On Optimum Quantization

ROGER C. WOOD, MEMBER, IEEE

Abstract—**The problem of minimizing mean-square quantization error is considered and simple closed form approximations based on the work of Max and Roe are derived for the quantization error and entropy of signals quantized by the optimum fixed-*N* quantizer. These approximations are then used to show that, when *N* is moderately large, it is better to use equi-interval quantizing than the optimum fixed-*N* quantizer if the signal is to be subsequently buffered and transmitted at a fixed bit rate. Finally, the problem of optimum quantizing in the presence of buffering is examined, and the numerical results presented for Gaussian signals indicate that equilevel quantizing yields nearly optimum results.**

THE REDUCTION of quantization error by tailoring the structure of the quantizer to the signal to be processed has received considerable theoretical attention in the past. We shall consider this concept for the special case of stochastic signals whose samples are independently and identically distributed. This problem of quantizing for minimum distortion for a signal of known probability density $p(x)$ was first considered in detail by Max [1] in 1961. By assuming the number of levels N to be fixed, Max derived equations for the optimum intervals (y_{k-1}, y_k) and levels x_k. When the criterion is minimum mean-square error, the appropriate equations are

$$x_k = \frac{\int_{y_{k-1}}^{y_k} xp(x)\,dx}{\int_{y_{k-1}}^{y_k} p(x)\,dx} \equiv \mu(y_{k-1}, y_k) \tag{1}$$

and

$$y_k = \tfrac{1}{2}(x_k + x_{k+1}). \tag{2}$$

Hence, the representative levels are the conditional means on the given intervals, and the interval boundaries are halfway between the levels. The analytical solution of these equations is impossible for all but trivial cases, but a numerical solution is straightforward. Moreover, Roe [2] has derived excellent approximate formulas based on Max's results.

The above equations will, in general, require an iteration technique for their solution. One such technique is given by Max, and many others are also feasible.

The purpose of this paper is to derive a simple estimate of the error saving to be obtained by Max's quantizer, which shall be labeled the optimum fixed-N quantizer; to examine the effect on signal entropy of such quantizing, and finally to examine the problem of optimum quantization in the presence of buffering (i.e., for fixed transmission rate and, therefore, fixed entropy).

A Convenient Approximation

Since all forms of the optimizing equations depend on the conditional mean $\mu(\eta_1, \eta_2)$, we shall generate an approximation for that function. To do so, we note that if $p(x)$ is sufficiently well behaved[1] on the interval $(\xi - \Delta/2, \xi + \Delta/2)$ we can generate Taylor's series expansions about ξ for both the numerator and denominator, and therefore, formally, we can write

$$\mu\left(\xi - \frac{\Delta}{2}, \xi + \frac{\Delta}{2}\right) \cong \frac{\xi p(\xi) + \frac{\Delta^2}{24}[\xi p''(\xi) + 2p'(\xi)]}{p(\xi) + \frac{\Delta^2}{24}p''(\xi)} \tag{3}$$

$$= \xi + \frac{\frac{\Delta^2}{12}p'(\xi)}{p(\xi) + \frac{\Delta^2}{24}p''(\xi)} \tag{4}$$

$$\cong \xi + \frac{\Delta^2}{12}\frac{p'(\xi)}{p(\xi)} \tag{5}$$

for Δ small enough.

Thus we have derived, for small intervals, an approximate expression for the conditional mean in terms of the midpoint and length of the given interval.

The Second-Order Moments of the Quantized Distribution

We can give a considerable amount of information about the behavior of the first two moments of the quantized variable. In particular we have the following theorem.

Theorem 1

When the optimum fixed-N quantizer is employed, the first moment of the quantized variable is given by μ, and the second moment can be approximated by

$$V(x^*) \cong \sigma^2 - \sum_k \frac{\Delta_k^3}{12}p(\xi_k)$$

Manuscript received November 17, 1967; revised August 9, 1968. This research was supported in part by the National Science Foundation under Grant GK 144 and by NASA under Research Grant G237-62.

The author is with the Department of Electrical Engineering, University of California, Santa Barbara, Calif. 93106.

[1] Since in all cases we truncate the Taylor's series after several terms, it will suffice for our purposes that the first few (at most, five) derivatives exist and are continuous. Moreover, for the approximation which will be developed to be close, the number of N levels must be large enough (i.e., the interval lengths Δ small enough) so that

$$p(x) >> \Delta p'(x) >> \Delta^2 p''(x) >> \cdots.$$

Thus the critical interval size, for application of the approximations, is seen to depend closely upon the nature of the probability density $p(x)$.

where Δ_k is the length and ξ_k the midpoint of the kth quantizer interval.

Proof: For the first moment, $E(x^*) = E[E(x \mid y_{k-1} < x < y_k)] = \mu$ so that the quantized variable has the same mean as the original continuous variable. Considering the second moment of x^*, we note first that we can write

$$E(x^2) \equiv \int_{-\infty}^{\infty} x^2 p(x)\, dx = \sum_k \int_{\xi_k - 1/2\Delta_k}^{\xi_k + 1/2\Delta_k} x^2 p(x)\, dx. \qquad (6)$$

Expanding $x^2p(x)$ about $x = \xi$ yields, after integrating over the intervals $(\xi_k - \frac{1}{2}\Delta_k, \xi_k + \frac{1}{2}\Delta_k)$,

$$E(x^2) = \sum_k \Bigg(\Delta_k \xi_k^2 p(\xi_k) + \frac{\Delta_k^3}{4 \cdot 3!}[\xi_k^2 p''(\xi_k) + 4\xi_k p'(\xi_k) + 2p(\xi_k)] + O(\Delta_k^5)\Bigg). \qquad (7)$$

Therefore, letting $\mu_k = E[x \mid y_{k-1} < x < y_k]$ and $p_k = P[y_{k-1} < x < y_k]$, we can write

$$\begin{aligned}
E[(x^*)^2] &= \sum_k \mu_k^2 p_k \\
&\cong \sum_k \left\{\left[\xi_k + \frac{\frac{\Delta_k^2}{12}p'(\xi_k)}{p(\xi_k) + \frac{\Delta_k^2}{24}p''(\xi_k)}\right] \cdot \left[\Delta_k \xi_k p(\xi_k) + \frac{\Delta_k^3 \xi_k p''(\xi_k)}{24} + \frac{2\Delta_k^3 p'(\xi_k)}{24}\right]\right\} \\
&\cong \sum_k \left(\Delta_k \xi_k^2 p(\xi_k) + \frac{\Delta_k^3}{24}[4\xi_k p'(\xi_k) + \xi_k^2 p''(\xi_k)]\right) \\
&\cong E(x^2) - \tfrac{1}{12} \sum_k \Delta_k^3 p(\xi_k), \qquad (8)
\end{aligned}$$

for Δ_k small enough. Hence, the variance of the quantized variable is less than that of the continuous variable and is given by

$$V(x^*) = \sigma^2 - \tfrac{1}{12} \sum \Delta_k^3 p(\xi_k) + O(\Delta_k^5) \qquad (9)$$

which completes the proof.

The significance of this result is that the variance of the quantized variable is less than that of the original signal. Hence, the signal and noise are dependent and no pseudo-independence of the sort considered by Widrow [3] is possible. Thus, the common additive noise model is not appropriate for the case of optimum fixed-N quantizing.

Closed Form Approximations for the Mean-Square Error and the Entropy of the Quantized Signal

Although the correction term for the second moment derived above did not possess a convenient closed form, it enabled us to demonstrate the lack of independence between signal and noise. We now develop a general technique for deriving a closed form approximation to the error, and therefore also the correction term for In addition, we derive an approximation to the entropy of the quantized sample.

For a mean-square error criterion, Roe has shown that the interval points for the optimum fixed-N quantizer can be approximated by

$$\int_0^{y_k} [p(x)]^{1/3}\, dx \cong 2C_1 k + C_2 \qquad (10)$$

where C_1 and C_2 are constants, provided only that, in the sense described previously, N is large and $p(x)$ sufficiently differentiable. Clearly, if (y_o, y_N) spans the domain of definition of $p(x)$, the quantity

$$\int_0^{y_N} [p(x)]^{1/3}\, dx - \int_0^{y_o} [p(x)]^{1/3}\, dx$$

depends only on $p(x)$ and $C_1 = O(1/N)$.

Theorem 2

For any signal with probability distribution $p(x)$ well enough behaved for Roe's approximations (10) to be applicable, the mean-square quantization error of the optimum fixed-N quantizer can be approximated, for large N (Δ small), by

$$\epsilon^2 = \tfrac{1}{12}(2C_1)^3 N \qquad (11)$$

where C_1 is given by evaluating (10) and is of the order N^{-1}.

Proof: For any $p(x)$ well enough behaved

$$\int_0^{y_k} [p(x)]^{1/3}\, dx \cong 2C_1 k + C_2\,.$$

If we now define $z(x)$ to be

$$z = \int_0^x [p(t)]^{1/3}\, dt,$$

$$\frac{dx}{dz} = \frac{1}{\frac{dz}{dx}} = [p(x)]^{-1/3}$$

and

$$\Delta_k = [z(y_k) - z(y_{k-1})] \left.\frac{dx}{dz}\right|_{(z=\xi_k)} = 2C_1[p(\xi_k)]^{-1/3}.$$

Hence, we can write

$$\begin{aligned}
\epsilon^2 &\cong \tfrac{1}{12}(2C_1)^2 \int_{y_o}^{y_N} [[p(x)]^{-1/3}]^2 p(x)\, dx \\
&= \tfrac{1}{12}(2C_1)^2 \int_{y_o}^{y_N} [p(x)]^{1/3}\, dx \\
&= \tfrac{1}{12}(2C_1)^2[2C_1 N + C_2 - 2C_1 \cdot O - C_2] \\
&= \tfrac{1}{12}(2C_1)^3 N
\end{aligned}$$

which concludes the proof.

For the case of a Gaussian signal, this procedure yields

$$\epsilon^2 = \frac{2.73N}{(N + 0.853)^3} \qquad (12)$$

	Optimum Fixed N				Equilevel			
	Mean-Square Error		Entropy		Mean-Square Error		Entropy	
N	Exact	Approximate	Exact	Approximate	Exact	Approximate	Exact	Approximate
5	0.0799	0.0797	2.20	2.24	0.176	0.213	1.50	1.37
10	0.0229	0.0232	3.13	3.12	0.0507	0.0533	2.40	2.37
15	0.0107	0.0109	3.68	3.67	0.0240	0.0237	2.97	2.95
20	0.00620	0.00628	4.07	4.07	0.0132	0.0133	3.37	3.36
25	0.00404	0.00408	4.38	4.38	0.00847	0.00852	3.69	3.69
30	0.00283	0.00287	4.64	4.64	0.00590	0.00592	3.95	3.95
35	0.00210	0.00212	4.86	4.86	0.00434	0.00435	4.17	4.17

We now derive asymptotic expressions for the entropy of the quantized signal that depend only upon the properties of the probability density $p(x)$, for both the optimum fixed N and the equilevel quantizer.

Theorem 3

For signals of finite range R and such that the entropy $H(x)$ of the continuous signal is finite, the entropy of the quantized signal, when the optimum fixed-N quantizer is employed, can be approximated by

$$H(x_o^*) \cong \tfrac{2}{3}H(x) - \tfrac{1}{3}\log\left(\frac{12\epsilon^2}{N}\right) \tag{13}$$

for large N.

If equilevel quantizing is performed, the entropy of the quantized signal approaches

$$H(x_e^*) = H(x) - \tfrac{1}{3}\log\left(\frac{12\epsilon^2}{N}\right) - \tfrac{1}{3}\log R \tag{14}$$

Proof: We note that

$$p_k \log p_k = p(\xi_k)\ \Delta_k[\log p(\xi_k) + \log \Delta_k]$$

so that

$$\begin{aligned}
H(x_o^*) &= -\sum_k p_k \log p_k \\
&\cong -(\sum_k p(\xi_k)\log p(\xi_k)\ \Delta_k + \sum_k (\log \Delta_k)p(\xi_k)\ \Delta_k) \\
&\cong H(x) - \sum_k (\log \Delta_k)p(\xi_k)\ \Delta_k \qquad \text{for small } \Delta_k \\
&\cong H(x) - \sum_k (\log [2C_1(p(\xi_k))^{-1/3}])p(\xi_k)\ \Delta_k \\
&\cong H(x) - \log 2C_1 + \frac{1}{3}\int_{y_o}^{y_N} p(\xi)\log p(\xi)\, d\xi \\
&= \tfrac{2}{3}H(x) - \log 2C_1 \\
&\cong \tfrac{2}{3}H(x) - \tfrac{1}{3}\log\left(\frac{12\epsilon^2}{N}\right),
\end{aligned}$$

since

$$\epsilon^2 \cong \frac{(2C_1)^3}{12}N,$$

and

$$2C_1 \cong \left(\frac{12\epsilon^2}{N}\right)^{1/3}.$$

Thus, we have derived an expression for the entropy of the quantized signal, which depends only on the properties of the probability density $p(x)$. We can, in a similar fashion, derive an expression for $H(x^*)$ when equi-interval quantizing is employed. To do this, we note that

$$\begin{aligned}
H(x_e^*) &= -\sum_k p_k \log p_k \\
&\cong -\sum_k p(\xi_k)\ \Delta[\log p(\xi_k) + \log \Delta] \\
&\cong H(x) - \log \Delta \\
&\cong H(x) - \tfrac{1}{3}\log\frac{12\epsilon^2}{N} - \tfrac{1}{3}\log R
\end{aligned}$$

since $\Delta = R/N$ and $\epsilon^2 = \frac{1}{12}\Delta^2$ for N large enough.

The rapid convergence of these approximations, for the case of Gaussian signals, is readily apparent from the data of Table I, which contains exact and approximate computations of entropy and mean-square error for equilevel and optimum fixed-N quantization. In performing the equilevel computations, R was taken to be 8 in the design of the quantizer and the approximations. The exact results, however, are based on the true (infinite) range.

The Application of Buffering and Encoding to the Quantized Signal

In the previous paragraphs, we derived expressions for the mean-square error and the entropy of the quantized signal for both the optimum fixed-N quantizer and for simple equi-interval quantizing. Those estimates are now employed to evaluate the effect of encoding the quantized signals and buffering so that the average bit rate is fixed. We assume, for purposes of comparison, that the mean-square error is fixed, and examine the difference in entropy between signals quantized by the above two devices. For this case, again under the assumption that the probability density $p(x)$ is well behaved, we are able to prove a quite startling and significant theorem about the relative asymptotic behavior of the two types of quantizers.

Theorem 4

Within the limits of our approximation, and therefore asymptotically for large N (given $p(x)$ well behaved and $H(x)$ finite) the output of the optimum fixed-N quantizer has entropy greater than or equal to that of the output of an equilevel quantizer yielding the same mean-square error, provided the range can be assumed to be finite. Therefore, assuming N is large, it is always better[2] to quantize with an equilevel quantizer than with an optimum fixed-N quantizer, if the output signal is to be encoded and transmitted at a fixed average bit rate.

Proof: We note that from (13) and (14), we can write

$$H(x_o^*) - H(x_e^*) \cong \tfrac{1}{3}\log R - \tfrac{1}{3}H(x) + \tfrac{1}{3}\log(N_o/N_e) \tag{15}$$

where the subscripts o and e represent optimum fixed N and equilevel quantizers, respectively. Since the errors are assumed to be equal

$$\frac{(2C_1)^3}{12}N_o = \frac{R^2}{12N_e^2}. \tag{16}$$

Now applying (10) for y_N and y_o we can write

$$2C_1 = \int_0^{y_N} [p(x)]^{1/3}\,dx/N_o\ . \tag{17}$$

Thus, by combining (16) and (17), we can solve for the ratio N_o/N_e and (15) becomes

$$\begin{aligned} &H(x_o^*) - H(x_e^*) \\ &\cong \tfrac{1}{3}H(x) + \tfrac{1}{2}\log\left[\int_{y_o}^{y_N}[p(x)]^{1/3}\,dx\right] \\ &= \frac{1}{2}\left[\frac{2}{3}\int_{y_o}^{y_N} p(x)\log p(x)\,dx + \log\left(\int_{y_o}^{y_N}[p(x)]^{1/3}\right)\right] \\ &= \tfrac{1}{2}[E(\log[p(x)]^{2/3}) + \log E([p(x)]^{-2/3})] \\ &\geqq \tfrac{1}{2}E[\log[p(x)]^{2/3} + \log[p(x)]^{-2/3}] = 0, \end{aligned} \tag{18}$$

since $\log x$ is a concave function. Moreover, equality is achieved if and only if $[p(x)]^{2/3}$ is a constant, that is, for the uniform distribution. For this case, however, there is no difference between the two devices, for the optimum fixed-N quantizer is, in fact, equi-interval. Thus we conclude that for fixed mean-square error, the entropy of a signal quantized by the optimum quantizer is not less than that which obtains if the same signal is quantized by an equi-interval device, at least to the order of our approximation. Since a signal can, by means of encoding, be transmitted at an average bit rate approaching the signal entropy, this implies that an optimum fixed-N quantizer should never by employed if encoding is also to be performed, and the theorem is proved.

To illustrate more fully the significance of these remarks, we will consider the case of unit-variance Gaussian signals in detail. For this case,

$$H(x) = \tfrac{1}{2}\log 2\pi e$$

and

$$H(x_o^*) = -0.3115 + \log(N + 0.853). \tag{19}$$

If the range is taken to be 8 (i.e., $\pm 4\sigma$), we have for the equi-interval case

$$H(x_e^*) \cong -0.953 + \log N. \tag{20}$$

The mean-square error for each case is

$$\epsilon_o^2 \cong \frac{2.73}{(N + 0.853)^2} \tag{21}$$

for the optimum fixed-N quantizer and

$$\epsilon_e^2 \cong 5.33/N^2 \tag{22}$$

for the equi-interval quantizer. If the entropies of the two methods are equated, i.e., if a transmission rate is fixed,

$$-0.3115 + \log(N_o + 0.853) \cong 0.953 + \log N_e$$

so that as N_o, N_e become large

$$0.64 + \log N_o \cong \log N_e$$

and $N_e \cong 1.559\,N_o$.

Thus, expressing the mean-square errors for each case in terms of N_o, we have

$$\epsilon_o^2 \cong \frac{2.73}{(N_o + 0.853)^2} \tag{23}$$

and

$$\epsilon_e^2 \cong \frac{5.33}{(1.559N_o)^2} = \frac{2.19}{N_o^2}. \tag{24}$$

Hence, for N_o moderately large, the error using equi-interval quantizing is less than that using the optimum fixed-N quantizer, if the signals are to be subsequently encoded and transmitted at a fixed bit rate. The reason for this apparent anomaly is that, for a given N, the entropy of the equi-interval quantized signal is considerably less than that of the optimum fixed-N quantized signal; and by employing more levels, this smaller entropy can be converted into lower mean-square error. Thus it is apparent that the optimum fixed-N quantizer loses more in terms of increased entropy than it gains in reduction of mean-square error, if encoding is to be practiced. It should be pointed out that it is necessary to use a buffer to achieve an advantage from any encoding scheme which involves words of variable length. There is therefore an apparent tradeoff between fixing N and using the more complex optimum quantizer but no buffer, and using a ordinary quantizer with a buffer.

It should be noted that as N becomes very large, so also does the quantized entropy. Thus, the indicated difference may be trivial compared to $H(x^*)$. However, for the case of Gaussian signals, (treated as an example in the following sections), there does exist a wide range of values over which the difference is appreciable.

The Optimizing Equations for the Case of Encoded Signals

It was shown in the previous section that if the quantized signal is to be encoded, buffered, and transmitted at a fixed average bit rate, the use of the optimum fixed-N quantizer yields suboptimum results; results which are worse, in fact, than for a simple equilevel quantizer. We

[2] From the viewpoint of minimum error. The practical questions

quantizer subject to the constraint of fixed average bit rate rather than fixed N. For simplicity, we assume that the encoding will be performed efficiently enough so that the entropy of the quantized signal is an adequate measure of the the average output bit rate. We again take the mean-square error as our optimization criterion, although other criteria might be more desirable for certain applications.[3] Under these assumptions, our optimization problem is to find the values of y_k, x_k, and N that minimize the mean-square quantization error

$$\epsilon^2 = \sum_{k=1}^{N} \int_{y_{k-1}}^{y_k} (x - x_k)^2 p(x)\, dx \tag{25}$$

subject to

$$\sum_{k=1}^{N} p_k \log p_k = C, \tag{26}$$

where

$$p_k = \int_{y_{k-1}}^{y_k} p(x)\, dx,$$

i.e., so that the entropy of x^* is constant.

It is easily shown that the x_k must be given by

$$x_k = \mu(y_{k-1}, y_k),$$

so that the problem is expressible solely in terms of the y_k and N.

As was the case for the optimum fixed-N quantizer, an analytic solution to these equations can be found only for trivial cases. Therefore, for the particular case of Gaussian signals, the optimizing equations were converted to steepest-descent equations, which were implemented and solved on the System Development Corporation Q-32 computer under the control of the TSS time-sharing system. The correctness of the computer program was checked by suppressing the entropy constraint, which gave results that agreed with Max's. Next, the entropy term was made dominant and the constraint set to that for maximum entropy. The results again agreed with the theoretical (i.e., equal p_k).

The convergence of the steepest-descent equations is very slow, demonstrating that, as is also true for the fixed-N optimum, the mean-square quantization error is very insensitive to moderately small deviations from the optimum interval structure.

The numerical results for Gaussian signals are rather surprising, in that the optimum fixed-entropy quantizer yields an error rate almost negligibly different from that of simple equilevel quantizing, except for very small N. These results are displayed in Fig. 1, from which we can see that the optimum fixed-entropy quantizer displays a marked improvement over the optimum fixed-N quantizer, but not over equilevel quantizing. In fact, the equilevel quantizer suffers in the comparison given because the range was assumed to be 8. For very small N, this obviously gives a poor choice of levels, and a better measure for comparison would be the optimum equilevel quantizer [1].

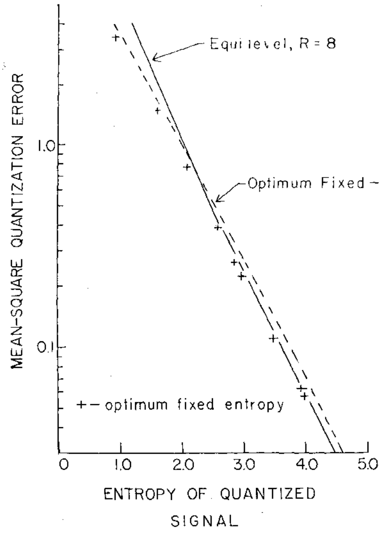

Fig. 1. Comparison of entropy versus mean-square error for equilevel, optimum fixed-N and optimum fixed entropy quantization.

Thus, equilevel quantizing, currently employed because of its ease of implementation, is superior to optimum fixed-N quantizing if the output signal is to be buffered, for all but very small values of N. Moreover, because of the insensitivity of the mean-square quantization error to moderate changes in interval structure, equilevel quantizing gives nearly optimum results for the special case of Gaussian signals.[4]

References

[1] J. Max, "Quantizing for minimum distortion," *IRE Trans. Information Theory*, vol. IT-6, pp. 7–12, March 1960.
[2] G. M. Roe, "Quantizing for minimum distortion," *IEEE Trans. Information Theory*, vol. IT-10, pp. 384–385, October 1964.
[3] B. Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory," *IRE Trans. Circuit Theory*, vol. CT-3, pp. 266–276, December 1956.
[4] R. C. Wood, "Optimum quantizing in hybrid computation," Ph.D. dissertation, Dept. of Engrg., University of California, Los Angeles, August 1966.

[3] Prof. Leo Breiman, of the University of California, Los Angeles, has suggested as an alternative formulation that the mutual information $I(x, x^*)$ be maximized subject to $H(x^*) = C$.

[4] It has been brought to my attention by Dr. G. M. Roe, in private communication, that it is possible to obtain an analytical derivation of these computer based conclusions, and to extend them to well-behaved density functions other than the Gaussian. Specifically, for the fixed-entropy quantizer discussed above, the optimum level spacing approaches the equal-interval case, with

$$\frac{d}{dk}(y_k) \cong \sqrt{\lambda}\left\{1 + \frac{\lambda}{24}\left[\left(\frac{p'}{p}\right)^2 - \frac{4}{5}\frac{p''}{p}\right] + O(\lambda^2)\right\}$$

with λ (a Lagrange multiplier) given by

$$\lambda = e^{2(H_1 - H_0)}$$

$$H_1 = \int_{y_0}^{y_N} p(x) \log p(x)\, dx$$

$$H_0 = \sum_{k=1}^{N} p_k \log p_k$$

and

$$\epsilon^2 \cong \tfrac{1}{12}\lambda$$