \times \left[-\frac{q}{2}, \frac{q}{2}\right] \).

In this case \( S \) is the sum of the two independent random variables \( S_1 \) and \( U \):

\[
S = S_1 + U
\]

The random variable \( U \) represents the dither.

### 7. Conclusion

In this paper was made a statistical analysis of the uniform quantization process. The Widrow's one-dimensional and two-dimensional quantization theorems were proved. These theorems indicate a way for the reconstruction of a signal from its quantified version. The hypotheses of these theorems are too restrictive. The signals satisfying these hypotheses cannot be found in practice. This is the reason why in this paper are presented some new results concerning the quantization of the signals members of some special classes. These classes are presented in the paragraphs 5 and 6.1 for the case of the Widrow's one-dimensional quantization theorem and 6.1 for the case of the Widrow's two-dimensional quantization theorem. The precision of the reconstruction after the quantization of the members of these classes is the same like the precision of reconstruction specified by the Widrow's quantization theorems but these signals do not satisfy the hypotheses of those theorems. To obtain such classes is sufficient to satisfy the conditions presented in the relation (13), that corresponds to the one-dimensional quantization theorem or the conditions presented in the relation (26), that correspond to the two-dimensional quantization theorem. These conditions are original, being presented for the first time in this paper. Between the signal classes presented in this paper, in the paragraphs already mentioned, there are some, large enough, frequently used in practice. This is the case for the class presented in the example E4 from the paragraph 5, or of the class of signals with dither, presented in the paragraph 6.1 and in the example E4 from the paragraph 6.1. The use of the dither at the input of a quantization system and of an averager at the output of this system is a very elegant solution for the artificial increasing of the resolution of an ADC. The relation between the correlation functions of the signals from the input and output of a quantization system presented in (24) is a very important result too. It can be used for the design of an ADC used for compression purposes. There are a lot of methods to decrease the values of the correlation function of the signal from the output of a quantization system. These methods represent the aim of different compression systems. The speech is a not-stationary signal. The sampling of such signals is described in [10]. This signal can be segmented obtaining stationary segments. Every such segment can be quantified using a specified resolution. To decrease the values of the correlation function at the output of every such quantization system several methods can be used. Some of them are based on orthogonal transforms. A very interesting orthogonal transform is presented in [11]. The signal to noise ratio of the signal obtained at the output of a quantization system can be enhanced using one of the methods presented in [12].
More about the quantization systems used for the compression of speech are presented in [13]. In fact the quantization is a coding procedure. Some such procedures are presented in [14]. More about the compression methods based on orthogonal transforms can be found in [15]. To compress the speech, a particular quantization system, built in accord with the particularities of this kind of signals, must be used. These particularities are very well presented in [16]. Other types of quantization can be used too for the speech compression, like the non-uniform quantization or the vector quantization. This last type of quantization is very modern, being presented in a lot of references, [17] being such an example.

References


Other useful references


