Statistical Theory of Quantization

Bernard Widrow, *Life Fellow, IEEE*, István Kollár, *Senior Member, IEEE*, and Ming-Chang Liu

Abstract— **The effect of uniform quantization can often be modeled by an additive noise that is uniformly distributed, uncorrelated with the input signal, and has a white spectrum. This paper surveys the theory behind this model, and discusses the conditions of its validity. The application of the model to floating-point quantization is demonstrated.**

I. Introduction

SIGNALS are represented in digital computers by sequences of finite bit-length numbers. Mapping of a continuous-time, continuous-amplitude signal to a sequence of computer numbers requires discretization both in time and amplitude: *sampling* and *quantization* must be performed (see Fig. 1).

Sampling theory has been elaborately described in the literature [1]–[3], and is well understood. Sampling is a linear operation; therefore linear system theory can be applied to the analysis of it. The sampled signal can be obtained from the continuous-time signal by multiplying it (*modulation*) with a certain impulse carrier [Fig. 2(a)] [4].

The spectrum of the original signal is repeated along the frequency axis [Fig. 2(c)]. When the repeated spectra do not overlap, the original spectrum [Fig. 2(b)] can be restored, and its inverse Fourier transform yields the original input signal. In other words, the sampled signal contains the same information as the continuous-time signal, and it can be used for the same purposes as the original signal. This statement in a precise mathematical form is the *sampling theorem.*

II. Quantization as Sampling of the Probability Density Function (PDF)

Quantization is generally less well understood than sampling. The reason is that it is a nonlinear operation; therefore most people believe that standard tools of linear system theory cannot be applied to it. In fact, we will show how linear system theory can be precisely used to analyze the effect of quantization on moments and other statistical properties of the signals.

Sampling discretizes time, and quantization discretizes amplitude. One would expect that quantization has a similar effect on functions of the amplitude as sampling has on functions of time. This recognition led Widrow to the study of probability density functions (PDF's) and to the development of a statistical theory of quantization in the late 1950's [5]–[7].

The characteristics of a uniform quantizer are pictured in Fig. 3(a), and a symbolic representation of quantization as an operator is shown in Fig. 3(b). The quantizer input is x, and the quantizer output is x'. Quantization is an operation on signals that is represented as a "staircase" function, a nonlinear relation between x' and x.

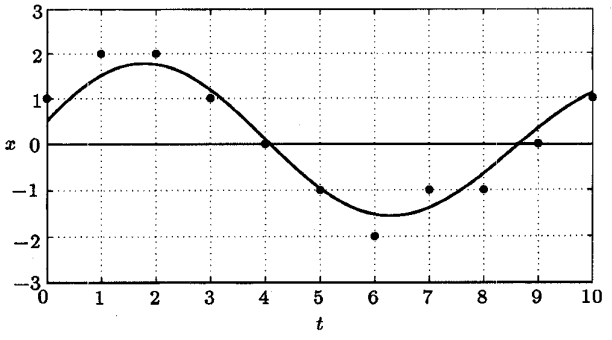


Fig. 1. Sampling and quantization.

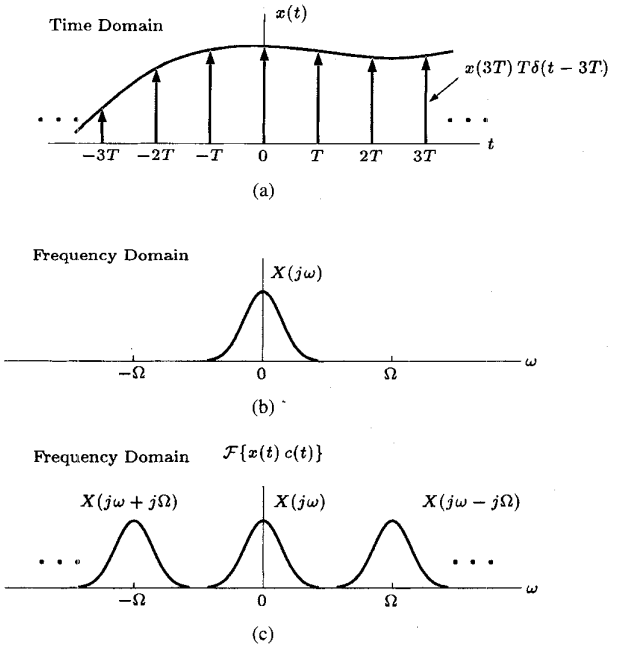


Fig. 2. The Fourier transform of a time function, and the Fourier transform of its samples: (a) a time function being sampled; (b) symbolic representation of Fourier transform of time function; and (c) symbolic representation of Fourier transform of samples of time function.

Manuscript received April 24, 1995; revised October 30, 1995. This work was sponsored by the National Science Foundation under Grant NSF IRI-9113491-A1, the Electric Power Research Institute under Grant 2DPM901, the Fulbright Program, and the U.S.–Hungarian Science and Technology Joint Fund in cooperation with the Hungarian Academy of Sciences and the National Institute of Standards and Technology under Project 299.

The authors are with the Department of Electrical Engineering, Information Systems Laboratory, Stanford University, Stanford, CA 94305-4055 USA.

Publisher Item Identifier S 0018-9456(96)02495-3.

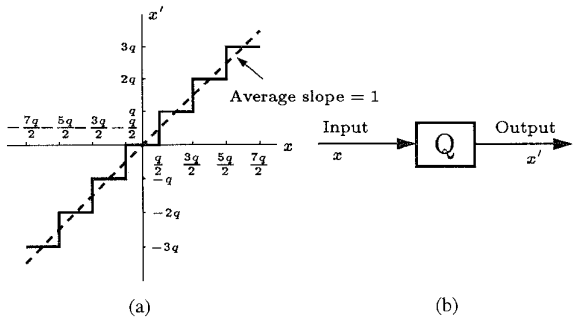


Fig. 3. A basic quantizer: (a) input-output characteristic and (b) block-diagram symbol of the quantizer.

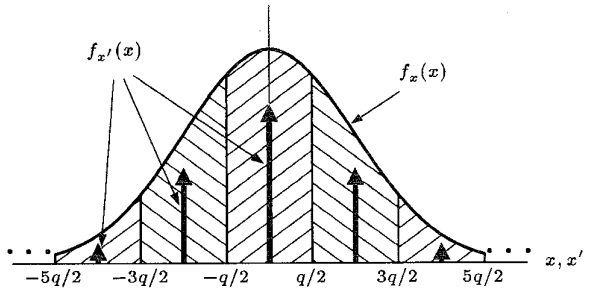


Fig. 4. Formation of the PDF of the quantizer output x': area sampling.

Fig. 4 is a sketch of a typical PDF of the quantizer input and output. The input PDF $f_x(x)$ is smooth, and the output PDF $f_{x'}(x)$ is discrete. The reason for this is that each input value is rounded toward the nearest allowable discrete level. The probability of each discrete output level equals the probability of the input signal occurring within the associated quantum band. For example, the probability that the output signal has the value zero equals the probability that the input signal falls between $\pm q/2$, where q is the quantization box size.

Careful study of quantization reveals that the PDF's of the input and output signals are related to each other through a special type of sampling. The output PDF is a string of Dirac delta functions whose areas correspond to the areas under the input PDF within the bounds of each quantum box. Cutting up the input PDF into strips as in Fig. 4, the area of each strip is compressed into an impulse in the center of the strip when forming the output PDF. This is like a sampling process, and we call it *area sampling*.

Area sampling can be accomplished by first convolving the input PDF $f_x(x)$ with a uniform pulse

$$f_n(x) = \begin{cases} \dfrac{1}{q}, & -\dfrac{q}{2} < x < \dfrac{q}{2} \\ 0 & \text{elsewhere,} \end{cases} \tag{1}$$

then following this with conventional sampling. A sketch of the input PDF is shown in Fig. 5(a). The function $f_n(x)$ is shown in Fig. 5(b). The convolution of $f_n(x)$ and $f_x(x)$ is illustrated in Fig. 5(c). To do conventional sampling, a uniform impulse train is represented in Fig. 5(d). Multiplying this train by the convolution gives the impulse train of Fig. 5(e). This

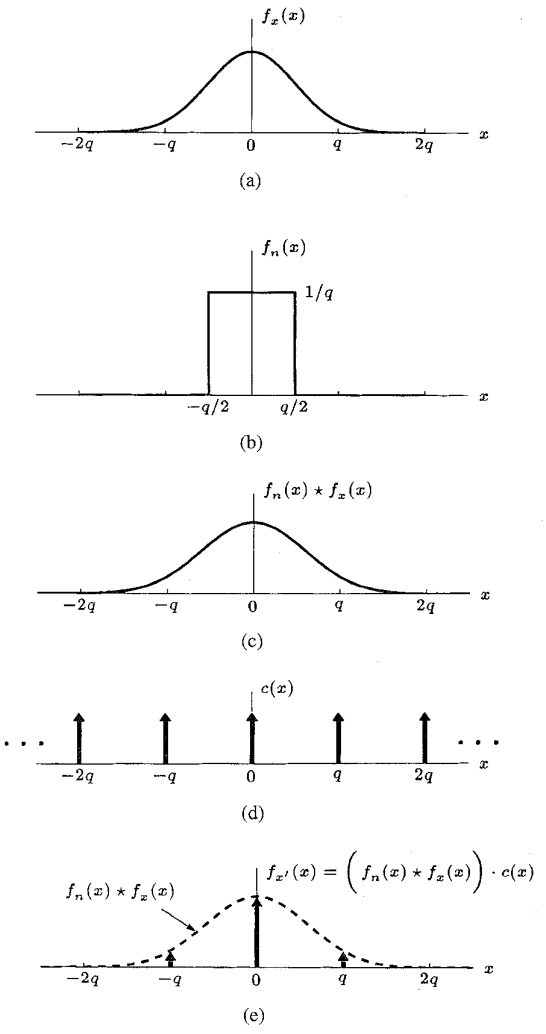


Fig. 5. Derivation of PDF of x' from area sampling of the PDF of x: (a) PDF of x; (b) rectangular pulse function; (c) convolution of (a) and (b); (d) the impulse train; and (e) PDF of x', the product of (c) and (d).

final impulse train is the PDF of the quantizer output, $f_{x'}(x)$. Every step of the way in going from $f_x(x)$ to $f_{x'}(x)$ involves linear operations.

The Fourier transform of the PDF is known in statistics as the characteristic function, the CF. The input CF is

$$\begin{aligned} \Phi_x(u) &= \int_{-\infty}^{\infty} f_x(x) e^{jux}\, dx \\ &= E\{e^{jux}\}. \end{aligned} \tag{2}$$

The CF is as useful in quantization theory as the Fourier transform of signals is in sampling theory. The input CF is sketched in Fig. 6(a), corresponding to the input PDF of Fig. 5(a). The Fourier transform of the rectangular pulse, a sinc function, is shown in Fig. 6(b), and this corresponds to the pulse of Fig. 5(b). The product of the Fourier transforms is shown in Fig. 6(c), corresponding to the convolution of Fig. 5(c). The product is repeated in the transform domain with a "frequency" of Ψ given by $\Psi \triangleq 2\pi/q$. The repetition

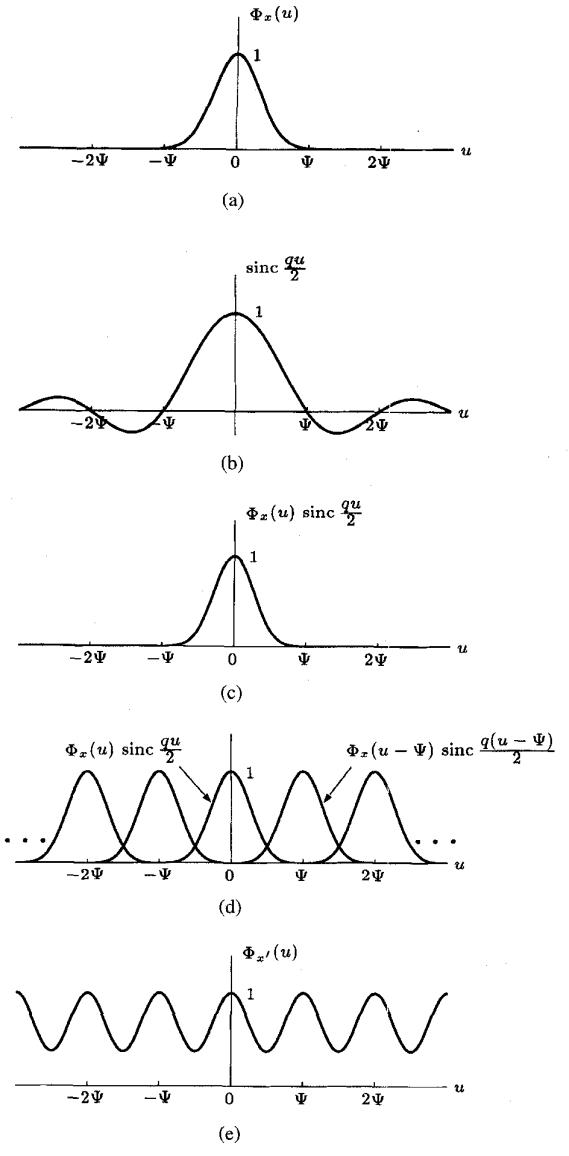


Fig. 6. Formulation of area sampling in the CF domain: (a) CF of x; (b) CF of n, the sinc function; (c) CF of $x + n$; (d) the repetition of (c); and (e) CF of x'.

is shown in Fig. 6(d), and the sum of the repetitions is shown in Fig. 6(e). This is a sketch of the Fourier transform of the output PDF of Fig. 5(e). A general expression for the CF of the quantizer output is

$$\Phi_{x'}(u) = \sum_{l=-\infty}^{\infty} \Phi_x(u + l\Psi) \operatorname{sinc}\left[\frac{q \cdot (u + l\Psi)}{2}\right] \tag{3}$$

where $\operatorname{sinc}(u) = \sin(u)/u$.

Equation (3) clearly shows the repetition at integer multiples of Ψ, the quantization frequency. This is analogous to the sampling radian frequency, $\Omega = 2\pi/T$, where T is the sampling period. The sampling period is analogous to the quantization box size q.

In terms of the bandlimitedness of the CF, a *Quantizing Theorem* can be formulated, as follows.

Quantizing Theorem I (QT I): If the CF of x is "band-limited," so that

$$\Phi_x(u) = 0 \quad \text{for} \quad |u| > \frac{\pi}{q} = \frac{\Psi}{2} \tag{4}$$

then

- the CF of x can be derived from the CF of x', and
- the PDF of x can be derived from the PDF of x'.

The proof is straightforward from (3). The quantizing theorem provides the condition for the output PDF of the quantized signal to contain all the information about the input PDF. In other words, we established a one-to-one connection between the statistical descriptions of the input and output signals of the quantizer.

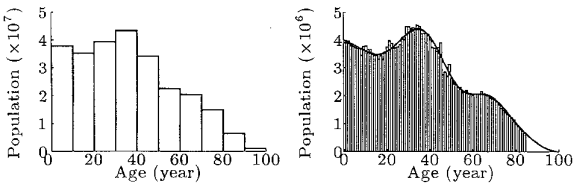
The above considerations lead to another very important consequence. By taking the central replica of the CF, or equivalently, by interpolating the output PDF, we obtain not the input PDF, but its convolution with a uniform PDF. Therefore, the central replica of the CF is the product of the CF of x and the CF of a uniform distribution.

The analogy between sampling and quantization is even more profound. When a signal is not band-limited, we usually apply an anti-aliasing filter to it before sampling. The anti-aliasing filter multiplies the spectrum by the transfer function of the filter, which is zero outside the desired passband. Similarly, in quantization we can find a way to multiply the CF by a desired function. A product of characteristic functions corresponds to convolution in the PDF domain, since CF's and PDF's are Fourier transform pairs. Convolution of PDF's corresponds to addition of independent random variables. Therefore, we can limit the band of the CF by adding an independent random variable with limited CF bandwidth to the input signal. This auxiliary signal is called *dither,* well-known in the practice of A/D conversion and digital signal processing [8]–[11]. This is a very important topic, but because of the limited space, we have to refer here to the literature for more detail.

III. Reconstruction of the Input PDF

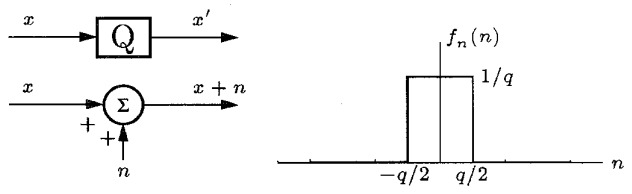
It follows from the model described above that as long as QT I is satisfied, the output and input PDF's are uniquely related to each other. Therefore, a crude histogram can be used for reconstruction of the input PDF. This is illustrated on age distribution of the 1992 US census data [12], see Fig. 7. The left-hand plots show histograms artificially made coarser than the usual one-year resolution; the right-hand side plots show the interpolated results superimposed on the bar graph of the original census data. It is striking how good the interpolation results are even with ten-year input resolution.[1]

[1]The distribution has a significant jump at zero, and this makes the CF wide, violating QT I. For the calculation of the reconstruction, we continued the histogram and the PDF by their mirror images in order to avoid these problems. Sinc function interpolation was performed, followed by deconvolution of the rectangular pulse.



(a) (b)

Fig. 7. Reconstruction of US age distribution from histograms based on 1992 census data: (a) ten-year histogram and (b) interpolation of the ten-year histogram, superimposed on the original one-year histogram data.



(a) (b)

Fig. 8. Comparison of quantization with addition of independent noise: (a) quantization and noise addition, and (b) PDF of the noise.

IV. Moments

The moments of a random variable x, such as the mean, mean square, mean cube, etc., can be determined by taking derivatives of the CF at the origin. The kth moment is

$$E\{x^k\} = \frac{1}{j^k} \left. \frac{d^k \Phi_x(u)}{du^k} \right|_{u=0}. \tag{5}$$

One can verify this by differentiating (2), making use of the definition

$$E\{x^k\} \triangleq \int_{-\infty}^{\infty} x^k f_x(x)\, dx. \tag{6}$$

V. The PDF of the Quantizer Output

It is well known that when two statistically independent signals are added together, the sum has a PDF which is the convolution of the PDF's of the two signals. Accordingly, the CF of the sum is the product of the CF's of the two signals. These facts are very important in the development of the statistical theory of quantization.

It is useful to compare quantization with the addition of uniform independent noise. Referring to Fig. 8(a), quantization of x yields x', and addition of independent noise n yields $x+n$. The PDF of the noise to be added is shown in Fig. 8(b). The PDF of x' is discrete, and the PDF of $(x+n)$ is smooth. The latter is equal to the convolution of x, i.e., $f_x(x)$, and the PDF of n, i.e., $f_n(x)$. By inspection of Fig. 5(e), one can deduce that the discrete PDF $f_{x'}(x)$ of the quantizer output is equal to the samples of the smooth PDF $f_{x+n}(x)$ of the sum of x and n. This is true in general, with or without satisfaction of the quantizing theorem.

VI. Moments of the Quantizer Output Signal

Referring once again to Fig. 8(a), it is useful to compare the moments of x' with those of $x+n$. In general, they do not correspond. But they do correspond exactly when certain quantizing theorems apply.

Refer now to Fig. 6. When QT I is satisfied, the replications of Fig. 6(d) do not overlap. When summing the replications, the derivatives at the origin in Fig. 6(e), related to the moments of x', will correspond exactly to the derivatives at the origin in Fig. 6(c), related to the moments of $x+n$. Thus, when QT I is satisfied,

$$E\{(x')^k\} = E\{(x+n)^k\}. \tag{7}$$

This relation will allow the moments of x to be calculated from the moments of x'.

Equation (7) will apply even when the replicas of Fig. 6(d) overlap, as long as the overlap does not impact on the derivatives at the origin. This leads to a second quantizing theorem that applies to moments.

Quantizing Theorem II (QT II): If the CF of x is band-limited so that

$$\Phi_x(u) = 0 \quad \text{for} \quad |u| > \frac{2\pi}{q} - \varepsilon = \Psi - \varepsilon, \tag{8}$$

with ε positive and arbitrarily small, then the moments of x can be calculated from the moments of x'.

QT I and QT II were first proved by Widrow [5]. He has also shown that if QT I or QT II holds, the moments of the quantized variable are equal to the moments of the sum of the input variable *and a uniformly distributed noise.* This noise has a mean of zero, a mean square of $q^2/12$, a mean cube of zero, a mean fourth of $q^4/80$, etc. A rearrangement of these relations yields Sheppard's famous corrections [13], [14], originally developed for grouped data under some smoothness conditions on the PDF. The most right-hand terms (in parentheses) are the Sheppard corrections

$$\begin{aligned}
E\{x\} &= E\{x'\} - (0) \\
E\{x^2\} &= E\{(x')^2\} - (\tfrac{1}{12}\, q^2) \\
E\{x^3\} &= E\{(x')^3\} - (\tfrac{1}{4}\, E\{x'\} q^2) \\
E\{x^4\} &= E\{(x')^4\} - (\tfrac{1}{2}\, q^2 E\{(x')^2\} - \tfrac{7}{240}\, q^4) \\
&\vdots
\end{aligned} \tag{9}$$

Sheppard's corrections allow one to recover the moments of x from the moments of x'.

The forms of QT I and also QT II resemble the sampling theorem. The similarity extends even further. Signals are usually not perfectly band-limited. None of the random variables which occur in practice have a perfectly band-limited CF, either. However, most of them are *approximately* band-limited, and a fine enough quantum size (large enough Ψ) can assure acceptable fulfillment of the conditions (4) or (8), by allowing the CF to be wide.

For Gaussian inputs, with standard deviation σ, a simple rule of thumb is that when $\sigma > q$, the conditions are fulfilled to good approximation. Let's consider the case $q = \sigma$, for example. The residual error of the second moment after Sheppard's correction is only $1.1 \times 10^{-8}\sigma^2$. For other

distributions the CF usually vanishes much more slowly than for the Gaussian CF; therefore q must be much smaller than the standard deviation for Sheppard's corrections to hold approximately.

The precise expressions for the residual errors in Sheppard's corrections are rather complex. Simple upper bounds for these errors can be obtained as follows. For many distributions, the envelope of the CF vanishes for large values of u as $\mathcal{O}[1/(Au)^p]$, where $p \geq 1$ depends on the probability distribution. Using this upper bound, it can be shown that the effect of the overlap at the origin in the characteristic function domain [see Fig. 6(e)], causing residual errors in Sheppard's corrections, can be approximated by a similar negative power function of the amplitude A, and this yields a simple expression of a minimum A (or a maximum q when A is fixed) [15].

VII. Statistics of Quantization Noise

We define the quantization noise ν to be the difference between the output and the input of the quantizer. Accordingly,

$$\nu \triangleq x' - x. \tag{10}$$

We would like to know about the statistical properties of the quantization noise. We do know that it is bounded between $\pm q/2$.

The PDF of the quantization noise $f_\nu(x)$ can be computed in the manner illustrated in Fig. 9. A given value of ν results from quantization of x falling at just the right places within all of the quantization boxes. The probability of getting a given value of ν is the sum of probabilities from all of the quantization boxes. The PDF of ν may therefore be constructed by cutting the PDF of x into strips, and stacking and adding them. It has been shown [6], [7] that the PDF of the quantization noise will be exactly uniform if either QT I or QT II is satisfied. As such, quantization noise has zero mean and a mean square of $q^2/12$.

The necessary and sufficient condition for the quantization noise to be uniform was developed by Sripad and Snyder [16]. The condition is satisfied when the CF is equal to zero at $2\pi l/q$, $l = \pm 1, \pm 2, \cdots$. This is a condition milder than QT II.

VIII. Cross-Correlation Between Quantization Noise and the Quantizer Input

Fig. 10 shows how one could measure quantization noise ν, defined by (10). It is of great interest to know the cross-correlation between the quantization noise and the quantizer input, to learn something about their relationship. It is clear, first of all, that the noise and the input are deterministically related. For a given input, there is a definite output and a definite difference between output and input. Although the quantization noise and the quantizer input are deterministically related, it is a curious fact that under certain circumstances, the input and noise are uncorrelated. It had been shown by Widrow [6], [7] that when either QT I or QT II is satisfied, quantization noise is uncorrelated with the signal being quantized. These conditions are met with Gaussian inputs to a very close approximation even when the quantization step size is as large as $q = \sigma$. Similar results are obtained with finer quantization for other input PDF's.

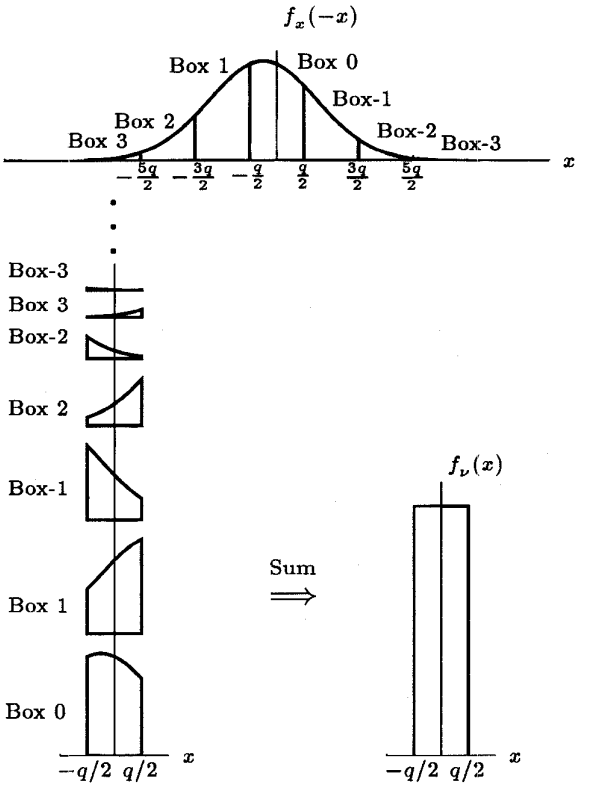


Fig. 9. Construction of the PDF of quantization noise.

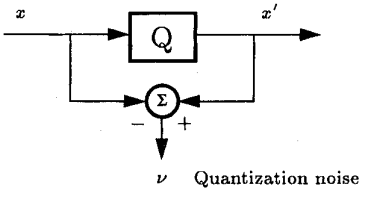


Fig. 10. Quantization noise, the difference between the quantizer output and its input.

IX. Pseudo Quantization Noise: The PQN Model

Refer once again to Fig. 8(a), where quantization is contrasted with the addition of independent uniformly distributed noise. The addition of independent noise and quantization are not the same, since the quantizer adds noise that is deterministically related to the signal being quantized. But when the conditions for QT I or QT II are met, all moments and joint moments correspond exactly for quantization and the addition of independent uniformly distributed noise. For example, when QT I or QT II is satisfied,

$$E\{(x')^k\} = E\{(x+n)^k\} \tag{11}$$

$$E\{\nu^k\} = E\{n^k\} \tag{12}$$

$$E\{x^k\nu^l\} = E\{x^k n^l\} = E\{x^k\}E\{n^l\} \tag{13}$$

for all positive integers k and l.

If the quantizer is embedded in a system with feedforward or feedback parts or both, the quantizer may be replaced for purposes of moment calculation when QT I or QT II is satisfied by a source of additive independent uniformly distributed noise. This noise is called pseudo quantization noise (PQN), and the additive noise model replacing quantization is called the PQN model.

X. High-Order Statistical Description of Quantization

Refer now to Fig. 1. To describe the statistics of the multiple-sample quantizer input, a multidimensional joint PDF would be required. These are high-order statistical descriptions.

High-order forms of QT I and QT II exist. They can be stated as follows.

Multidimensional Quantizing Theorem I (QT I): If the CF of a sequence of quantizer input samples is "band-limited" in N-dimensions, so that

$$\Phi_{x_1,\cdots,x_N}(u_1,\cdots,u_N)=0 \quad \text{when} \quad |u_k|>\frac{\pi}{q}=\frac{\Psi}{2} \quad \text{for any} \quad k\in[1,N] \tag{14}$$

then

- the CF of $x_1,\cdots,x_N$ can be derived from the CF of $x'_1,\cdots,x'_N$, and
- the PDF of $x_1,\cdots,x_N$ can be derived from the PDF of $x'_1,\cdots,x'_N$.

Multidimensional Quantizing Theorem II (QT II): If the CF of $x_1,\cdots,x_N$ is band-limited in N-dimensions, so that

$$\Phi_{x_1,\cdots,x_N}(u_1,\cdots,u_N)=0 \quad \text{when} \quad |u_k|>\frac{2\pi}{q}-\varepsilon=\Psi-\varepsilon \quad \text{for any} \quad k\in[1,N] \tag{15}$$

with ε positive and arbitrarily small, then the moments of $x_1,\cdots,x_N$ can be calculated from the moments of $x'_1,\cdots,x'_N$.

When conditions are met for either high-order QT I or QT II, a high-order PQN model applies, and the moments related to quantization will correspond exactly to those pertaining to the addition of independent noise, noise samples that are independent of the input signal and independent of each other over time. Thus, when either QT I or QT II is satisfied, the quantization noise among other things is uniformly distributed in multidimensions, white, and uncorrelated with the quantizer input. It has a mean of zero and a mean square of $q^2/12$.

The exact condition of whiteness was given by Sripad and Snyder [16] in terms of the joint CF of two input samples as[2]

$$\Phi_{x_1,x_2}\left(\frac{2\pi l_1}{q},\frac{2\pi l_2}{q}\right)=0 \tag{16}$$

[2]The condition in the original paper of Sripad and Snyder contains a typo; we give here the correct version.

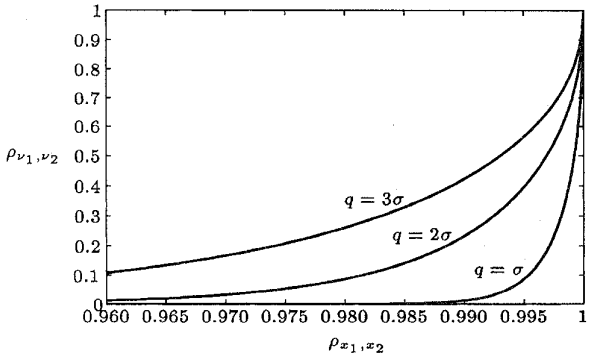


Fig. 11. Correlation coefficient of the quantization noise as a function of the correlation coefficient of the quantizer input. The input is Gaussian.

for every integer value of l_1 and l_2, except $(l_1, l_2) = (0, 0)$. This condition is quite difficult to apply in practice; therefore alternative formulations are of great importance (see below).

In practice, input CF's are not exactly band-limited, and the quantizing theorems apply only approximately. High-order CF's have some overlap with their repetitive parts, and this impacts their moments. If the quantizer input is Gaussian with q as big as σ, correlations among input samples as high as 99% will cause correlations among corresponding quantization noise samples of only 1%. The quantization noise will be essentially white, having a flat spectrum and an impulsive autocorrelation function almost without regard to the autocorrelation function of the quantizer input. Fig. 11 shows theoretical plots of correlation coefficients of quantization noise samples versus correlation coefficients of corresponding quantizer input samples. Similar curves were derived by Widrow in 1956 [6].

An approximate condition of whiteness was developed in [17] as

$$\Omega<2\pi K\frac{E\{|\dot{x}(t)|\}}{q} \tag{17}$$

where the constant K may be equal to $1\cdots 2$, depending on the input PDF; a uniformly applicable value is $K=1$. Condition (17) is easy to check by measurement.

XI. Application of Uniform Quantization Theory to Analysis of Floating-Point Quantization

Scientific calculations are almost exclusively done using floating-point number representation, and also more and more digital signal processors contain floating-point arithmetic. Therefore, it is of high importance to develop models that account for floating-point roundoff effects.

Floating-point quantization was extensively discussed in the literature [18]–[24]. Generally the properties of the *relative error* are investigated: the quantization error of the input x is approximated by

$$\nu_{FL}\approx x\cdot\epsilon, \tag{18}$$

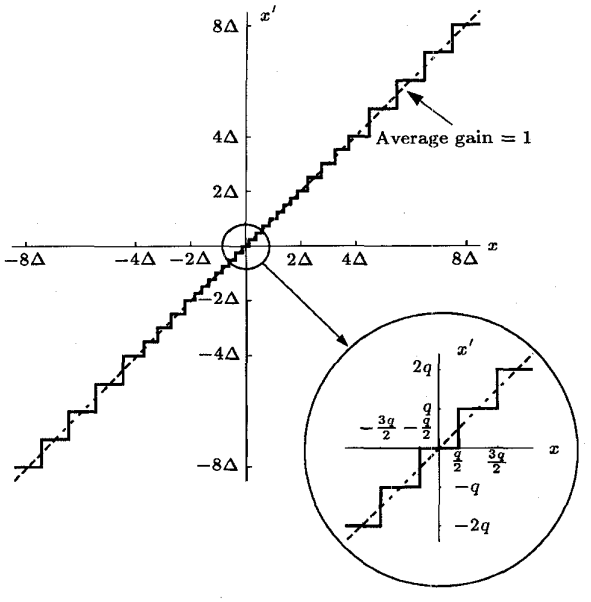


Fig. 12. Input-output staircase function for a floating-point quantizer with a 3-bit mantissa, i.e., $p = 3$.

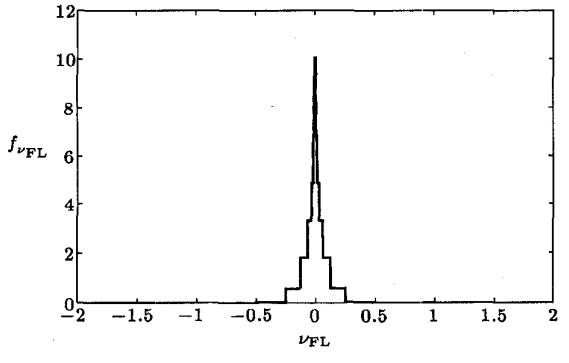


Fig. 13. The PDF of floating-point quantization noise with a zero-mean Gaussian input, $\sigma = 64\Delta$, and with a 2-bit mantissa.

where ϵ is independent of x, has a special (trapezoid-like) distribution, and its width is determined by the bit length of the mantissa. The results of uniform quantization are usually not applied to floating-point quantization, because of the nonuniform overall characteristic of the latter. In this section we will demonstrate that application of the statistical theory of quantization to floating-point quantizers is a very fruitful approach. There is not enough space here to go into all details; instead, we will highlight the most important ideas.

A floating-point quantizer characteristic is illustrated in Fig. 12. This is clearly a nonuniform quantizer. Its quantization noise, $\nu_{FL} = x' - x$, has a strange type of PDF that we call a "skyscraper distribution" (Fig. 13).

However, it is possible to represent the floating-point quantizer by the combination of a piecewise linear compressor, a uniform quantizer (the "hidden quantizer"), and a piecewise linear expander (Figs. 14 and 15), like in the compandors for speech coding. The expander is the inverse of the compressor.

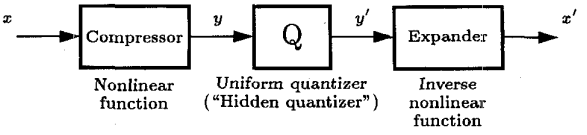


Fig. 14. A model of a floating-point quantizer.

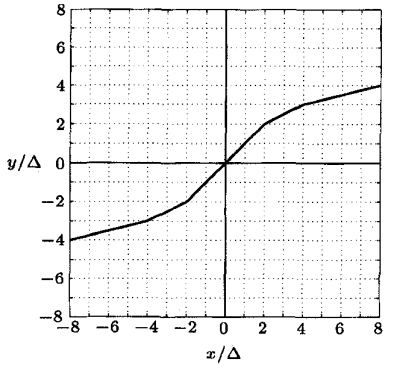


Fig. 15. The input-output characteristic of the compressor.

If we can determine the properties of the compressed signal, uniform quantization theory can be applied to the uniform quantizer, and, after application of the expander, we have a working model.

The PDF of the compressed signal contains huge jumps at the breakpoints of the compressor [Fig. 16(a)]. This seems to prevent the use of the quantizing theorems. However, it turns out that for bit numbers used in practice, especially for the IEEE single- and double-precision standards [25], the conditions of the quantization theorems are fulfilled very closely. This is illustrated for much coarser quantization, with $p = 4$ and $p = 8$ in Fig. 16(b) and (c), where p is the number of bits of the mantissa. Then, the hidden quantizer can be replaced by its PQN model. The expander, which is the exact inverse of the compressor (Fig. 15), is approximately an exponential function. Therefore, with reasonable approximation, which can be analyzed more closely, the floating-point noise can be expressed as

$$\nu_{FL} \approx \nu \cdot |x| \cdot \left(\frac{\ln 2}{\Delta}\right) \tag{19}$$

where Δ is the width of the linear sections of the expander. This approximate model supports the idea of representing the floating-point quantization noise by means of a relative error as in (18).

Systematic use of the PQN model leads to simple and useful results. Some highlights are the following:

- ν_{FL} is "skyscraper-distributed" (see Fig. 13)
- $E\{\nu_{FL}\} = 0$
- for most distributions, $E\{\nu_{FL}^2\} \approx 0.180 \times 2^{-2p} E\{x^2\}$
- $\text{cov}\{x, \nu_{FL}\} = 0$
- ν_{FL} is white for all practical cases.

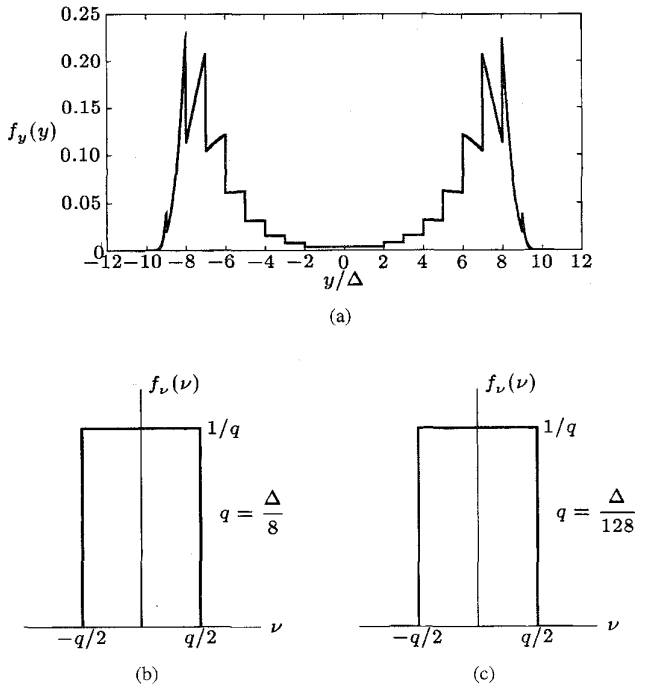


Fig. 16. PDF of compressor output and of hidden quantization noise when x is zero-mean Gaussian with $\sigma = 100\Delta$: (a) $f_y(y)$; (b) $f_\nu(\nu)$ for $p = 4$ $(q = \Delta/8)$; and (c) $f_\nu(\nu)$ for $p = 8$ $(q = \Delta/128)$.

These and further results can be applied to scientific computations and floating-point signal processing, as floating-point digital filters, floating-point FFT, and so on. Derivations, proofs, and applications will be given in a forthcoming Prentice-Hall book entitled "Quantization Noise," by Widrow and Kollár [26]. Other useful and related references are [27]–[35].

XII. Conclusions

A brief survey of the statistical theory of quantization was presented. The most important results were summarized, and application of the theory to floating-point quantization was presented. This theory is a very powerful tool to analyze statistical properties of quantized variables and of estimators calculated from them.

References

[1] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE,* vol. 47, pp. 10–21, 1949.

[2] A. J. Jerri, "The Shannon sampling theorem—Its various extensions and applications: A tutorial review," *Proc. IEEE,* vol. 65, no. 11, pp. 1565–1596, Nov. 1977.

[3] R. J. Marks, II, *Introduction to Shannon Sampling and Interpolation Theory.* New York: Springer-Verlag, 1991

[4] W. K. Linvill, "Sampled-data control systems studied through comparison of sampling with amplitude modulation," *AIEE Trans.,* vol. 70, pp. 1779–1788, 1951.

[5] B. Widrow, " A study of rough amplitude quantization by means of Nyquist sampling theory," Sc.D. thesis, Department of Electrical Engineering, MIT, June 1956.

[6] ——, "A study of rough amplitude quantization by means of Nyquist sampling theory," *IRE Trans. Circuit Theory,* vol. 3, no. 4, pp. 266–276, Dec. 1956.

[7] ——, "Statistical analysis of amplitude-quantized sampled-data systems," *Trans. AIEE, Part II: Applicat. Ind.,* vol. 79, no. 52, pp. 555–568, Jan. 1961 Section.

[8] D. T. Sherwood, "Some theorems on quantization and an example using dither," in *Conf. Rec. 19th Asilomar Conf. Circuits, Systems and Computers,* Pacific Grove, CA, Nov. 6–8, 1986, 86CH2331-7, pp. 207–212.

[9] S. P. Lipshitz and R. A. Wannamaker, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.,* vol. 40, no. 5, pp. 355–375, May 1992.

[10] P. Carbone *et al.,* "Effect of additive dither on the resolution of ideal quantizers," *IEEE Trans. Instrum. Meas.,* vol. 43, no. 3, pp. 389–396, June 1994.

[11] R. M. Gray and T. G. Stockham, Jr., "Dithered quantizers," *IEEE Trans. Inform. Theory,* vol. 39, no. 3, pp. 805–812, May 1993.

[12] Statistical Abstract of the United States. *The National Data Book,* 114th ed. U.S. Dept. of Commerce, Economics and Statistics Administration, Bureau of the Census, 1994.

[13] W. F. Sheppard, "On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant divisions of scale," *Proc. London Math. Soc.,* vol. 29, pp. 353–380, 1898.

[14] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics, Vol. 1, Distribution Theory,* 6th ed. London: Edward Arnold–New York: Wiley, 1994.

[15] I. Kollár, "Bias of mean value and mean square value measurements based on quantized data," *IEEE Trans. Instrum. Meas.,* vol. 43, no. 5, pp. 733–739, Oct. 1994.

[16] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust, Speech, Signal Processing,* vol. ASSP-25, no. 5, pp. 442–448, Oct. 1977.

[17] I. Kollár, "The noise model of quantization," in *Proc. 1st IMEKO TC4 Symp. Noise in Electrical Measurements,* Como, Italy, June 19–21, 1986; Budapest, OMIKK-Technoinform, 1987, pp. 125–129.

[18] R. W. Hamming, "On the distribution of numbers," *Bell System Tech. J.,* vol. 49, no. 8, pp. 1609–1625, Oct. 1970.

[19] T. Kaneko and B. Liu, "On local roundoff errors in floating-point arithmetic," *J. Assoc. Comp. Mach.,* vol. 20, no. 3, pp. 391–398, July 1973.

[20] J. Kontro, K. Kalliojärvi, and Y. Neuvo, "Floating-point arithmetic in signal processing," in *Proc. IEEE Int. Symp. Circuits and Systems,* San Diego, CA, May 10–13, 1992, 92CH3139-3, vol. 4, pp. 1784–1791.

[21] A. Lacroix and F. Hartwig, "Distribution densities of the mantissa and exponent of floating-point numbers," in *IEEE Int. Symp. Circuits and Systems,* San Diego, CA, May 1992, pp. 1792–1795.

[22] B. Liu and T. Kaneko, "Error analysis of digital filters realized with floating-point arithmetic," *Proc. IEEE,* vol. 57, no. 10, pp. 1735–1747, Oct. 1969.

[23] A. V. Oppenheim and C. J. Weinstein, "Effects of finite register length in digital filtering and the fast fourier transform," *Proc. IEEE,* vol. 60, no. 8, Aug. 1972, pp. 957–976.

[24] A. B. Sripad and D. L. Snyder, "Quantization errors in floating-point arithmetic," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-6, no. 5, pp. 456–463, Oct. 1978.

[25] The Institute of Electrical and Electronics Engineers, "IEEE Standard for binary floating-point rrithmetic," ANSI/IEEE Standard 754-1985, New York, Aug. 1985; "IEEE Standard for radix-independent floating-point arithmetic," ANSI/IEEE Standard 854-1987, New York, Oct. 1987.

[26] B. Widrow and I. Kollár, *Quantization Noise.* Englewood Cliffs, NJ: Prentice-Hall, in preparation.

[27] D. Bellan, A. Brandolini, and A. Gandelli, "Quantization theory in electrical and electronic measurements," in *Instrumentation and Measurement Tech. Conf., IMTC'95,* Waltham, MA, Apr. 24–26, 1995.

[28] W. R. Bennett, "Spectra of quantized signals," *Bell System Tech. J.,* vol. 27, no. 3, pp. 446–472, July 1948.

[29] T. A. C. M. Claasen and A. Jongepier, "Model for the power spectral density of quantization noise," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. ASSP-29, no. 4, pp. 914–917, 1981.

[30] T. Dobrowiecki, "Quantized error spectra at high frequencies for a certain class of signals," in *Proc. 2nd IMEKO TC7 Symp. Application of Statistical Methods in Measurement,* Leningrad, May 16–19, 1978, pp. Dobrowiecki/1–8.

[31] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Information Theory,* vol. 36, no. 6, pp. 1220–1244, Nov. 1990.

[32] J. Katzenelson, "On errors introduced by combined sampling and quantization," *IRE Trans. Automat. Contr.,* vol. AC-7, pp. 58–68, 1962.

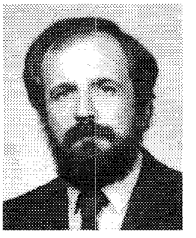
[33] I. Kollár, "Statistical theory of quantization: Results and limits," *Periodica Polytechnica Ser. Elect. Eng.,* vol. 28, no. 2/3, pp. 173–190, 1984.

[34] G. H. Robertson, "Computer study of quantizer output spectra," *Bell System Tech. J.,* vol. 48, no. 5, pp. 2391–2403, 1969.
[35] A. I. Velichkin, "Correlation function and spectral density of a quantized process," *Telecommunications and Radio Engineering, Part II: Radio Engineering,* pp. 70–77, July 1962.

Bernard Widrow (M'58–SM'75–F'76–LF'95) received the S.B., S.M., and Sc.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1951, 1953, and 1956, respectively.

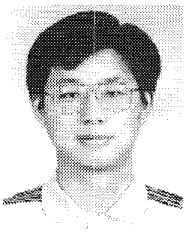
He was with MIT until he joined the faculty, Stanford University, Stanford, CA, in 1959, where he is now professor of electrical engineering. He is presently engaged in research and teaching in neural networks, pattern recognition, adaptive filtering, adaptive control systems, and quantization theory. He is associate editor of the journals *Adaptive Control and Signal Processing, Neural Networks, Information Sciences,* and *Pattern Recognition,* is coauthor with S. D. Stearns of *Adaptive Signal Processing* (Prentice-Hall), and coauthor with E. Walach of *Adaptive Inverse Control* (Prentice-Hall).

Dr. Widrow is a member of the National Academy of Engineering, the American Association of University Professors, the Pattern Recognition Society, Sigma Xi, and Tau Beta Pi. He is a Fellow of the American Association for the Advancement of Science, and is past president of the International Neural Network Society. Professor Widrow received the IEEE Centennial Medal in 1984, and the IEEE Neural Networks Pioneer Medal in 1991. In 1986, he received the IEEE Alexander Graham Bell Medal for exceptional contributions to the advancement of telecommunications.



István Kollár (M'87–SM'93) was born in Budapest, Hungary in 1954. He graduated in electrical engineering from the Technical University of Budapest in 1977, and in 1985 received the degree "Candidate of Sciences" (the equivalent of Ph.D.) from the Hungarian Academy of Sciences, and the degree dr.tech. from the Technical University of Budapest.

From September 1993 to June 1995, he was a Fulbright scholar and visiting associate professor in the Department of Electrical Engineering, Stanford University. He is associate professor of electrical engineering, Department of Measurement and Instrument Engineering, Technical University of Budapest. His research interests span the areas of digital and analog signal processing, measurement theory, and system identification. He has published about 50 scientific papers, and is coauthor of the book *Technology of Electrical Measurements,* (L. Schnell, Ed., Wiley, 1993). He authored the *Frequency Domain System Identification Toolbox* for Matlab.



Ming-Chang Liu was born in Taipei, Taiwan in 1962. He received the B.S. degree from the National Taiwan University in 1984, and M.S. degree of mechanical engineering from Stanford University in 1989. He is currently a Ph.D. candidate in electrical engineering at Stanford University. His research interests are in the areas of statistical signal processing, quantization noise, and adaptive control.