Abstract—Quantization is the process of replacing analog samples with approximate values taken from a finite set of allowed values. The approximate values corresponding to a sequence of analog samples can then be specified by a digital signal for transmission, storage, or other digital processing. In this expository paper, the basic ideas of uniform quantization, companding, robustness to input power level, and optimal quantization are reviewed and explained. The performance of various schemes is compared using the ratio of signal power to mean-square quantizing noise as a criterion. Entropy coding and the ultimate theoretical bound on block quantizer performance are also compared with the simpler zero-memory quantizer.

QUANTIZATION

ALLEN GERSHO, MEMBER, IEEE

I. INTRODUCTION

The processing and transmission of digital signals is rapidly approaching a dominant role in communication systems. Nevertheless, the physical origin of many information-bearing signals (speech, image, telemetry, seismic, etc.) is intrinsically analog and continuous-time in nature. Therefore, an effective interface between the analog and digital worlds is of crucial importance in modern signal processing. Very often the quality of analog-to-digital (A/D) conversion is the critical limiting factor in overall system performance. A clear understanding of quantization, the essential mechanism of A/D conversion, is needed to answer such questions as how many bits per second (or bits per sample) are really needed, or how much distortion (or quantizing noise) is inevitable for a given bit rate.

Analog-to-digital conversion may be viewed as being made up of four operations: prefiltering, sampling, quantizing, and coding. In this paper we focus on quantization, and specifically on "zero-memory" quantization.

Quantization begins with the availability of analog samples. Each sample may in general take on any of a continuum of amplitude values ranging from $-\infty$ to $+\infty$. The quantizer replaces each of these sample values with an output value which is an approximation to the original amplitude. The key feature is that each output value is one of a *finite* set of real numbers. Hence a symbol from a finite alphabet can be used to represent and identify the particular output value that occurs. A distinct n-bit binary word can be associated with each output value if the set of output values contains no more than 2^n members. With this procedure a sequence of analog samples can be transformed into a sequence of binary words suitable for storage, transmission, or some other form of *digital* signal processing. A receiver having the table of output values (sometimes called "quanta" or "quantum levels") associated with the set of binary words can then reconstruct an approximation to the original sequence of samples. Hence with some appropriate form of interpolation, a continuous waveform can be created which approximates the waveform originally applied to the A/D system. The reconstruction process is called digital-to-analog (D/A) conversion.

The simplest and most common form of quantizer is the *zero-memory* quantizer. In this case the output value is determined by the quantizer only from one corresponding input sample, independent of the values taken on by earlier (or later) analog samples applied to the quantizer input. More sophisticated (but less well understood theoretically) is the *block* quantizer which looks at a group or "block" of input samples simultaneously and produces a block of output values, chosen from a finite set of possible output blocks, approximating the corresponding input samples. In general, for a given number of bits per sample representing the output values, a better quality approximation can be achieved by block quantization. Of theoretical interest is the limiting case where the block length approaches infinity. Studying this limiting situation provides information about the ultimate quality of approximation achievable for a given bit rate. Another class of quantizers which could be described as *sequential quantizers* includes

This work was performed while the author was visiting the Department of System Science, University of California, Los Angeles, CA. He is with Bell Laboratories, Murray Hill, NJ 07974.

such well-known digitization schemes as delta modulation, differential PCM, and other adaptive versions. A sequential quantizer stores some information about the previous samples and generates the present quantized output using both the current input *and* the stored information. In this paper we shall focus primarily on zero-memory quantization. Quantization with memory will be discussed only for the purpose of examining how much can be gained through the use of memory.

II. ZERO-MEMORY QUANTIZATION

A zero-memory N-point quantizer Q may be defined by specifying a set of $N+1$ *decision levels* $x_0, x_1, \ldots, x_N$ and a set of N *output points* $y_1, y_2, \cdots, y_N$. When the value X of an input sample lies in the ith quantizing interval, namely,

$$R_i = \{x_{i-1} < X < x_i\},$$

the quantizer produces the output value y_i. Since y_i is used to approximate samples contained in the interval R_i, y_i is itself chosen to be some value in the interval R_i. The end levels x_0 and x_N are chosen equal to the smallest and largest values, respectively, that the input samples may have. Usually, the sample values are unbounded, which we henceforth assume, so that $x_0 = -\infty$ and $x_N = +\infty$. The N output points always have finite values. If $N = 2^n$, a unique n-bit binary word can be associated with each output point, yielding an "n-bit quantizer."

The input-output characteristic $Q(x)$ of a quantizer has a staircase form. The midtread characteristic shown in Fig. 1 produces zero output for input samples that are in the neighborhood of zero; the midriser characteristic shown in Fig. 2 has a decision level located at zero. A quantizer is simply a memoryless nonlinearity whose characteristic may be viewed as a staircase approximation to the "identity" operation $y = x$.

When the input sample is located in the end regions R_1 or R_N the quantizer is said to be *overloaded.* All other quantizing intervals R_i are finite in size.

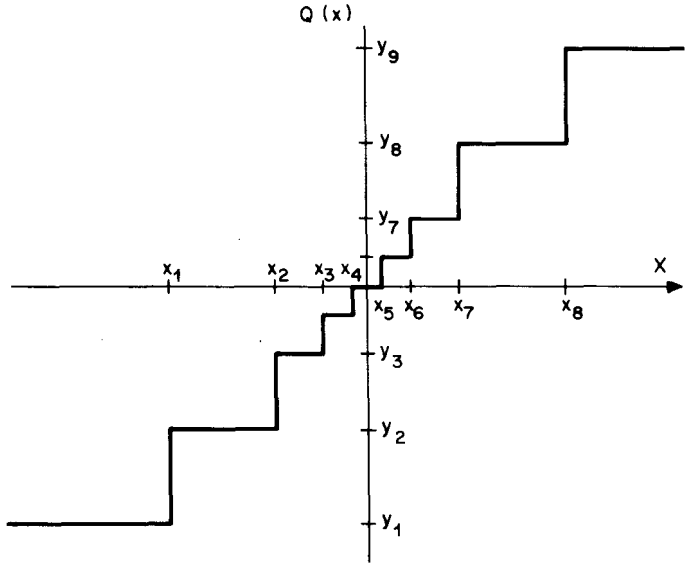


Fig. 1. Input-output characteristic of a midtread quantizer with N=9.

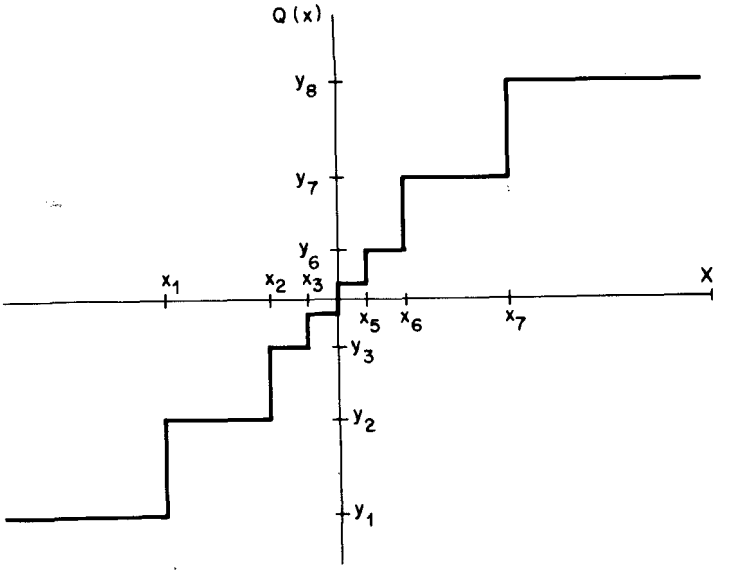


Fig. 2. Input-output characteristic of a midriser quantizer with N=8.

Fundamental to an analytical study of quantization is the recognition that the input samples must be regarded as random in character. The input samples are not known in advance and thus can be regarded as information-bearing. Quantization is actually a mechanism whereby information is thrown away, keeping only as much as is really needed to allow reconstruction of the original signal to within a desired accuracy as measured by some *fidelity criterion.* We define $p(x)$ as the first-order probability density function (hereafter pdf) of each input sample to the quantizer. Assume for convenience that the mean value of the input samples is zero and that $p(x)$ has even symmetry about zero. The zero mean assumption implies that any dc bias has been removed. The symmetry assumption is satisfied by most common density functions including the Gaussian (normal) density. With the symmetry assumption, the quantizer characteristic $Q(x)$ is normally chosen to have odd symmetry.

The quantization process can be modeled as the addition of a random noise component $e = Q(x) - x$ to the input sample, as indicated in Fig. 3. Unlike the usual signal-plus-noise models in communication theory, here the noise is actually dependent on the signal amplitude. The quantization noise may be regarded as the response when the input sample is applied to the nonlinear characteristic

$$U(x) = Q(x) - x$$

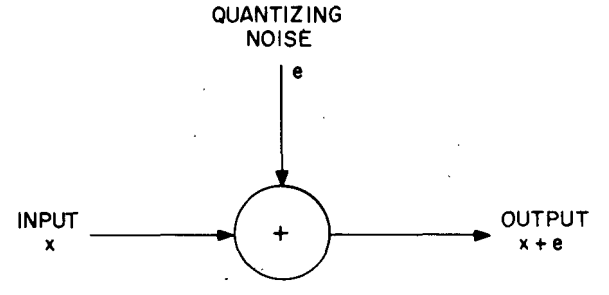


Fig. 3. Additive noise model of quantization. The quantizing noise e is often approximated as being independent of the input samples when the number of levels is large.

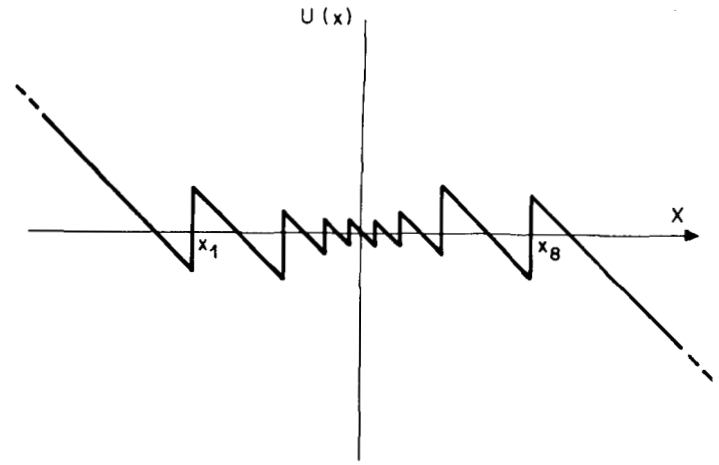


Fig. 4. Quantizing error as a function of input sample value for the quantizer of Fig. 1.

shown in Fig. 4. When the input sample lies within the interval $x_1 < x < x_{N-1}$, the output noise is described as *granular* noise and is bounded in magnitude. When the input lies outside this interval, the output is described as *overload* noise and the amplitude is unbounded. It is often convenient to artificially model quantization noise as the sum of granularity and overload noise as if they were two separate noise sources.

An effectively designed quantizer should be "matched" to the particular input probability density function, to the extent that this density is known to the designer. In particular, for a fixed number N of levels, the choice of overload levels x_1 and x_{N-1} controls a tradeoff between the relative amounts of granularity and overload noise.

In modeling quantization error as an additive noise source as in Fig. 3, it is often convenient to treat the noise as having a flat spectral density and as being uncorrelated with the input samples. This idea was used by Widrow [1] for uniformly spaced quantization levels. More generally, it may be shown that the quantizing noise is approximately white (i.e., successive noise samples are uncorrelated) and uncorrelated with the input process if: 1) successive input samples are only moderately correlated, 2) the number of output points N is large, and 3) the output points are very close to the midpoints of the corresponding quantization intervals. For a more precise treatment of the spectrum of quantizing noise, see Bennett [2].

III. PERFORMANCE MEASURES

Since the quantization error is modeled as a random variable, a measure of the performance of a quantizer must be based on a statistical average of some function of the error. Most common is the mean-square distortion measure D, defined by the usual expectation of the square of $U(x)$ above:

$$D = \int_{-\infty}^{\infty} [Q(x)-x]^2 \, p(x)dx. \tag{1}$$

This quantity can be used to measure the degradation introduced by the quantizer for a fixed input pdf $p(x)$. Frequently, it is more useful to describe the quantizer's performance by the "signal-to-noise ratio," often defined as

$$\text{SNR} = 10 \log_{10} (\sigma^2/D) \tag{2}$$

where σ^2 is the variance of the input samples. Other error criteria have also been considered in the study of quantization, such as the expectation of the kth power of the error magnitude. Frequently, the performance measure adopted is a subjective evaluation, and psychological studies are used to determine preferred quantization schemes among a set of schemes considered. Another approach is to consider the quality of approximation of a segment of the reconstructed waveform to the original waveform. Mean-square distortion may be viewed as a special case of this approach where the performance measure is the expectation of the sum of the squared errors for all sampling instants of the waveform segment. However, this measure does not distinguish between different approximations having the same total squared error. For example, it might be subjectively preferable to have a very high squared error at one isolated sampling instant than to have moderately high squared errors at several adjacent sampling instants. Hence, a more sophisticated distortion measure might be more meaningful than the usual mean-square distortion criterion.

In most applications of quantization, the number of levels N is very large so that a sufficiently high SNR is obtained. A useful formula for mean-squared error can then be used. Equation (1) can be written in the form

$$D = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (y_i - x)^2 \, p(x)dx \tag{3}$$

by breaking up the region of integration into the separate intervals R_i and noting that $Q(x) = y_i$ when x is in R_i. For large N, each interval R_i can be made quite small (with the exception of the overload intervals R_1 and R_N which are unbounded). Then it is reasonable to approximate the probability density $p(x)$ as being constant within the interval R_i. On setting $p(x) \cong p(y_i)$ when x is in R_i and approximating $p(x) \cong 0$ for x in the overload regions, the integral for each term of the sum (3) is readily found, and we get

$$D = \frac{1}{12} \sum_{i=2}^{N-1} p(y_i) \, \Delta_i^3 \tag{4}$$

where $\Delta_i = x_i - x_{i-1}$, the length of interval R_i. This approximate formula is based on the assumption that, for N large, a sufficient number of quantizing levels are available for both the granularity and overload noise to be very small. Equation (4) implies that the overload points x_0 and x_N are chosen so that overload noise is negligible compared to granular noise. Equation (4) will be used later to derive an integral formula for distortion.

Of frequent interest is the special case of *uniform* quantization where the decision levels are equally spaced so that the intervals R_i are of constant length, i.e., $\Delta_i = \Delta$, sometimes called the *step size* of the quantizer. In this case, the staircase quantizer characteristic of Fig. 1 has equal width and equal height steps. The expression for mean-square error simplifies to

$$D = \frac{\Delta^2}{12} \sum_{i=2}^{N-1} p(y_i)\Delta.$$

But

$$\sum p(y_i)\Delta \approx \int p(s)ds = 1,$$

so that

$$D \approx \frac{\Delta^2}{12} \tag{5}$$

Thus the mean-square distortion of a uniform quantizer grows as the square of the step size. This is perhaps the most often used result concerning quantization. This expression may be obtained directly by regarding the granularity noise as a uniformly distributed random variable over the interval $-\Delta/2$ to $+\Delta/2$ and neglecting overload noise.

A symmetric uniform quantizer is fully described by specifying the number of levels and either the step size Δ or the overload level V where $V = x_N = -x_0$. To avoid significant overload distortion, the overload level is chosen to be a suitable multiple, $y = V/\sigma$, called the *loading factor,* of the rms signal level σ. A common choice is the so-called *four-sigma loading* where $y = 4$. Then the step size is $\Delta = 8\sigma/(N-2)$ since the total amplitude range of the quantizing intervals is 8σ and there are $N-2$ levels in that range. Then, for an n-bit quantizer with $N = 2^n$ and $N \gg 2$, we find using (2) and (5) that

$$\text{SNR} = 6n - 7.3. \tag{6}$$

This linear increase of SNR with the number of bits of quantization was noted by Oliver, Pierce, and Shannon [3] in 1948. Note that changing the loading factor modifies the constant term 7.3, but does not alter the rate of increase of SNR with n. (The rate is actually $20 \log_{10} 2 \approx 6.0$.)

Varying the loading factor for a particular input power level σ^2 is equivalent to varying the input power level for a fixed loading factor. In Fig. 5, the dependence of signal-to-noise ratio on input power level is sketched for a uniform quantizer with $N = 128$. The curve takes into account the effect of overload noise which rapidly becomes dominant as the signal level reaches a critical value. The curve is based on the assumption of a Laplacian pdf,

$$p(x) = \frac{1}{\lambda} e^{-2|x|/\lambda}$$

which is occasionally used to approximate the pdf of speech. In this case it may be seen that the best performance is achieved when the loading factor of 6.1 (15.7 dB) is used. If the input power level deviates a few decibels from the anticipated value (used in designing the quantizer), a substantial drop in SNR will result.

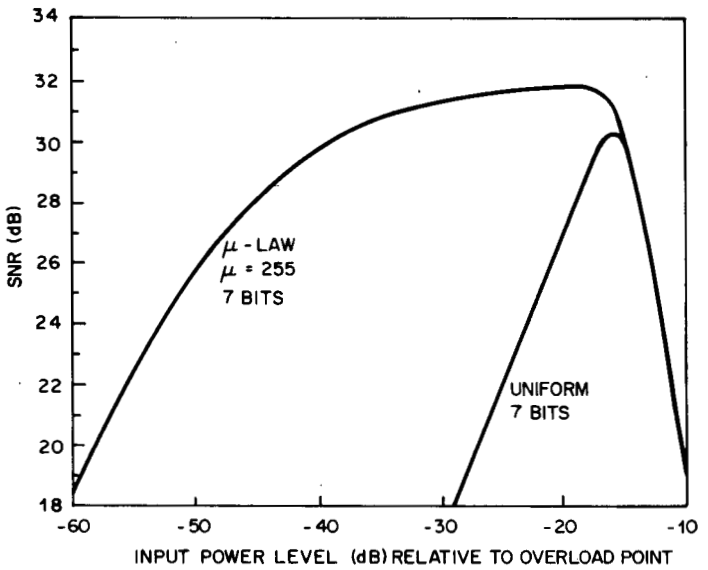


Fig. 5. Dependence of signal-to-noise ratio on the input power level for uniform and μ-law quantizers both having 7 bits of quantization (128 levels). For a minimum acceptable quality of 25 dB, it can be seen that the μ-law quantizer has a dynamic range of about 40 dB, while the uniform quantizer has a range of about 10 dB. The curves may also be used to show how SNR depends on the choice of overload point when the input power level is fixed. Curves are based on a Laplacian input pdf.

IV. COMPANDING

Uniform quantization is not in general the most effective way to achieve good performance. For a given number of quantizing intervals, taking into account the input probability density, nonuniform spacing of the decision levels can yield lower quantizing noise and less sensitivity to variations in input signal statistics. An effective technique for studying nonuniform quantization, used by Bennett [2], is to model the quantizer as a memoryless nonlinearity $F(x)$, the "compressor," followed by a uniform quantizer as shown in Fig. 6. The nonlinearity spreads out low-amplitude sample values over a larger range of amplitudes while shrinking the higher amplitude values into a smaller range. This compressed signal is then uniformly quantized. The effect is to allocate more quantizer levels to the lower amplitudes, which generally have higher probability, and fewer levels to the less frequently occurring higher amplitudes. The output values are then applied to the inverse nonlinearity $F^{-1}(x)$, producing an approximation to the signal originally applied to the compressor. The overall scheme in Fig. 6 is called *companding,* a term combining the words "compressing" and "expanding."

The characteristic $F(x)$ is a monotonically increasing function having odd symmetry, ranging from values $-V$ to $+V$, and

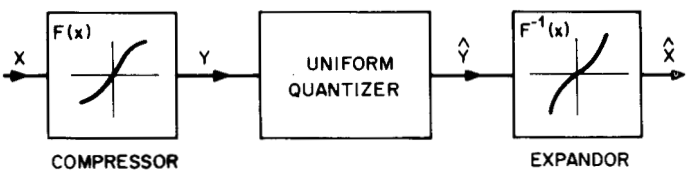


Fig. 6. Companding model of nonuniform quantization.

with $F(V)=V$ and $F(0)=0$. This nonlinear operation, being monotonic, is completely invertible. That is, an input sample x applied to the compressor produces the response value $F(x)$; the original value x could be recovered by applying the value $y = F(x)$ to the inverse nonlinearity, the "expandor" $F^{-1}(y)$, and obtaining x again. Thus, there is no loss of information due to the nonlinear operation itself. The uniform quantizer is chosen to have $N-2$ ($\approx N$) intervals, not including overload regions, so that $\Delta = 2V/N$. The combined effect of the compressor and the uniform quantizer is equivalent to the operation of a particular nonuniform quantizer whose decision levels and output points are determined by the shape of the compressor. Every possible nonuniform quantizer can be modeled in this way by a suitable choice of the function $F(x)$. Fig. 7 shows how the nonuniform quantizer decision levels are related to the uniform quantizer levels.

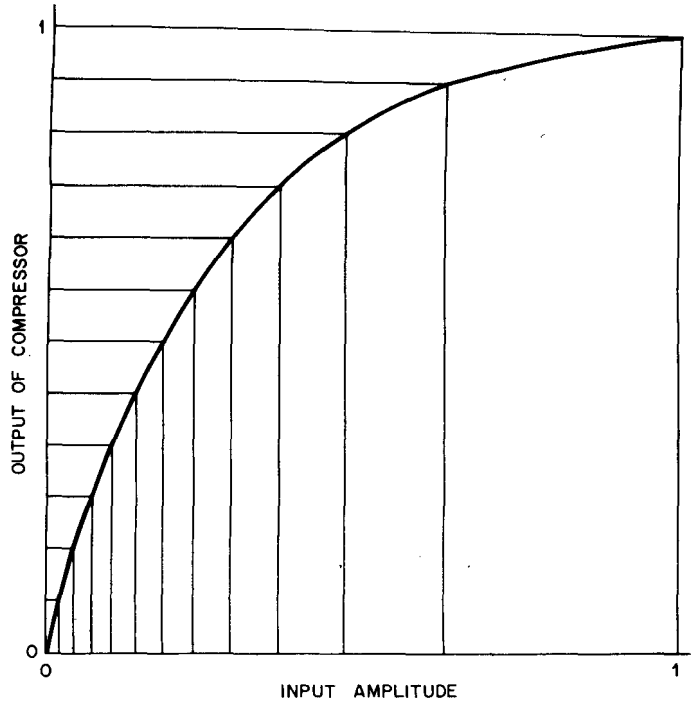


Fig. 7. Compressor mapping of decision levels. (Shown for positive amplitudes only.)

An important approximate formula for mean-square error in nonuniform quantizers can be derived based on the preceding model of nonuniform quantization. For large N, we approximate the curve of $F(y)$ in the ith quantizing interval by a straight-line segment with slope $F'(y_i)$, the derivative of $F(y)$ evaluated at y_i, where y_i is the output point of the equivalent nonuniform quantizer.

Then

$$F'(y_i)\Delta_i \approx F(x_i) - F(x_{i-1}) \approx 2V/N$$

so that, defining the slope of the compressor curve,

$$g(y) \triangleq F'(y),$$

we have

$$\Delta_i \approx \frac{2V}{Ng(y_i)}. \tag{7}$$

Now applying (4) yields

$$D = \frac{V^2}{3N^2} \int_{-V}^{V} \frac{p(s)}{[g(s)]^2}\, ds. \tag{8}$$

This formula, due to Bennett [2], is based on the assumption that N is large and that the overload distortion is negligible. Given a proposed compressor characteristic $F(x)$ and choice of overload point V, the formula (8) can be used to evaluate the resulting quantizer distortion. The formula is also of analytic value for optimizing the compressor characteristic. (See Section VI.)

For speech signals as well as many other analog sources, lower amplitude values occur with higher probability than the higher amplitude values so that it would be reasonable to have quantizer levels more densely packed in the low signal region. For very low signal levels, the relevant step sizes will be approximately uniform with size

$$\Delta_o = \frac{2V}{Ng(0)}.$$

The improvement in performance of the nonuniform quantizer for low signal level inputs over the uniform quantizer is then determined by the ratio

$$c_A = \frac{\Delta}{\Delta_o} = g(0)$$

which is called the *companding advantage.* This quantity is frequently used in comparing different compressor characteristics. Increasing the companding advantage concentrates more levels in the low amplitude region and improves the SNR for weak signal inputs. At the same time, a higher companding advantage means fewer levels in the high amplitude region, tending to reduce the SNR for strong signal inputs.

V. ROBUST QUANTIZATION

In certain applications, notably in speech transmission, the same quantizer must accommodate signals with widely varying power levels. The use of "robust" quantizers, which are relatively insensitive to changes in the probability density of the input samples, has become of great practical importance.

To obtain robust performance, the signal-to-noise ratio of the quantizer should ideally be independent of the particular pdf of the input signal. If the slope of the compressor curve were chosen to be

$$g(x) = \frac{V}{b|x|} \tag{9}$$

then (8) reduces to

$$D = \frac{b^2}{3N^2}\sigma^2$$

so that the signal-to-noise ratio σ^2/D reduces to the constant $3N^2/b^2$, which is in fact independent of $p(x)$. Integrating (9) gives

$$F(x) = V + c \log(x/V) \tag{10}$$

for $x > 0$ where c is a constant. This result shows that such a logarithmic compressor curve would give the desired robust performance. Of course, the formula (8) neglects overload noise so that the SNR will not remain constant but will begin to drop when the input power level becomes large enough. Also, the compressor curve (10) is not in fact realizable since $F(0)$ is not finite. To circumvent the latter difficulty, a modified compressor curve is used which behaves well for small values of x and retains the logarithmic behavior elsewhere.

A compressor curve widely used for speech digitization is the *μ-law* curve (see Fig. 8) given by

$$F(x) = V \frac{\log(1 + \mu x/V)}{\log(1 + \mu)} \tag{11}$$

for $x > 0$. As always, $F(x)$ is an odd function so that $F(x) = -F(-x)$ for negative x. This characteristic was first described in the literature by Holzwarth [5], studied extensively by Smith [12], and reportedly was used by Bennett as early as

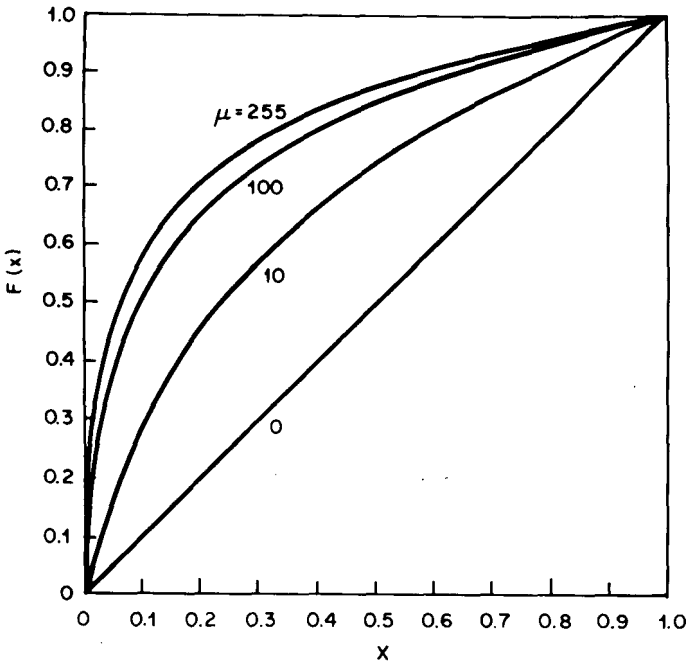


Fig. 8. μ-law compressor curve.

1944 in unpublished work. For $\mu \gg 1$ and $\mu x \gg V$, $F(x)$ approximates the form (10). From (8), the mean-square granular quantizing noise can be calculated, leading to the result

$$D/\sigma^2 = \frac{[\log (1+\mu)]^2}{3N^2} \left\{1 + 2\alpha\, y/\mu + (y/\mu)^2\right\} \qquad (12)$$

where α is the ratio of mean absolute value to rms value of the input samples and y is the loading factor defined earlier. The effects of different choices of μ (corresponding to different companding advantages) has been examined by Smith [12]. Typical values of μ are 100 for 7-bit and 225 for 8-bit speech quantizers. PCM systems in the United States, Canada, and Japan use μ-law companding.

Another robust logarithmic characteristic due to Cattermole [6] is *A-law* companding where

$$F(x) = \begin{cases} \dfrac{Ax}{1+\log A}, & 0 \le x \le V/A \\ \dfrac{V + V \log (Ax/V)}{1+\log A}, & V/A \le x \le V. \end{cases} \qquad (13)$$

A typical value for A is 87.6 for a 7-bit speech quantizer. The A-law characteristic is used in European PCM telephone systems. Both A-law and μ-law have the desired robust quality and can achieve more or less the same performance.

To illustrate the advantage of a robust quantizer, Fig. 5 shows curves of SNR versus input signal power level for both uniform and μ-law quantizers when the number of levels is 128. For a wide range of power levels, a high SNR of the μ-law quantizer is maintained, while the SNR of the uniform quantizer drops rapidly with diminishing power levels. In order to achieve the same quality over a significant dynamic range, an 11-bit uniform quantizer must be used. Thus a saving of 4 bits per sample is achieved by using nonuniform quantization.

In practice, companders are now designed as piecewise linear approximations to a desired characteristic. These "segmented" companding laws are conveniently implemented with digital circuitry. The coded binary word has certain bits that identify to which segment the analog sample belongs and the remaining bits identify which level within the segment represents the analog sample.

VI. OPTIMUM QUANTIZATION

For applications where one particular probability density function is known to describe adequately the distribution of input samples to be quantized, it is natural to seek the best possible quantizer characteristic for that density. Two approaches have been taken to this problem: one uses the assumption that N is large and leads to explicit solutions; the other is valid for any N, and leads to algorithmic procedures for finding the optimum decision levels and output points. We begin with the latter approach.

In a little known Polish article, Lukaszewicz and Steinhaus [7] in 1955 found necessary conditions for optimality of a set of decision levels and output points for both the mean-square and the mean-absolute error criterion. [In the latter case, $[Q(x)-x]^2$ is replaced by $|Q(x)-x|$ in (1).] Independently, in 1957 Lloyd [4], using the mean-square error criterion, found necessary conditions for optimality and an effective algorithm for computing the optimal solution. In 1960, Max [8] independently formulated the necessary conditions for optimality for a kth absolute mean error criterion (including k=2), and rediscovered the same algorithm used by Lloyd. In addition, Max examined the optimization of the step size for uniform quantization. Max also tabulated the optimum quantizer levels for the Gaussian distribution for various values of N.

For the mean-square error criterion, with some fixed value of N, the necessary conditions for optimality on the values of $x_1, x_2, \cdots, x_{N-1}$ and $y_1, y_2, \cdots, y_N$ are found simply by setting derivatives of D as given in (3) with respect to each of these parameters to zero. The resulting conditions are as follows.

1) Each output level of y_j must be the *centroid* or center or mass of the interval R_j with respect to the input density $p(x)$. In other words, y_j is the conditional mean value of the input random variable x given that x is in the region R_j.

2) Each decision level x_j must be halfway between the two adjacent output points.

These conditions do not give the optimum values explicitly, since the value of the output point y_j for an interval R_j depends on the value of the decision levels x_{j-1} and x_j defining R_j, and the decision levels x_j depend on the output levels y_j and y_{j+1}. However, these conditions are used in the Lloyd-Max algorithm (see Max [8]) for computing iteratively a set of parameters that simultaneously satisfy both conditions. Using the Lloyd-Max algorithm, Paez and Glisson [9] tabulated the optimum quantizer parameters for the Laplacian and a particular form of the gamma density.

Lloyd also observed that the conditions, while necessary, are not sufficient conditions for a minimum. In fact, he gave a

counterexample of a probability density function and an associated quantizer that satisfies the conditions and is not optimal. Fleischer [10] obtained sufficient conditions which, if satisfied, will guarantee that the quantizer is in fact optimal. In particular, he showed that, if the input density $p(x)$ satisfies the property that

$$\frac{d^2}{dx^2}[\log p(x)] < 0 \tag{14}$$

for all x, in other words, if $\log p(x)$ is concave, then only one quantizer exists which satisfies the Lloyd-Max conditions 1) and 2) and that quantizer is indeed optimal. It should be noted that the converse is not true, so that it is possible to have a density $p(x)$ not satisfying (14) and yet a unique optimal quantizer may exist. Nonetheless, condition (14) holds for the Gaussian density as well as for many other common densities. Hence, the tabulated quantizer parameters given by Max for the Gaussian density are in fact unique and optimal.

An alternate approach to the search for optimal quantizers begins with the use of Bennett's formula (8), which is based on the assumption that N is large. Minimization of (8) over the class of all curves of compressor slope $g(x)$ that satisfies a suitable constraint yields the result that the optimum compressor slope $g^*(x)$ is proportional to the cube root of the pdf:

$$g^*(s) = c_1[p(s)]^{1/3}.$$

By integrating $g^*(s)$, one obtains the compressor characteristic

$$F^*(s) = c_1 \int_0^s [p(\alpha)]^{1/3}\, d\alpha, \qquad \text{for } s > 0 \tag{15}$$

where c_1 is the constant chosen so that $F(V) = V$. Equation (15) was first obtained by Panter and Dite [11] in a classic and often overlooked paper. Their approach started with (4) and did not make use of Bennett's formula. Direct minimization of (8) was first examined by Smith [12]. Roe [13], while unaware of the works of Panter and Dite and Smith, derived a formula for the optimal decision levels that is equivalent to the result (15), but does not use the companding model. Algazi [14] used the companding model to obtain results on optimal quantizers for a general class of error criteria. His results include (15).

Finally, we note that (15) determines the optimum quantizer for a given choice of overload point V. A separate one-dimensional minimization of D can be used to obtain the best overload point. (See [14].) From Fig. 7, it is evident that once the compressor curve is known, the decision levels and output points are readily obtained by a mapping of the uniform quantizer parameters. Computation of the minimum mean-square error obtained with this approach leads to values in good agreement with Max's tabulations (for the Gaussian pdf), even for values of N as small as 6. For $N = 6$, the individual decision levels are within 3 percent of the correct values (see Roe [12]). Naturally, as N increases, the discrepancy approaches zero, since (15) is based on the assumption that N is large.

An example of optimum quantization studied by Smith [12] is based on the Laplacian pdf. The optimum compressor according to (15) has the form

$$F(x) = \frac{V(1-e^{-mx})}{1-e^{-mV}}, \qquad \text{for } x > 0$$

giving rise to the "m-law" quantizer. Fig. 9 shows the dependence of SNR on input power level for the robust μ-law quantizer with $\mu = 255$ and for the optimum m-law quantizer with $m = 10$ when the input density is Laplacian. Comparison of

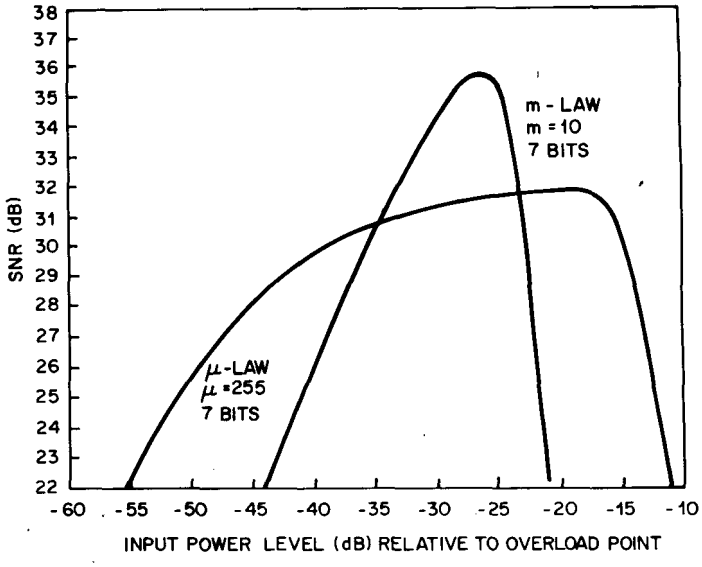


Fig. 9. Dependence of SNR on input power level for μ-law and m-law quantizers when the input pdf is Laplacian and the number of levels is 128 (7 bits). The m-law quantizer is optimal for the Laplacian pdf with input power level 26.5 dB. At this power level there is a 4 dB improvement in SNR over the suboptimum but more robust μ-law quantizer.

SNR performance of μ-law and m-law quantizers shows that, for $\mu = 255$ and $m = 10$, the m-law curve has a 5 dB advantage at the power level for which it is designed. The μ-law quantizer maintains its reasonably high SNR over a broad range of power levels, while the m-law quantizer becomes inferior for input power levels about 10 dB below or 5 dB above the designed value. Comparing the m-law SNR curve in Fig. 9 with the uniform quantizer SNR curve in Fig. 5 shows that there is less than a 6 dB improvement in using the optimum quantizer rather than the uniform quantizer. In some applications this gain might not justify the extra cost of implementing a specially designed nonuniform quantizer as opposed to the simpler uniform quantizer.

A convenient and general way to describe the performance of optimal quantizers is based on the application of the optimal compressor slope $g^*(y)$ to the Bennett formula (8) for mean-square distortion. The result is that the minimum granular distortion for optimal N-point quantization and for large N is given by

$$D = \left\{ \frac{1}{12N^2} \left[\int_{-y}^{y} [p_0(x)]^{1/3} dx \right]^3 \right\} \sigma^2 \tag{16}$$

where $p_0(x)$ denotes the input density normalized to have unit variance, and y is the loading factor discussed above. This formula, first derived by Panter and Dite [11], is useful for estimating the number of quantization levels needed for a desired performance (i.e., SNR specification). The integral in

square brackets, L, depends only on the shape of the input density function and not on the actual power level. It can be seen from (16) that the SNR for an optimal quantizer has the form $10 \log_{10} N^2 - C$ where C is a constant determined by $p_o(x)$ using (16). Letting $n = \log_2 N$ gives the result

$$\text{SNR} = 6n - C$$

where $C = 10 \log_{10} (L^3/12)$. The 6 dB per bit improvement in SNR is the same as for the nonoptimum uniform quantizer where the SNR is given by (6). However, the value of C obtained from (16) is as small as possible since (16) gives the minimum granular distortion attainable for any zero-memory quantizer. For the Gaussian density, C is found to be 4.35 dB by approximating (16) using $V = \infty$.

VII. QUANTIZATION WITH MEMORY

In block quantization, more commonly considered for image digitization rather than speech, a block of k input samples $(x_1, x_2, \cdots, x_k) = \boldsymbol{x}$ (which may be regarded as a vector in k dimensions) is simultaneously quantized, producing an output "point" or vector $(y_{i1}, y_{i2}, \cdots, y_{ik}) = \boldsymbol{y}_i$ approximating $\boldsymbol{x}$. Thus, the output y_{ij} is an approximation to x_j for each $j = 1, 2, \cdots, k$. An N-point quantizer selects one of N output "points" $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N$ to approximate $\boldsymbol{x}$. Unlike zero-memory quantization, the value y_{ij} depends not only on the corresponding input sample x_j, but also on the values of all other samples x_i in the block. Even if the input samples are statistically independent, an advantage can be gained by quantizing a block at a time rather than one sample at a time. A convenient measure of the distortion of the block quantizer is

$$D = \frac{1}{k} \sum_{i=1}^{k} \overline{e_i^2}$$

where $\overline{e_i^2}$ is the mean-square error in the ith sample. The performance of block quantization could be compared with zero-memory quantization by examining how the *bit rate* or *average number of bits per sample*, $B = (\log_2 N)/k$, depends on D, the distortion per sample. Clearly, as the block length k increases, the minimum bit rate needed for a given distortion will decrease. In the limit as $k \to \infty$, the minimum bit rate B approaches a limiting value R depending on D. The function $R(D)$ is the *rate-distortion* function due to Shannon (who defined it in a different way). For certain classes of input process x_i, explicit solutions for $R(D)$ have been found and for many other cases upper and lower bounds are available. For a treatment of rate-distortion theory, see Berger [15].

One simple technique for reducing the bit rate without the full complexity of block quantization is by *entropy-coding* the successive output symbols of a zero-memory quantizer. The output of an N-point zero-memory quantizer is one of N different symbols $y_1, y_2, \cdots, y_N$, each having a corresponding probability $p_1, p_2, \cdots, p_N$ of occurring. Instead of transmitting $\log_2 N$ bits per sample (or the next largest integer if $\log_2 N$ is not an integer) to identify each output sample, variable-length codes such as the Huffman code can be used. Such a code assigns a word with more bits to a low probability symbol and fewer bits to a high probability symbol. The resulting average number of bits per sample attainable approaches or equals the entropy of the quantizer output,

$$H = -\sum_{1}^{N} p_i \log_2 p_i.$$

This scheme requires buffering in order to produce a steady output bit stream.

In general, optimal quantizers do not result in equal probabilities for the output symbols, in which case H is always smaller than $\log_2 N$. For example, a 16-point optimal quantizer for Gaussian samples produces output symbols with entropy 4.73 bits (from Max [8]) compared to the 5 bits per symbol needed for equal-length coding. Once entropy coding is to be used, the preceding optimization theory is no longer relevant. It is more appropriate to find a compressor curve which leads to minimum mean-square error for a constraint on output entropy rather than on number of output points. This leads to the surprising result that the uniform quantizer is nearly optimal! See Gish and Pierce [16].

Finally, another class of quantizers with memory are the sequential quantizers such as delta modulation, differential PCM, and the various adaptive versions of these schemes. In essence, all of these schemes take advantage of correlation in the successive input samples by using a feedback loop around the quantizer. However, this is a subject for a separate paper.

VIII. QUANTIZER PERFORMANCE

From a user's viewpoint, the performance of a quantizer is determined by the number of bits per sample needed to digitize a given analog source so that it can be reproduced with a prescribed maximum amount of distortion (or minimum SNR). Alternatively, the performance is determined by how high an SNR can be achieved for a prescribed average bit rate B measured in bits per sample. We take the latter approach here and survey some key results on achievable quantizer performance. For convenience, we focus only on the case of input samples with a Gaussian pdf and the mean-square distortion measure. The issue of robustness is not considered in this discussion.

From rate-distortion theory it is known that, in the limit as the block length approaches infinity, block quantization of a Gaussian source with statistically independent samples can achieve the bit rate

$$B = \frac{1}{2} \log_2 (\sigma^2/D)$$

where D is the average mean-square distortion per sample (and $D < \sigma^2$). Converting to SNR then gives the result

$$\text{SNR}_1 = 6B, \tag{17}$$

which is also a lower bound on attainable SNR for any realizable quantization scheme regardless of the input pdf as long as the samples are independent. If the source samples are correlated, a higher SNR can always be achieved. See Berger [15].

If a zero-memory uniform quantizer is used with entropy coding of the output symbols, an efficient quantization scheme is achieved. By optimizing the overload point, Goblick and Holsinger [17] found that H*, the highest output entropy

attainable with uniform quantizing, satisfies the equation

$$H^* = \frac{1}{4} + \frac{1}{2}\log_2 (\sigma^2/D).$$

With entropy coding, H^* may be taken as the attainable bit rate B so that solving for SNR gives

$$\text{SNR}_2 = 6B - 1.50. \quad (18)$$

Hence the uniform quantizer with entropy coding achieves an SNR only 1.5 dB below the very best attainable performance with block quantization.

If the best nonuniform quantizer for minimizing distortion is combined with entropy coding, taking the bit rate as the output entropy gives the result

$$\text{SNR}_3 = 6B - 2.45 \quad (19)$$

where (19) was empirically found to fit the data tabulated by Max for quantizers with more than eight levels. Clearly, the nonuniform quantizer is inferior when entropy coding is being used.

Of course, entropy coding adds a significant amount of complexity to the implementation of a quantizer. Without entropy coding, we have seen that the highest SNR achievable with nonuniform quantization is given using (16) by

$$\text{SNR}_4 = 6B - 4.35. \quad (20)$$

Recall that (16) is based on the assumption of large N and it neglects overload noise. The exact SNR values for N between 2 and 36 can be obtained from Max's tables. It turns out that (20) is about 3.6 percent too small for $N = 12$ and becomes progressively more accurate as N increases. See Fig. 10.

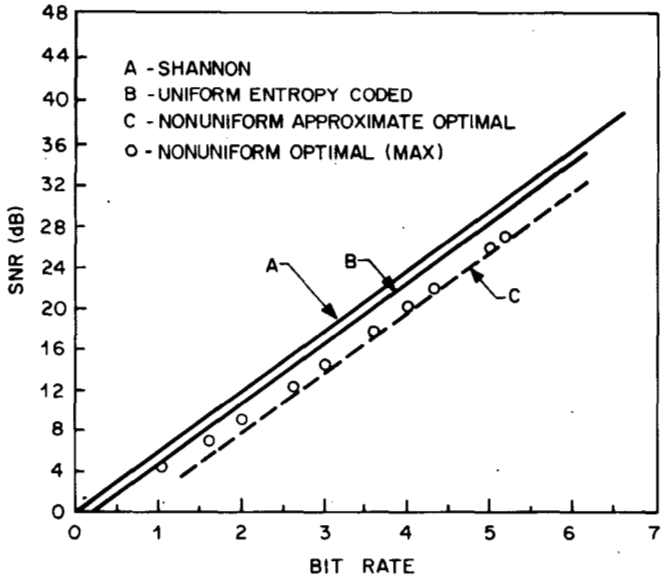


Fig. 10. Quantizer performance in SNR as a function of the average number of bits per sample needed to encode an analog source. Curve A is the best theoretically attainable performance for a Gaussian source with independent samples. It is also a lower bound for any source with independent samples. Curve B is the performance achieved for Gaussian samples with a uniform quantizer followed by entropy encoding. Curve C, based on (16), is asymptotically for a large number of levels the optimal performance obtainable with nonuniform quantization of Gaussian samples without entropy coding. Circled points are based on Max's tabulated values for optimal nonuniform quantization of Gaussian samples without entropy encoding.

Finally, the simplest quantization scheme, using a uniform quantizer without entropy coding, gives the least favorable performance. Using a loading factor of 4 and neglecting overload noise led to the SNR formula (6):

$$\text{SNR}_5 = 6B - 7.3. \quad (21)$$

However, the optimum loading factor depends on the number of levels used and the effect of overload distortion. Goblick and Holsinger [17] fitted the curve

$$B = 0.125 + 0.6\log_2(\sigma^2/D)$$

to Max's tabulated data for uniform quantizers with optimized loading factor. Converting this expression to a SNR formula gives

$$\text{SNR}_6 = 5B - 0.63. \quad (22)$$

Since Max's tables go up to $N = 36$, it is not known how accurate (22) is for $B > 5.2$. For $B < 6.7$, (22) gives higher SNR values than (21), which shows that four-sigma loading is not an optimal choice for a Gaussian pdf.

Summarizing, we have seen that uniform quantizing followed by entropy coding can achieve SNR values within 1.5 dB of the best performance theoretically attainable with any quantization scheme whatever. For an additional 3 dB penalty in SNR, an optimum nonuniform quantizer without the complexity of entropy coding can be used. Simplest of all, the uniform quantizer can achieve an SNR within 7 dB or so of the best performance theoretically attainable.

It should be emphasized that this modest difference in performance between the simplest and most complex quantization schemes is based on the assumption that the input samples are statistically independent. Zero-memory quantization can be grossly inadequate when there is substantial correlation between successive input samples. However, the utility of zero-memory quantizers does not end when the input is correlated. In such situations, the zero-memory quantizer is still used as a component part of more sophisticated quantization schemes. Sequential quantization schemes all use a zero-memory quantizer of one form or another imbedded in a feedback loop. Also, block quantization schemes generally attempt to transform the vector of input samples into a new vector with independent components. These components are then individually quantized with a zero-memory quantizer. Indeed, the basic zero-memory quantizer plays a ubiquitous role in the digital coding of analog sources.

REFERENCES

[1] B. Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory," *IRE Trans. Circuit Theory,* vol. CT-3, pp. 266-276, 1956.

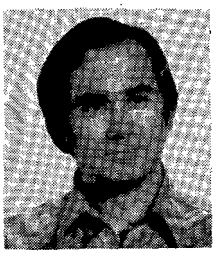
[2] W. R. Bennett, "Spectrum of quantized signals," *Bell Syst. Tech. J.,* vol. 27, pp. 446-472, July 1948.

[3] B. M. Oliver, J. R. Pierce, and C. E. Shannon, "The philosophy of PCM," *Proc. IRE,* vol. 36, pp. 1324-1331, 1948.

[4] S. P. Lloyd, "Least-squares quantization in PCM," unpublished memorandum, Bell Laboratories, 1957 (copies available from the author).

[5] H. Holzwarth, "PCM and its distortions by logarithmic quantization" (in German), *Arch. Elekt. Ubertragung,* vol. 3, pp. 277-285, 1949.

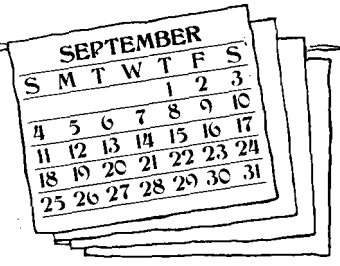
[6] K. W. Cattermole, *Principles of Pulse Code Modulation.* Elsevier, London: Iliffe, 1969.
[7] J. Lukaszewicz and H. Steinhaus, "On measuring by comparison" (in Polish), *Zastos. Mat.,* vol. 2, pp. 225-231, 1955.
[8] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory,* vol. IT-6, pp. 7-12, 1960.
[9] M. D. Paez and T. H. Glisson, "Minimum mean-squared-error quantization in speech PCM and DPCM systems," *IEEE Trans. Commun.,* vol. COM-20, pp. 225-230, Apr. 1972.
[10] P. Fleischer, "Sufficient conditions for achieving minimum distortion in a quantizer," in *IEEE Int. Conv. Rec.,* 1964, pp. 104-111.
[11] P. F. Panter and W. Dite, "Quantizing distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE,* vol. 39, pp. 44-48, 1951.
[12] B. Smith, "Instantaneous companding of quantized signals," *Bell Syst. Tech. J.,* vol. 27, pp. 446-472, 1948.
[13] G. M. Roe, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory,* vol. IT-10, pp. 384-385, 1964.
[14] V. R. Algazi, "Useful approximation to optimum quantization," *IEEE Trans. Commun. Technol.,* vol. COM-14, pp. 297-301, 1966.
[15] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.
[16] H. Gish and J. N. Pierce, "Asymptotically efficient quantization," *IEEE Trans. Inform. Theory,* vol. IT-14, pp. 676-681, 1968.
[17] T. J. Goblick and J. L. Holsinger, "Analog source digitization: A comparison of theory and practice," *IEEE Trans. Inform. Theory,* vol. IT-13, pp. 323-326, Apr. 1967.



Allen Gersho (S'58–M'64) was born in Canada in 1940. He received the B.S. degree from M.I.T. in 1960 and the Ph.D. degree from Cornell University in 1963.

Since 1963 he has been a member of the Technical Staff at Bell Laboratories, Murray Hill, NJ, where his research work has been primarily related to theoretical aspects of communication systems with occasional digressions to circuit theory, social systems modeling, physics, and biophysics. He has recently returned from an academic year (1976-1977) at the System Science Department of UCLA where he taught linear systems and probability theory, as well as giving a series of seminar lectures on the digital coding of analog sources.

Dr. Gersho is a member of the Communication Theory Committee of the IEEE Communications Society. □



Calendar of Events

OCTOBER 1977

International Symposium on Measurements in Telecommunications, Oct. 4-6, 1977, Lannion, France.

Sponsor: URSI, Comm's A.C. *Information:* Mr. J. LeMezec, Colloque URSI-Mesures, CNET, 22301 Lannion, France.

International Symposium on Information Theory, Oct. 10-14, 1977, Cornell University, Ithaca, NY.

Sponsor: IEEE IT. *Information:* Prof. Toby Berger, Dep. Elec. Eng., Cornell University, Ithaca, NY 14853.

Radar-International-RADAR '77, Oct. 25-28, London.

Sponsors: IEE, IEEE AES. *Information:* IEE Conference Dep., Savoy Place, London, WC2R OBL, England.

Lawrence Symposium on Systems and Decisions Sciences, Oct. 3-4, 1977, Berkeley, CA.

Sponsors: ERDA and Lawrence Livermore Lab. *Topics:* Applications of modeling, control, identification, etc., to energy, engineering, and socioeconomic systems. *Information:* D. D. Siljak, School of Engineering, University of Santa Clara, Santa Clara, CA 95053.

ITC/USA (International Telemetering Conference), Oct. 18-20, 1977, Los Angeles, CA.

Sponsor: Int. Foundation for Telemetering. *Topics include:* Communication theory, digital systems, devices, and instruments, and oceanography. *Information:* Robert Dixon, Spectrack Systems Inc., P.O. Box 1164, Cypress, CA 90630.

NOVEMBER 1977

COMPSAC '77 (Computer Software and Applications Conference), Nov. 8-11, 1977, Chicago, IL.

Sponsor: IEEE Computer Society. *Topics:* Information management systems, data communications and computer networking, simulation, etc. *Information:* Prof. Stephen S. Yau, Dep. Computer Sci., Northwestern Univ., Evanston, IL 60201.

DECEMBER 1977

National Telecommunications Conference (NTC), Dec. 5-7, 1977, Los Angeles, CA.

Sponsor: ComSoc CSCB. *Information:* Stanley Butman, JPL, 4800 Oak Grove Drive, Pasadena, CA 91103, (213) 354-2759.

1978

National Radio Science Meeting, Jan. 9-13, 1978, University of Colorado, Boulder, CO.

Sponsor: USNC/URSI, IEEE AP, CAS, IT, etc. *Information:* T. E. Mode, Dep. Elec. Eng., University of Colorado, Boulder, CO 80309.

28th Vehicular Technology Conference, Mar. 22-24, 1978, Denver, CO.

Sponsor: IEEE. *Topics:* Command & control; safety & security; propulsion; communications: voice, data, control, etc. *Information:* John Shafer, U.S. Dep. Commerce, NBS, 325 Broadway, Boulder, CO 80302.

1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, Apr. 10-12, 1978, Tulsa, OK.

Sponsor: IEEE ASSP. *Topics:* Signal & speech processing, underwater acoustics, seismic signal analysis, noise measurements. *Information:* T. H. Crystal, IDA, Thanet Rd, Princeton, NJ 08540.

1978 IEEE International Symposium on Circuits and Systems, May 17-19, 1978, New York, NY.

Sponsor: IEEE CAS. *Topics:* Analysis and design of circuits & systems, analog & digital signal processing, multidimensional filters, computer-aided design, modeling, etc. *Information:* H. E. Meadows, Dep. Elec. Eng. and Comput. Sci., Columbia University, New York, NY 10027. □