the assumptions $a_1 > a_0$ and $a_0, a_1 > 1.0$; then (II-12) becomes

$$\log [X(\omega)] = \log [S(\omega)] + \log [a_1 e^{-j\omega t_1} (1 + a_0/a_1 e^{j\omega(t_1-t_0)} + 1/a_1 e^{j\omega t_1}) \quad \text{(II-15)}$$

or

$$\log [X(\omega)] = \log [S(\omega) a_1 e^{-j\omega t_1}] + \log (1 + a_0/a_1 e^{j\omega(t_1-t_0)} + 1/a_1 e^{j\omega t_1})$$

and application of the same log series will give

$$\begin{aligned}\log [X(\omega)] = \log [S(\omega) a_1 e^{-j\omega t_1}] &+ a_0/a_1 e^{j\omega(t_1-t_0)} \\ &+ 1/a_1 e^{j\omega t_1} - a_0{}^2/2a_1{}^2 e^{j2\omega(t_1-t_0)} \\ &- \tfrac{1}{2} a_1{}^2 e^{j2\omega t_1} - a_0/a_1{}^2 e^{j\omega(2t_1-t_0)} + \cdots . \quad \text{(II-16)}\end{aligned}$$

The complex cepstrum is

$$\begin{aligned}&F^{-1}\{\log [X(\omega)]\} \\ &= F^{-1}\{\log [S(\omega) a_1 e^{-j\omega t_1}]\} + a_0/a_1\, \delta[t + (t_1 - t_0)] \\ &\quad + 1/a_1\, \delta(t + t_1) - a_0{}^2/2a_1{}^2\, \delta[t + 2(t_1 - t_0)] \\ &\quad - \tfrac{1}{2} a_1{}^2\, \delta(t + 2t_1) - a_0/a_1{}^2\, \delta[t + (2t_1 - t_0)] + \cdots . \quad \text{(II-17)}\end{aligned}$$

Therefore, the complex cepstrum for multiple echoes with amplitudes greater than that of the wavelet consists of the inverse transform of the complex logarithm of the transform of the echo with the greatest amplitude, plus delta functions all located in either positive or negative time, depending upon whether $t_0 > t_1$ or $t_1 > t_0$, respectively. It is also apparent that the cases concerned with amplitudes greater than unity are maximum phase while those with amplitudes less than unity are minimum phase situations.

References

[1] D. G. Childers, R. S. Varga, and N. W. Perry, Jr., "Composite signal decomposition," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 471–477, Dec. 1970.

[2] S. Senmoto and D. G. Childers, "Adaptive decomposition of a composite signal of identical unknown wavelets in noise," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, pp. 59–66, Jan. 1972.

[3] B. P. Bogert, M. J. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proc. Symp. Time Series Analysis*, M. Rosenblatt, Ed. New York: Wiley, 1963, ch. 15, pp. 209–243.

[4] A. M. Noll, "Short-time spectrum and cepstrum techniques for vocal-pitch detection," *J. Acoust. Soc. Amer.*, vol. 36, pp. 296–302, Feb. 1964.

[5] B. P. Bogert and J. F. Ossanna, "The heuristics of cepstrum analysis of a stationary complex echoed Gaussian signal in stationary Gaussian noise," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 373–380, July 1966.

[6] A. V. Oppenheim, "Superposition in a class of nonlinear systems," Res. Lab. Electron., M.I.T., Cambridge, Mass., Tech. Rep. 432, Mar. 31, 1965.

[7] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, Jr., "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264–1291, Aug. 1968.

[8] R. W. Schafer, "Echo removal by discrete generalized linear filtering," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 1968.

[9] T. J. Cohen, "Source-depth determinations using spectral, pseudo-autocorrelation and cepstral analysis," *Geophys. J. Roy. Astron. Soc.*, vol. 20, pp. 223–231, 1970.

[10] T. J. Ulrych, "Application of homomorphic deconvolution to seismology," *Geophysics*, vol. 36, pp. 650–660, Aug. 1971.

[11] J. C. Prabhakar and S. C. Gupta, "Separation of Rayleigh and Poisson density functions through homomorphic filtering," *Nat. Electronics Conf.*, pp. 605–610, Dec. 1970.

[12] G. D. Bergland, "A guided tour of the fast Fourier transform," *IEEE Spectrum*, vol. 6, pp. 41–52, July 1969.

[13] R. C. Kemerait, "Signal detection and extraction by cepstrum techniques," Ph.D. dissertation, Univ. Florida, Gainesville, 1971.

Optimum Quantizers and Permutation Codes

TOBY BERGER, MEMBER, IEEE

Abstract—**Amplitude quantization and permutation encoding are two of the many approaches to efficient digitization of analog data. It is shown in this paper that these seemingly different approaches actually are equivalent in the sense that their optimum rate versus distortion performances are identical. Although this equivalence becomes exact only when the quantizer output is perfectly entropy coded and the permutation code block length is infinite, it nonetheless has practical consequences both for quantization and for permutation encoding. In particular, this equivalence permits us to deduce that permutation codes provide a readily implementable block-coding alternative to buffer-instrumented variable-length codes. Moreover, the abundance of methods in the literature for optimizing quantizers with respect to various criteria can be translated directly into algorithms for generating source permutation codes that are optimum for the same purposes.**

The optimum performance attainable with quantizers (hence, permutation codes) of a fixed entropy rate is explored too. The investigation reveals that quantizers with uniformly spaced thresholds are quasi-optimum with considerable generality, and are truly optimum in the mean-squared sense for data having either an exponential or a Laplacian distribution. An attempt is made to provide some analytical insight into why simple uniform quantization is so good so generally.

Manuscript received January 3, 1972; revised February 25, 1972.

The author is with the School of Electrical Engineering, Cornell University, Ithaca, N.Y. 14850.

I. Introduction and Synopsis

Although communication and information theorists have suggested many novel digitization techniques, simple quantization continues to be used almost universally in practice. The widespread preference for quantization has a sound basis. Quantizers are relatively easy to implement and, moreover, their encoding performance usually is nearly optimum. For example, in the case of minimum-meansquare digitization of a Gaussian sequence, quantizers with uniformly spaced levels have entropies that exceed the rate-distortion function lower bound by only one fourth of a bit [1].

The main drawback to quantization is that a variable-length code must be employed if one wishes to ensure that the actual bit rate only barely exceeds the quantizer entropy. Moreover, if very accurate reproduction is required, the quantizer must have many levels, some of which are much more probable than others.[1] This means that certain words in the variable-length code have to be much longer than others, which leads to difficult buffering problems [2].

Permutation codes for sources [3], [4] provide a synchronous alternative to buffer-instrumented variable-length encoding. In this paper we show that, given any quantizer, there exists a permutation code whose entropy rate R and average distortion D approximate those of the quantizer as closely as desired. When D is very small, the block length n of the permutation code in question has to be very large. Intuition notwithstanding, however, we show that the principal task in permutation encoding, that of partially ordering the n source outputs, actually becomes easier to perform as n gets large. As a result, permutation codes become so easy to implement for large n that they offer an attractive alternative to buffer-instrumented variable-length encoding in those applications in which the associated block-coding delay is tolerable.

The optimum R versus D performance attainable with quantizers (hence, permutation codes) is explored too. In the case of the squared-error distortion measure, the optimum quantizer is specified by a set of simultaneous nonlinear equations that can be solved recursively. Investigation reveals that quantizers with uniformly spaced thresholds perform effectively as well as do the optimum quantizers generated by the recursive solution procedure. This result holds not only in the limit as $D \to 0$, when the optimum quantizer itself is known to have threshold spacings that tend toward uniformity [5], but also for moderate and large values of D. Moreover, if the source outputs are governed by either an exponential density or a Laplacian density, then the optimum quantizer is shown to be characterized by threshold spacings that are exactly uniform. Some analytical insight into why uniform quantization is so good so generally is provided by an examination of optimum quantizers for piecewise-constant probability densities and rth-power distortion measures.

II. Basic Equations of Quantization

Let X denote the real random variable to be digitized and let $F(\cdot)$ denote its cumulative distribution function. A device with input X and output Y will be called a quantizer if $a_{i-1} < X \leq a_i$ implies $Y = y_i$. The a_i are called the quantization thresholds and the y_i are called the reconstruction levels. We shall assume without any loss of generality that for all i

$$p_i = F(a_i) - F(a_{i-1}) > 0. \tag{1}$$

Although in practice there are only finitely many levels $\{y_i\}$ and thresholds $\{a_i\}$, we shall allow for the possibility of a countable infinity of levels and thresholds.

With each quantizer we associate two quantities called the entropy rate R and the average distortion D. These are defined by

$$R = -\sum_i p_i \log p_i \tag{2}$$

and

$$D = E|Y - X|^r = \sum_i \int_{a_{i-1}}^{a_i} |x - y_i|^r \, dF(x). \tag{3}$$

Here, and in all that follows, it is assumed that $E|X|^r < \infty$.[2] An optimum quantizer is one that minimizes D for fixed R. Optimization of quantizers is discussed in Sections VI and VII.

III. Basic Equations of Permutation Codes

A permutation code of block length n is a collection of real n-vectors, called codewords, with the following structure. For some set $\{n_i\}$ of nonnegative[3] integers satisfying

$$\sum_i n_i = n \tag{4}$$

and some strictly increasing set $\{\mu_i\}$ of real numbers, the code consists of all n-vectors that have n_i of their components equal to μ_i for each i. Clearly, the codewords all are permutations of one another and the number of words in the code is

$$N = \frac{n!}{\prod_i n_i!}. \tag{5}$$

Let $B = \{y_1, \cdots, y_N\}$ be a permutation code of block length n with parameters $\{n_i\}$ and $\{\mu_i\}$, and define

$$S_i = \sum_{j \leq i} n_j. \tag{6}$$

Theorem 1: Given any real n-vector $\boldsymbol{x} = (x_1, \cdots, x_n)$, the codeword $\boldsymbol{y} = (y_1, \cdots, y_n) \in B$ that minimizes

$$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{k=1}^{n} |x_k - y_k|^r, \qquad r \geq 1, \tag{7}$$

is obtained by replacing the $S_{i-1} + 1$ through S_i smallest components of $\boldsymbol{x}$ by μ_i for all i.

Proof: This is a special case of Theorem 1 of Berger *et al.* [4].

Now consider a random n-vector $\boldsymbol{X} = (X_1, \cdots, X_n)$ with

[1] For a broad and interesting class of distortion measures, the thresholds of the quantizer whose entropy is minimum for a specified average distortion D become uniformly spaced as $D \to 0$ (cf. Section VII). This phenomenon accounts for the highly nonequiprobable nature of the output levels when an accurate reproduction is required.

[2] Although $E|Y - X|^r$ can be made to be finite even when $E|X|^r$ does not exist by employing an appropriate infinite-level quantizer, such cases are of limited interest.

[3] Obviously, at most n of the n_i are nonzero. Allowing the n_i to be zero is notationally convenient in what follows because it avoids explicit reindexing of the n_i as n varies; in general, we have countably many n_i indexed both negatively and positively even for finite n.

statistically independent components each of which is distributed as the random variable X of Section II. It should be clear that, if the permutation code B is used to encode the value x assumed by X, then each of the codewords has probability $1/N$ of being used. The number of bits per component needed to encode X with B (i.e., to specify the index of the resulting codeword) therefore is

$$R = n^{-1} \log_2 N = n^{-1} \left(\log_2 n! - \sum_i \log_2 n_i!\right). \quad (8)$$

The average distortion per component that results from encoding X with B is

$$D = E\left[n^{-1} \sum_i \sum_{j=S_{i-1}+1}^{S_i} |X_n^j - \mu_i|^r\right], \quad (9)$$

where X_n^j is the jth smallest component of X.

IV. Equivalence of Quantizers and Permutation Codes

Although quantization and permutation encoding are two seemingly different approaches to source digitization, the following theorem establishes that they actually are equivalent in the sense that their optimum R versus D performances are identical.

Theorem 2: Let X be a random variable with cumulative distribution function $F(\cdot)$, and let $\{X_k\}$ be a sequence of independent random variables identically distributed as X. Given any quantizer $(\{a_i\},\{\gamma_i\})$ that encodes X with finite rate R and finite distortion D, there exists a sequence of permutation codes B_n of block length n, $n = 1,2,\cdots$, that encode $(X_1,\cdots,X_n)$ with respective rates R_n and per-component average distortions D_n that satisfy both $\lim_{n\to\infty} R_n = R$ and $\lim_{n\to\infty} D_n = D$.

Proof: For all n let the parameter set $\{\mu_i\}$ of the code B_n equal the set $\{\gamma_i\}$ of output levels of the quantizer. Let the other parameter set $\{n_i\}$ of the code B_n vary with n in such a way that

$$\lim_{n\to\infty} n^{-1}S_i = F(a_i) \quad (10a)$$

or equivalently in such a way that

$$\lim_{n\to\infty} n_i/n = \lim_{n\to\infty} n^{-1}(S_i - S_{i-1}) = F(a_i) - F(a_{i-1}) = p_i. \quad (10b)$$

Since n_i grows linearly with n because $p_i > 0$, we know that $\log_2 n_i! \sim n_i \log_2 n_i - n_i \log_2 e + o(n)$, so from (4), (8), (10b), and the fact that R is finite, we have

$$R_n \sim \log_2 n - \sum_i (n_i/n) \log_2 n_i + \left(1 - n^{-1} \sum_i n_i\right) \log_2 e$$
$$= \sum_i p_i \log_2 n - \sum_i (n_i/n) \log_2 n_i \to -\sum p_i \log p_i = R.$$

Upon comparing (3) and (9), we see that the proof will be complete if we can show that

$$\lim_{n\to\infty} E\left[n^{-1} \sum_i \sum_{j=S_{i-1}+1}^{S_i} |X_n^j - \gamma_i|^r\right] = \sum_i \int_{a_{i-1}}^{a_i} |x - \gamma_i|^r \, dF(x). \quad (11)$$

We do this in the Appendix by establishing both that the limit on the left side of (11) is finite and that the desired convergence in fact holds with probability one, namely

$$n^{-1} \sum_i \sum_{j=S_{i-1}+1}^{S_i} |X_n^j - \gamma_i|^r \xrightarrow[\substack{\text{with}\\ \text{probability 1}}]{} \sum_i \int_{a_{i-1}}^{a_i} |x - \gamma_i|^r \, dF(x). \quad (12)$$

These two results together imply the validity of (11), thereby completing the proof.

We see from Theorem 2 that the best permutation code is at least as good as the best quantizer in the (R,D) sense. Conversely, Theorems 1 and 2 together imply that the performance of the best permutation code for $r \geq 1$ is no better in the limit of infinite block length than is that of the best quantizer. Although a proof is lacking, it seems reasonable to conjecture that the performance of the optimum permutation code of block length n and rate R or less can only improve with increasing n; this indeed has been the case in all examples investigated to date. The validity of this conjecture would imply that, at least for $r \geq 1$, the best quantizer is as good as the best permutation code, too. The source coding significance of the intimate relationship between quantizers and permutation codes is explored further in the next section.

V. Permutation Codes Versus Variable-Length Codes

Since we now know that the R versus D performance of an optimum quantizer and variable-length code is also attainable via permutation coding, we must address the question of which of the two techniques is better suited to a given application. If a small value of D is required, then n must be made very large in order for the rate of the permutation code to approach that of the entropy-coded quantizer (cf. [4]). In certain applications the concomitant coding delay may become intolerable, in which case buffer-instrumented variable-length coding of the quantizer outputs is probably the more desirable alternative. We say "probably" rather than "certainly" because buffer overflows usually occur after $|X|$ has assumed large improbable values on several successive samples. The average distortion incurred per sample lost because of a buffer overflow is therefore inordinately large compared to $E|X|^r$. This means that a very long buffer must be employed in order truly to realize a small required value of D. This, in turn, results in a large average coding delay, especially if the probability of buffer underflow must be kept very small also in order to ensure operation at a rate that only barely exceeds the quantizer output entropy. Since detailed analytical investigation of the average distortion and average coding delay associated with buffer-instrumented variable-length encoding of quantizer outputs is lacking at present, it is not entirely clear that this technique is superior to permutation coding even from the standpoint of coding delay.

For applications in which large coding delays are tolerable, we submit that permutation codes are preferable to variable-length codes. Since permutation codes are a subclass of block codes, they operate synchronously and thereby avoid all the buffering problems discussed above. Perhaps even more important, and certainly more surprising, is the fact that permutation codes become increasingly simpler to implement as the block length n increases. In this regard it has been shown [4] that the effort required for (noiseless) channel encoding and decoding of the index of the selected permutation grows only linearly with block length. The potentially troublesome operation is that of partially ordering the source outputs in the manner prescribed by Theorem 1 in order to effect optimum source encoding. A complete ordering would require a number of comparisons that grows as $n \log n$ [6], but this difficulty can be circumvented in the case of the desired partial ordering.[4] In particular, for large n we can capitalize on the law of large numbers as follows. Instead of partially ordering the source outputs in the prescribed manner, we simply quantize them individually with the quantizer that corresponds in the sense of Theorem 2 to the permutation code being employed. Although the random number N_i of outputs that fall in the ith quantization bin $[a_{i-1},a_i)$ usually will not be exactly n_i, $|N_i - n_i|$ will be $0(\sqrt{n})$ in the limit of large n with probability 1. Hence, we can closely approximate the codeword that corresponds to the desired partial ordering simply by replacing the $S_{i-1} + 1$ through S_i smallest quantized source outputs by μ_i. Ties may be broken according to any scheme whatever when ordering the quantizer outputs, so no additional ordering need be done beyond the partial ordering already effected by the quantization itself. In other words, we force the desired composition $\{n_i\}$ by a procedure, which in effect removes certain of the quantized samples from their actual quantization bins and places them in neighboring bins. This results in an average distortion D_n that of course exceeds the average distortion D between the source outputs and the quantizer outputs. However, since the number M of quantized samples that have to be moved out of their bins satisfies $n^{-1}M \to 0$ with probability 1, we have $D_n \to D$ with probability 1. Asymptotically in n, then, the scheme in question circumvents the partial-ordering problem entirely with no degradation in performance. For moderately large values of n, it may prove advisable to establish guard bands $(a_i - \delta, a_i + \delta)$ around the quantization thresholds and then to move the samples that fall in these bands first when breaking ties, thereby yielding a D_n somewhat closer to D.

The preceding discussion strongly suggests that permutation coding is a very promising technique for source digitization when large coding delays are tolerable because the encoding effort per source output does not increase with the block length. Additional light has been shed on the intimate relationship between permutation codes and quantizers, with particular emphasis on the sense in which permutation coding provides a possible replacement not for the quantizer itself, but rather solely for the variable-length coding of the quantizer outputs.

VI. Optimum Quantizers

Since optimization of a permutation code is tantamount to optimization of the quantizer that corresponds to it in the sense of Theorem 2, it is of interest to be able to determine the parameters of an optimum quantizer. In this regard the reconstruction levels $\{\gamma_i\}$ have no effect on the entropy rate R, so they always should be chosen to minimize D. A simple calculation reveals that the optimum γ_i is specified uniquely in terms of a_{i-1}, a_i, and r by the requirement

$$\int_{a_{i-1}}^{\gamma_i} (\gamma_i - x)^{r-1}\, dF(x) = \int_{\gamma_i}^{a_i} (x - \gamma_i)^{r-1}\, dF(x). \quad (13)$$

Since it usually is very difficult to solve (13) for γ_i explicitly for general r, we shall specialize to the important case $r = 2$. In this case (13) reduces to the well-known result that γ_i is the mean of X conditional on the fact that $a_{i-1} < X \leq a_i$, namely

$$\gamma_i = \int_{a_{i-1}}^{a_i} x\, dF(x) \Big/ \int_{a_{i-1}}^{a_i} dF(x) = (1/p_i) \int_{a_{i-1}}^{a_i} x\, dF(x). \quad (14)$$

For $r = 2$, then, the task of designing an optimum quantizer of rate R reduces to that of choosing the thresholds $\{a_i\}$ so as to minimize D of (3) subject to (14) and to the fact that $-\sum p_i \log p_i$ must equal the specified value of R. Toward this end we use (14) and (1)–(3) to express the quantity $J = D + \lambda^{-1}R$ solely in terms of the $\{a_i\}$, and then set $dJ/da_i = 0$. This yields a set of simultaneous nonlinear equations indexed by i that can be put in the form

$$p_{i+1} = p_i \exp\left[\lambda(\gamma_{i+1} - \gamma_i)(\gamma_{i+1} + \gamma_i - 2a_i)\right], \quad (15)$$

where the Lagrange multiplier λ must be selected to achieve the desired value of R. It is very difficult to solve (15) for the $\{a_i\}$ because p_i and γ_i are themselves rather complicated functions of a_{i-1} and a_ι via (1) and (14), respectively. This probably explains why optimum quantizers of a fixed entropy rate were not determined long ago. It turns out, however, that (15) can be solved recursively as follows. If it is assumed that a_{i-1}, a_i, and λ are known, then p_i and γ_i can be computed from (1) and (14). The only unknowns that then remain in (15) are p_{i+1} and γ_{i+1}, both of which are increasing functions of a_{i+1}. It follows that, by gradually increasing our guess of the amount by which a_{i+1} exceeds a_i, we eventually reach the value of a_{i+1} for which the two sides of (15) are equal. With a_{i+1} now known, the same procedure can be used to determine a_{i+2}, and so forth.

The recursive procedure previously described yields a three-parameter family of quantizers that satisfy (15), the parameters being a_0, a_1, and λ. (It should be clear how $a_{-1}, a_{-2}, \cdots$ can be determined recursively in a manner similar to that just described for determining $a_2, a_3, \cdots$.)

[4] If the number of nonzero n_i remains bounded as $n \to \infty$, which corresponds in the sense of Theorem 2 to a finite-level quantizer, then effecting the desired partial ordering consists of locating the positions of a fixed number of prescribed quantiles in a sample of size n. The number of comparisons needed to accomplish this is known to grow only linearly with n [7].

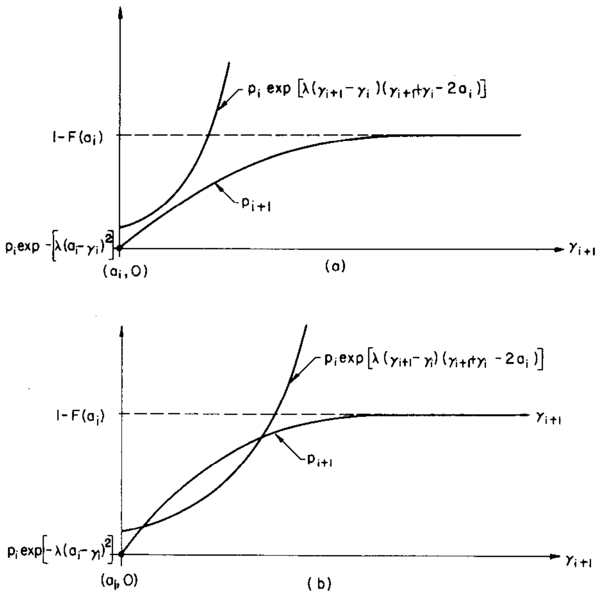

Fig. 1. Graphical solution of (15). (a) No solution. (b) Two solutions.

However, since (15) is only a necessary condition for optimality, not all the quantizers in this three-parameter family are optimum. Thus, it is necessary to determine which of the many quantizers in this family that have the desired rate R has the least average distortion. Fortunately, in many cases of interest the situation is not quite this desperate. If, for example, $F(\cdot)$ possesses a density $f = F'$ that is symmetric about its mean, then it is clear that either the interval (a_0,a_1) should be centered about the mean or a_0 should equal the mean. This reduces the problem to investigation of a pair of two-parameter families of quantizers. If there is a finite number c such that $F(x) = 0$ for $x < c$ and $F(x) > 0$ for $x > c$, then it is clear that we may take $a_0 = c$ and need determine the a_i only for i positive; again, we obtain a two-parameter family of quantizers indexed by a_1 and λ. Similarly, if there exists d such that $F(x) = 1$ for $x \geq d$ and $F(x) < 1$ for $x < d$, then we may set $a_0 = d$ without loss of generality, thereby obtaining a two-parameter family of quantizers $\{a_i, i \leq 0\}$ indexed by a_{-1} and λ. Finally, if both c and d exist with the above properties, then all quantizers of interest have only finitely many thresholds of the form $c = a_0 < a_1 < a_2 < \cdots < a_K = d$ for some K. It follows that only certain choices of the pair (a_1,λ) will yield a recursively determined set of $\{a_i\}$, one of which equals d exactly. Hence, in such an instance there are only a finite number of quantizers that satisfy (15) for each value of λ, but we have to solve a two-point boundary value problem in order to determine them.

To make matters worse, the solution of (15) for a_{i+1} in terms of a_{i-1}, a_i, and λ is unique only if $F(\cdot)$ is continuous and λ is negative, whereas the best quantizers we have been able to find in the examples we have studied to date all correspond to positive values of λ. For $\lambda > 0$, (15) can have several solutions $a_{i+1}(a_{i-1},a_i,\lambda)$. Usually, however, one of the two situations sketched in Fig. 1 prevails. In Fig. 1(a) there are no solutions, which means that $a_{i+1} = \infty$, i.e., a_i is the last finite threshold. In Fig. 1(b) there are two solutions, the smaller of which seems to have yielded somewhat better quantizers in the examples that we have studied. Moreover, it even is not entirely clear at present whether the same solution should be used for all i or the smaller solution should be used for some values of i and the larger one for others.

VII. Uniform Quantizers

Study of (15), despite the myriad difficulties chronicled above, has proved to be rewarding. Perhaps the most surprising discovery was that, although (15) produced quantizers with rather nonuniform threshold separations $a_i - a_{i-1}$, in none of the cases we explored were their rates ever found to be more than 0.005 bits lower than those of uniform quantizers that achieved the same D. We knew that the optimum quantizers would tend toward uniformity in the limit of small D (large R), since Gish and Pierce [5] already had established that uniform quantizers are asymptotically optimum in this limit for all $r > 0$.[5] The unexpected phenomenon was that, even at moderate and large values of D, uniform quantizers had rates that for all intents and purposes were as low as those of the nonuniform quantizers that actually satisfied (15).[6] Moreover, in the special case of the exponential probability density $f(x) = \alpha \exp(-\alpha|x|)$, $x > 0$, and the Laplacian probability density $f(x) = (\alpha/2) \exp(-\alpha|x|)$, the computer solution indicated that quantizers with uniformly spaced thresholds $a_i = i\Delta$ satisfied (15) exactly. An analytical check immediately verified this fact. Further analysis then yielded the following parametric expression for the R versus D performance curve of the optimized quantizers in the exponential case:

$$R = (1-\theta)^{-1}[-\theta \log \theta - (1-\theta) \log (1-\theta)] \quad (16a)$$

$$D = \alpha^{-2}[1 - \theta(1-\theta)^{-2} \log^2 \theta]. \quad (16b)$$

As the parameter $\theta = e^{-\alpha\Delta}$ runs from 0 to 1, D runs from α^{-2} to 0 and R runs from 0 to ∞. In the Laplacian case, R is greater than in the exponential case by log 2, while D remains unchanged. It is of interest to note that in the limit of small D (i.e., $\theta \rightarrow 1$), the asymptotic behavior of (16) is $R \sim \log (e/\sqrt{12\alpha^2 D})$, whereas the absolute lower limit on all source-encoding systems set by the asymptotic behavior of the rate-distortion function $R(D) \sim \log \sqrt{e/2\pi\alpha^2 D}$ is only $\log \sqrt{\pi e/6} \approx 0.51$ bits lower [9, sect. 4.3.4].

Some appreciation for why uniform quantization is so good so generally can be gleaned from considering the following problem. Suppose that F possesses a density $f = F'$ that is piecewise constant, say $f(x) = c_k$ for all x in the interval I_k, $k = 1,2,\cdots$. Further suppose that our

[5] The Gish–Pierce result implies that uniform quantizers should satisfy (15) in the limit as the interthreshold width $\delta \rightarrow 0$. In this regard asymptotic analysis reveals that for $\lambda = 6/\delta^2$ the difference between the two sides of (15) for a uniform quantizer with threshold spacing δ vanishes like δ^5 at all points at which $F(\cdot)$ is twice differentiable.

[6] The near optimality of uniform quantization for all D had been observed previously by Wood [8] in the case of Gaussian signals, but the phenomenon apparently prevails quite generally.

task is to choose the number N_k of quantization bins to be assigned to I_k in such a way as to minimize $D = E|Y - X|^r$ subject to the requirement that the quantizer entropy may not exceed R. Although intuition may suggest that the density with which we should pack quantization levels into I_k should be an increasing function of c_k, the following analysis reveals that the same density of levels should be used everywhere. Since (15) is satisfied by uniformly spaced levels when X is uniformly distributed, the reconstruction levels of the bins assigned to I_k should be equally spaced within I_k. Hence, if we let L_k denote the length of I_k,

$$J \triangleq D + \mu R = \sum_k c_k L_k \left[\frac{L_k^{r+1}}{(r+1)2^r N_k^{r+1}} - \mu \log \frac{c_k L_k}{N_k}\right].$$

Setting $dJ/dN_k = 0$ yields $(L_k/N_k)^{r+1} = 2^r\mu$, so the width L_k/N_k of the quantization bins in I_k is independent of k. That is, the optimum quantizer has uniformly spaced thresholds. Of course, a truly uniform quantizer cannot be constructed if the L_k are incommensurate and can be constructed only for certain values of R and D if the L_k are commensurate. In the limit as $D \to 0$, however, the thresholds have to crowd together, so the uniform solution can be approximated as closely as desired even if the L_k are incommensurate. Since any $f(\cdot)$ can be expressed as the limit of a sequence of piecewise-constant density functions, the present analysis can be extended to provide an alternative derivation of the result of Gish and Pierce [5] that uniform quantization is asymptotically optimum as $D \to 0$ for arbitrary $f(\cdot)$ and arbitrary $r > 0$.

Appendix

Proof of (11)

We follow the approach outlined in the discussion embodying (12). Since the left side of (11) clearly is nonnegative, we can establish its finiteness by bounding it from above. For this purpose we employ the inequality [10]

$$|a + b|^r \le c_r|a|^r + c_r|b|^r, \qquad r > 0,$$

where $c_r = 1$ if $r \le 1$ and $c_r = 2^{r-1}$ if $r \ge 1$. It follows that

$$\begin{aligned} E&\left[n^{-1} \sum_i \sum_{j=S_{i-1}+1}^{S_i} |X_n^j - \gamma_i|^r\right] \\ &\le c_r E\left[n^{-1} \sum_i \sum_{j=S_{i-1}+1}^{S_i} |X_n^j|^r + |\gamma_i|^r\right] \\ &= c_r E\left[n^{-1} \sum_{j=1}^{n} |X_n^j|^r\right] + c_r n^{-1} \sum_i (S_i - S_{i-1})|\gamma_i|^r \\ &= c_r E\left[n^{-1} \sum_{j=1}^{n} |X_j|^r\right] + c_r n^{-1} \sum_i n_i|\gamma_i|^r \\ &= c_r E|X|^r + c_r \sum_i (n_i/n)|\gamma_i|^r. \end{aligned}$$

Now $E|X|^r < \infty$ by assumption, while (10b) implies that

$$\sum_i (n_i/n)|\gamma_i|^r \to \sum_i p_i|\gamma_i|^r = E|Y|^r,$$

where Y is the quantized version of X. Since

$$E|Y|^r \le c_r E|X|^r + c_r E|Y - X|^r = c_r E|X|^r + c_r D < \infty$$

the desired finiteness has been established.

It remains only to establish (12), which is of the form

$$n^{-1} \sum_i \sum_{j=S_{i-1}+1}^{S_i} |X_n^j - \gamma_i|^r \xrightarrow[\substack{\text{with}\\ \text{probability 1}}]{} E[|X - g(X)|^r], \quad \text{(A-1)}$$

where g is the quantizer function $g(x) = \gamma_i$, $a_{i-1} < x \le a_i$. Toward this end, introduce the empirical cumulative distribution functions F_n, $n = 1,2,\cdots$, according to the usual definition

$$F_n(x) = N(x)/n, \quad \text{(A-2)}$$

where $N(x)$ is the number of values of j between 1 and n inclusive for which $X_j \le x$. Let F_i and $F_{n,i}$ denote $F(a_i)$ and $F_n(a_i)$, respectively. Also, select the parameters $\{n_i\}$ (equivalently $\{S_i\}$) of the permutation code B_n according to the prescription $S_i = [nF_i]$, where $[y]$ denotes the integral part of y. Note that this choice of the $\{S_i\}$ is consistent with (10a). Next write

$$\begin{aligned} n^{-1} &\sum_{j=S_{i-1}+1}^{S_i} |X_n^j - \gamma_i|^r \\ &= n^{-1} \sum_{j=[nF_{i-1}]+1}^{nF_{n,i-1}} |X_n^j - \gamma_i|^r + n^{-1} \sum_{j=nF_{n,i-1}+1}^{nF_{n,i}} |X_n^j - \gamma_i|^r \\ &\quad + n^{-1} \sum_{j=nF_{n,i}+1}^{[nF_i]} |X_n^j - \gamma_i|^r. \end{aligned} \quad \text{(A-3)}$$

Since $F_{n,i} \to F_i$ with probability 1 by the Borel strong law, the number of terms in the first and third sums on the right side of (A-3) is $o(n)$ with probability 1 for each i. Were the terms bounded, it would follow that both of these terms approach 0 with probability 1. The potential difficulty stemming from the unboundedness of $|X_n^j - \gamma_i|^r$ is circumvented easily, however. The third sum, for example, is devoid of terms with probability 1 if $F_i = 0$ or 1. If $0 < F_i < 1$, then we can find a finite $\delta > 0$ such that $F(a_i - \delta) < F_i < F(a_i + \delta)$. The Borel strong law then implies that $X_n^{[nF_i]} \in (a_i - \delta, a_i + \delta)$ for n sufficiently large with probability 1. It follows therefrom that $X_n^j \in (a_i - \delta, a_i + \delta)$ for all j between $nF_{n,i} + 1$ and $[nF_i]$ for all but finitely many n with probability 1. Since the number of such j is $o(n)$ with probability 1 and $|X_n^j - \gamma_i|^r \le |a_i + \delta - \gamma_i|^r < \infty$ for each of them, the third term approaches 0 with probability 1. Similar arguments imply that the first term approaches 0 with probability 1. The task of establishing (A-1) therefore has been reduced to showing that

$$n^{-1} \sum_i \sum_{j=nF_{n,i-1}+1}^{nF_{n,i}} |X_n^j - \gamma_i|^r \xrightarrow[\substack{\text{with}\\ \text{probability 1}}]{} E[|X - g(X)|^r]. \quad \text{(A-4)}$$

From the definition of $F_{n,i}$, we know that $X_n^j \in (a_{i-1},a_i]$ for $nF_{n,i-1} + 1 \le j \le nF_{n,i}$, so an alternative way of expressing (A-4) is

$$n^{-1} \sum_i \sum_{j=nF_{n,i-1}+1}^{nF_{n,i}} |X_n^j - g(X_n^j)|^r \xrightarrow[\substack{\text{with}\\ \text{probability 1}}]{} E[|X - g(X)|^r]. \quad \text{(A-5)}$$

Since each value of $j = 1,\cdots,n$ appears in one and only one of the ranges $nF_{n,i-1} + 1 \le j \le nF_{n,i}$, (A-5) reduces to

$$n^{-1} \sum_{j=1}^{n} |X_j - g(X_j)|^r \xrightarrow[\text{probability 1}]{\text{with}} E[|X - g(X)|^r]$$

the validity of which is a direct consequence of the pointwise ergodic theorem.

References

[1] T. J. Goblick, Jr., 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.

[2] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 490–501, May 1968.

[3] J. G. Dunn, "The performance of a class of *n*-dimensional quantizers for a Gaussian source," in *Proc. Columbia Symp. Signal Transmission and Processing*, Columbia Univ., New York, N.Y., pp. 76–81, May 1965.

[4] T. Berger, F. Jelinek, and J. K. Wolf, "Permutation codes for sources," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 160–169, Jan. 1972.

[5] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676–683, Sept. 1968.

[6] C. A. R. Hoare, "Quicksort," *Comput. J.*, vol. 5, pp. 10–15, 1962.

[7] M. Blum, "Computational complexity," presented at the 1972 IEEE Int. Symp. Information Theory, Asilomar, Calif., Jan. 31–Feb. 3, 1972.

[8] R. C. Wood, "On optimum quantization," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 248–252, Mar. 1969.

[9] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

[10] M. Loève, *Probability Theory*, 3rd ed. Princeton, N.J.: Van Nostrand, 1963, p. 155.

On Variable-Length-to-Block Coding

FREDERICK JELINEK, SENIOR MEMBER, IEEE, AND KENNETH S. SCHNEIDER, MEMBER, IEEE

***Abstract*—Variable-length-to-block codes are a generalization of run-length codes. A coding theorem is first proved. When the codes are used to transmit information from fixed-rate sources through fixed-rate noiseless channels, buffer overflow results. The latter phenomenon is an important consideration in the retrieval of compressed data from storage. The probability of buffer overflow decreases exponentially with buffer length and we determine the relation between rate and exponent size for memoryless sources. We obtain codes that maximize the overflow exponent for any given transmission rate exceeding the source entropy and present asymptotically optimal coding algorithms whose complexity grows linearly with codeword length. It turns out that the optimum error exponents of variable-length-to-block coding are identical with those of block-to-variable-length coding and are related in an interesting way to Renyi's generalized entropy function.**

I. Introduction

ENCODING of variable-length sequences of source outputs into codewords of constant length is called *variable-length-to-block coding*. It can be considered a generalization of run-length encoding [3] and is a technique of data compression that seems especially attractive for a skew source (where the frequency of some output letters very much exceeds that of others) or for retrieval situations that require block formatting of data. Variable-length-to-block coding was recently considered by Tunstall [7] who described an encoding construction and proved it optimal in a certain sense (see Section III).

In this paper we will apply variable-length-to-block coding to fixed-rate sources and channels. We will be concerned with the problems analyzed by Jelinek [2] for block-to-variable-length encoding: buffer overflow, construction of optimal codeword sets, and coding theorems. The overflow problem is important in real-time transmission of quantized data that are then encoded to minimize the overall rate. Gish and Pierce [9] have shown that when this approach is applied to Gaussian data, its performance is close to the rate-distortion optimum.

It will be shown in the Appendix that the fixed-rate source and channel concept can also serve as a model of an important problem in fast retrieval from storage of encoded (compressed) data. Thus, the applicability of buffer overflow results is not limited to communication situations.

Let us begin by considering Fig. 1, which consists of three objects; a constant memoryless source (henceforth abbreviated CMS), a fixed-rate noiseless channel (henceforth abbreviated FRC), and a user. The CMS emits digits z in the c-ary alphabet $J_c = (0,1,\cdots,c-1)$ at the rate of one every second. These are independent and identically distributed random variables under the common probability distribution $\{Q(z)\}$. (The convention will be adopted that $Q(0) \leq Q(1) \leq \cdots \leq Q(c-1)$.) The FRC can accept digits in the d-ary alphabet $J_d = (0,1,\cdots,d-1)$ at its input and transmit them to its output without error. However, the channel can only accept digits for transmission at the rate of one every $(\log d)/R$ seconds. The parameter R is called the channel rate. Finally, the user is interested in learning the outputs of the CMS.

The task of the communication engineer is to employ the FRC as a link by which the user may learn the outputs of

Manuscript received September 15, 1970; revised February 6, 1972. This work was supported in part by NASA Contract NAS 2-5643.

F. Jelinek is with the School of Electrical Engineering, Cornell University, Ithaca, N.Y. 14850.

K. Schneider was with the School of Electrical Engineering, Cornell University, Ithaca, N.Y. 14850. He is now with the M.I.T. Lincoln Laboratory, Lexington, Mass.