Lossy Source Coding

Toby Berger, *Fellow, IEEE*, and Jerry D. Gibson, *Fellow, IEEE*

(Invited Paper)

Abstract— **Lossy coding of speech, high-quality audio, still images, and video is commonplace today. However, in 1948, few lossy compression systems were in service. Shannon introduced and developed the theory of source coding with a fidelity criterion, also called rate-distortion theory. For the first 25 years of its existence, rate-distortion theory had relatively little impact on the methods and systems actually used to compress real sources. Today, however, rate-distortion theoretic concepts are an important component of many lossy compression techniques and standards. We chronicle the development of rate-distortion theory and provide an overview of its influence on the practice of lossy source coding.**

Index Terms—**Data compression, image coding, speech coding, rate distortion theory, signal coding, source coding with a fidelity criterion, video coding.**

I. Introduction and Prologue

THE concept of specifying the rate required to represent a source with some less-than-perfect fidelity was introduced by Shannon in his landmark 1948 paper. In Part V thereof, Shannon describes the idea of “continuous information” and defines “The Rate for a Source Relative to a Fidelity Evaluation.” Furthermore, he states the first theorem concerning such lossy representations (his Theorem 21) and outlines its proof via an AEP-like argument. Shannon then writes the expression for the rate for the desired “valuation” (distortion) and poses the constrained optimization problem to be solved for the transition probabilities. Then he gives a general form of the solution to this optimization problem (now widely called the backward test channel), and specializes it to the important special case of difference distortion measures. In Theorem 22 he gives the exact rate for an ideal bandlimited Gaussian source relative to a mean-squared error (MSE) fidelity criterion, and in Theorem 23 he bounds the MSE information rate of a bandlimited non-Gaussian source in terms of now-classic expressions involving the source power and the entropy rate power. A most auspicious beginning indeed!

In 1948, although pulse-code modulation (PCM) was being developed for speech applications [259] and Dudley’s vocoder had been around for about ten years [260], actual implementations of lossy digital compression systems were nonexistent. This testifies to the power of Shannon’s insights but also helps explain why he would delay further consideration of lossy compression until 10 years later. By 1959, work in scalar quantization and PCM was well underway [196] and differential encoding had received considerable attention [180], [186], [215].

Shannon coined the term “rate-distortion function” when he revisited the source-coding problem in 1959 [2]. The insights and contributions in that paper are stunning. In particular, rate-distortion terminology is introduced, the rate-distortion function $R(D)$ is carefully defined, positive and negative coding theorems are proved, properties of $R(D)$ are investigated, the expression for $R(D)$ in several important cases is derived, some numerical examples are presented, the important lower bound to $R(D)$, now called the Shannon lower bound, is derived, and the duality between $R(D)$ and a capacity cost function is noted. A lifetime of results in two papers!

We treat Shannon’s seminal contributions in greater detail below, also emphasizing how they inspired others to begin making significant contributions both to rate-distortion theory and to laying the groundwork for advances in the practice of lossy source coding. Specifically, we survey the history and significant results of rate-distortion theory and its impact on the development of lossy source-compression methods. A historical overview of rate-distortion theory is presented in the first part of the paper. This is followed by a discussion of techniques for lossy coding of speech, high-quality audio, still images, and video. The latter part of the paper is not intended as a comprehensive survey of compression methods and standards. Rather, its emphasis is on the influence of rate-distortion theory on the practice of lossy source coding.

There is both logic and historical precedent for separating the treatment of lossy source coding into a theory component and a practice component. Davisson and Gray took this approach in the Introduction of their 1976 compilation of papers on Data Compression [183]. Additionally, there is a continuity in the development of rate-distortion theory and, similarly but separately, in the development of the practice of lossy source coding. These continuities deserve preservation, since appreciation for research and development insights is enhanced when they are embedded in their proper historical contexts.

II. In the Beginning

Shannon’s [1] motivations for writing “Section V: The Rate for a Continuous Source” likely included the following:

1) It provided the source coding complement to his treatment of the input-power limited AWGN channel.

Manuscript received December 1, 1997; revised July 1, 1998. This work was supported in part by NSF under Grants ECS-9632266 and NCR-9796255.

T. Berger is with the School of Electrical Engineering, Cornell University, Ithaca, NY 14853 USA.

J. D. Gibson is with the Department of Electrical Engineering, Southern Methodist University, Dallas, TX 75275 USA.

Publisher Item Identifier S 0018-9448(98)06886-2.

2) It provided the means by which to extend information theory to analog sources. Such an extension was necessary because all analog sources have infinite entropy by virtue of their amplitude continuity and, therefore, cannot be preserved error-free when stored in or transmitted through practical, finite-capacity media.
3) Shannon considered the results to have inherent significance independent of their analogies to and connections with channel theory.

A. A Brief Detour into Channel Theory

Shannon's most widely known and most widely abused result is his formula for the capacity of an ideal bandlimited channel with an average input power constraint and an impairment of additive, zero-mean, white Gaussian noise, namely,

$$C = W \log_2 (1 + P/N) \text{ bits/s.} \tag{1}$$

Here, P is the prescribed limitation on the average input power, W is the channel bandwidth in positive frequencies measured in hertz, and N is the power of the additive noise. Since the noise is white with one-sided power spectral density N_0 or two-sided power spectral density $N_0/2$, we have $N = N_0 W$. Of course, the result does not really require that the noise be truly white, just that its spectral density be constant over the channel's passband. Common abuses consist of applying (1) when

i) The noise is non-Gaussian.
ii) The noise is not independent of the signal and/or is not additive.
iii) Average power is not the (only) quantity that is constrained at the channel input.
iv) The noise is not white across the passband and/or the channel transfer function is not ideally bandlimited.

Abuse i) is conservative in that it underestimates capacity because Gaussian noise is the hardest additive noise to combat. Abuse ii) may lead to grossly underestimating or grossly overestimating capacity. A common instance of abuse iii) consists of failing to appreciate that it actually may be peak input power, or perhaps both peak and average input power, that are constrained. Abuse iv) leads to an avoidable error in that the so-called "water pouring" result [3], generalizing (1), yields the exact answer when the noise is not white, the channel is not bandlimited, and/or the channel's transfer function is not flat across the band. (See also [6] and [7].)

B. Coding for Continuous Amplitude Sources

There is a pervasive analogy between source-coding theory and channel-coding theory. The source-coding result that corresponds to (1) is

$$R = W \log_2 (S/N) \text{ bits/s.} \tag{2}$$

It applies to situations in which the data source of interest is a white Gaussian signal bandlimited to $|f| \le W$ that has power $S = S_0 W$, where S_0 denotes the signal's one-sided constant power spectral density for frequencies less than W. The symbol, N, although often referred to as a "noise," is actually an estimation error. It represents a specified level of mean-squared error (MSE) between the signal $\{X(t)\}$ and an estimate $\{\hat{X}(t)\}$ of the signal constructed on the basis of data about $\{X(t)\}$ provided at a rate of R bits per second. That is,

$$N = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} dt E[\hat{X}(t) - X(t)]^2.$$

It was, and remains, popular to express MSE estimation accuracy as a "signal-to-noise ratio," S/N, as Shannon did in (2). It must be appreciated, however, that $\{\hat{X}(t) - X(t)\}$ is not noise in the sense of being an error process that is independent of $\{X(t)\}$ that nature adds to the signal of interest. Rather, it is a carefully contrived error signal, usually dependent on $\{X(t)\}$, that the information theorist endeavors to create in order to conform to a requirement that no more than R bits per second of information may be supplied about $\{X(t)\}$. In modern treatises on information theory, the symbol "D," a mnemonic for average distortion, usually is used in place of N. This results in an alternative form of (2), namely,

$$R(D) = W \log_2 (S/D) \text{ bits/s} \tag{3}$$

which is referred to as the MSE rate-distortion function of the source.

Formula (3) gets abused less widely than formula (1), but probably only because it is less widely known. Abuses consist, analogously, of applying it to situations in which

i) The signal is non-Gaussian.
ii) Distortion does not depend simply on the difference of $\hat{X}(t)$ and $X(t)$.
iii) Distortion is measured by a function of $\hat{X}(t) - X(t)$ other than its square.
iv) The signal's spectral density is not flat across the band.

Again, abuse i) is conservative in that it results in an overestimate of the minimum rate R needed to achieve a specified MSE estimation accuracy because white Gaussian sources are the most difficult to handle in the sense of bit rate versus MSE. Abuses ii) and iii), which often stem in practice from lack of knowledge of a perceptually appropriate distortion measure, can result in gross underestimates or overestimates of R. Abuse iv) can be avoided by using a water-pouring generalization of (3) that we shall discuss subsequently. In anticipation of that discussion, we recast (3) in the form

$$R(D) = \max [0, W \log(S_0 W/D)]. \tag{4}$$

This form of the equation explicitly reflects the facts that i) the signal spectrum has been assumed to be constant at level S_0 across the band of width W in which it in nonzero, and ii) $R(D) = 0$ for $D \ge S_0 W$, because one can achieve an MSE of $S = S_0 W$ without sending any information simply by guessing that $X(t) = 0$. (If $\{X(t)\}$ has a nonzero mean $m(t)$, then of course one guesses $m(t)$ instead of zero. In general, adding a deterministic mean-value function to the signal process does not change its rate-distortion function with respect to any fidelity criterion that measures average distortion as some functional of the difference process $\{\hat{X}(t) - X(t)\}$.) The base of the logarithm in (4) determines the information unit—bits for $\log_2$ and nats for $\log_e$. When we deal with continuously distributed quantities, it is more "natural" to employ natural logs. When no log base appears, assume that a natural log is intended.

C. Deterministic Processes Have Nonzero Rate-Distortion Functions

It is appropriate at this juncture to comment on the relationship between rate-distortion theory and the theory of deterministic processes. The Wold decomposition theorem assures us, among other things, that any bandlimited random process is deterministic in the following sense: it can be predicted with zero MSE infinitely far into the future and infinitely far into the past once one knows the values it has assumed in an arbitrarily small open interval. This is because the sample paths of such processes are analytic functions with probability one, which implies that they have derivatives of all orders at every instant. Knowledge of the process over an arbitrarily small open interval allows each such derivative to be computed with perfect accuracy at any point within the interval by taking the limit of an appropriate difference quotient. This, in turn, permits using Taylor series or other techniques to extrapolate the process with perfect accuracy into the arbitrarily remote past and future. This suggests that the ideal bandlimited Gaussian process we have been studying should have an MSE rate-distortion function that is identically zero for all D because one needs to supply information about the process only during an arbitrarily short interval, after which it becomes known perfectly for all time. Yet, Shannon's formula $R(D) = W \log(S/D)$ says that one must keep supplying information about it for all time at a rate of $R(D) > 0$ in order to be able to reconstruct it with a MSE of $D < S$.

This apparent contradiction is readily resolved. The sticking point is that it requires an infinite amount of information to specify even a single continuously distributed random variable exactly, let alone the uncountable infinity of them indexed by all the points in an open interval. Accordingly, when information is provided at a finite rate, which is always the case in practice, one never learns the values in any interval perfectly no matter how long one gathers information about them. Determinism in the above sense thus is seen to be a purely mathematical concept that is devoid of practical significance. The operative, physically meaningful measure of the rate at which a random process, even a so-called deterministic random process, produces information subject to a fidelity criterion is prescribed by Shannon's rate-distortion theory.

D. The Basic Inequality

A basic inequality of information theory is

$$D \geq R^{-1}(C) \tag{5}$$

sometimes referred to as the *information transmission inequality*. It says that, if you are trying to transmit data from a source with rate-distortion function $R(D)$ over a channel of capacity C, you can achieve only those average distortions that exceed the inverse of the rate-distortion function evaluated at C. (Not surprisingly, the inverse rate-distortion function is called the distortion-rate function.)

Suppose, for example, that we wish to send data about the aforementioned bandlimited white Gaussian process $\{X(t)\}$ over an average-input-power-limited, ideally bandlimited AWGN channel. Assume our task is to construct on the basis of what we receive at the channel output an approximation $\{\hat{X}(t)\}$ that has the least possible MSE. The source and the channel have the same frequency band $|f| \leq W$. Since $R(D) = W \log_2(S/D)$, the distortion-rate function is

$$D(R) = S2^{-R/W}$$

so (1) and (4) together tell us that

$$D \geq D(C) = S \exp\left[-\frac{W \log(1 + P/N)}{W}\right]$$

or

$$D/S \geq (1 + P/N)^{-1}. \tag{6}$$

This tells us that the achievable error power per unit of source power (i.e., the achievable normalized MSE) is bounded from below by the reciprocal of one plus the channel signal-to-noise ratio (SNR).

E. An Optimum System via a Double Coincidence

There happens to be a trivial scheme for achieving equality in (6) when faced with the task of communicating the source of Section II-B over the channel of Section II-A. It consists of the following steps:

Step 1: Tranmit $X(t)$ scaled to have average power P; that is, put $\sqrt{P/S}X(t)$ into the channel.

Step 2: Set $\hat{X}(t)$ equal to the minimum mean-square error (MMSE) estimate of $X(t)$ based solely on the instantaneous channel output $\sqrt{P/S}X(t) + N(t)$ at time t.

Since the signal and the channel noise are jointly Gaussian and zero mean, the optimum estimate in Step 2 is simply a linear scaling of the received signal, namely,

$$\hat{X}(t) = \alpha[\sqrt{P/S}X(t) + N(t)].$$

The optimum α is found from the requirement that the error of the optimum estimator must be orthogonal to the data

$$E[\hat{X}(t) - X(t)][\sqrt{P/S}X(t) + N(t)] = 0.$$

This may be written as

$$E[(\alpha\sqrt{P/S} - 1)X(t) + \alpha N(t)][\sqrt{P/S}X(t) + N(t)] = 0.$$

Using $ES^2(t) = S$, $EN^2(t) = N$, and $EX(t)N(t) = 0$, we obtain $\alpha = \sqrt{PS}/(P + N)$. The resulting minimized normalized MSE is easily computed to be

$$D/S = (1 + P/N)^{-1} \tag{7}$$

which means we have achieved equality in (6).

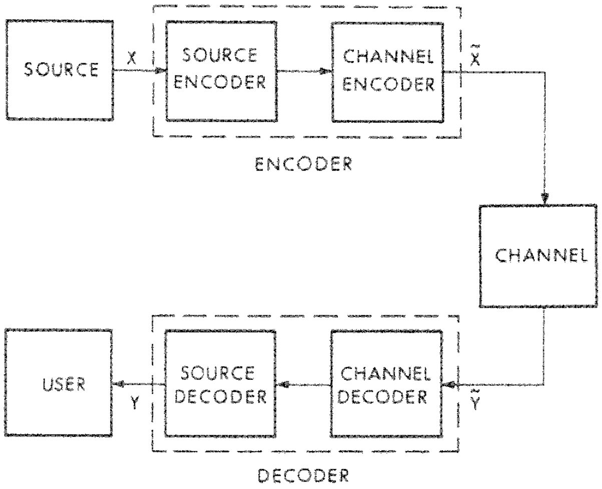

Fig. 1.

Thus the simple two-step scheme of instantaneously scaling at the input and at the output results in an end-to-end communication system that is optimum. No amount of source and/or channel coding could improve upon this in the MSE sense for the problem at hand. This fortuitous circumstance is attributable to a double coincidence. The first coincidence is that the source happens to be the random process that drives the channel at capacity. This is, the given source, scaled by $\sqrt{P/S}$, is that process of average power not exceeding P which maximizes the mutual information between the input and output of the channel. The second coincidence is that the channel just happens to provide precisely the transition probabilities that solve the MSE rate-distortion problem for the given source. That is, when the channel is driven by the scaled source, its output minimizes mutual information rate with the source over all processes from which one can calculate an approximation to the source that achieves a normalized MSE not in excess of $(1+P/N)^{-1}$.

We are operating at a saddle point at which the mutual information rate is simultaneously maximized subject to the average power constraint and minimized subject to the average distortion constraint. The slightest perturbation in any aspect of the problem throws us out this saddle—unequal source and channel bandwidths, non-Gaussianness of the source or channel, an error criterion other than MSE, and so on. The result of any such perturbation is that, in order to recover optimality, it is in general necessary to code both for the source and for the channel as depicted in Fig. 1.

The source encoder and channel decoder usually have to implement complicated many-to-one mappings that depend on the values their inputs assume over long durations, not just at one instant. Hence, whereas a surface perusal of Shannon's founding treatise [1] might, via the key formulas discussed above, instill the illusion that all one ever has to do to build an optimum communication system is simply to insert into the channel a version of the given source trivially accommodated to whatever channel input constraints may prevail, nothing could be further from the truth. The goals of this tutorial paper include exorcising any such misconception and surveying the major developments in rate-distortion theory over the fifty years from 1948 to 1998.

III. The Fifties

A. The Russian School

From 1949 to 1958 no research was reported on rate-distortion theory in the United States or Europe. However, there was a stream of activity during the 1950's at Moscow University by members of Academician A. N. Kolmogorov's probability seminar. Kolmogorov, a renowned mathematician who founded axiomatic probability theory and contributed many of its fundamental limit laws, saw an application for Shannon's information theory in the long-standing isomorphism problem of ergodic theory. That problem concerns necessary and sufficient conditions for when two "shifts" can be placed in a one-to-one, measure-preserving correspondence. It includes as an important special case the question of whether or not two given random processes, or sources, can be viewed as perhaps intricately disguised rearrangements of the same information stream. Shannon's theory showed that each discrete-amplitude information source has an entropy rate H that measures in a fundamental way the rate at which it produces information, and that any two sources of the same entropy rate can be "coded" into one another losslessly. Thus entropy (more exactly, entropy rate) emerged as a promising candidate to serve as the long-sought invariant in the isomorphism problem. However, "coding" in Shannon theory differs from "coding" in ergodic theory. Shannon's coding concerns operations on possibly long but always finite blocks of information, thereby honoring a tie to practice, whereas the codes of ergodic theory operate on the entirety of each infinite sequence that constitutes a realization of an ergodic flow. Thus it was not a trivial matter to establish that entropy rate could indeed serve as an invariant in the sense of ergodic theory. Kolmogorov and Sinai [5], [8] succeeded in showing that equal entropy rates were a necessary condition for isomorphism. Years later, Ornstein [9] proved sufficiency within an appropriately defined broad class of random stationary processes comprising all finite-order Markov sources and their closure in a certain metric space that will not concern us here. With the Moscow probability seminar's attention thus turned to information theory, it is not surprising that some of its members also studied Section V, The Rate for a Continuous Source. Pinsker, Dobrushin, Iaglom, Tikhomirov, Oseeyevich, Erokhin, and others made contributions to a subject that has come to be called ϵ-entropy, a branch of mathematics that subsumes what we today call rate-distortion theory. ϵ-entropy is concerned with the minimal cardinality of covers of certain spaces by disks of radius ϵ. As such, it is a part of topology if a complete cover is desired. If, however, a probability measure is placed on the space being covered, then one can consider covering all but a set of measure δ, where $\delta = 0$ is also a value of considerable significance [10], [11], [1, p. 656]. It also becomes interesting to consider the expected distance from a point in the space to the closest disk center, which is the approach usually adopted in rate-distortion theory.

Most of the attention of scholars of ϵ-entropy was focused on the asymptotic case in which $\epsilon \to 0$. This doubtless accounts for why the symbol ϵ was selected instead of, say, D for distortion. It was appreciated that ϵ-entropy would diverge as $\epsilon \to 0$ in all continuous-amplitude scenarios. The problem was to determine the rate of divergence in particular cases of interest. Thus when invited to address an early information theory symposium, Kolmogorov [3] emphasized in the portion of his report dealing with ϵ-entropy Iaglom's expression for the limiting information rate of Wiener processes as $\epsilon \to 0$ and extensions thereof to more general diffusions. However, he also reported the exact answer for the ϵ-entropy of a stationary Gaussian process with respect to the squared L_2-norm for *all* ϵ, not just $\epsilon \to 0$ (his equations (17) and (18)). That result, and its counterpart for the capacity of a power-constrained channel with additive colored Gaussian noise, have come to be known as the "water-pouring" formulas of information theory. In this generality the channel formula is attributable to [12] and the source formula to Pinsker [13], [14]. We shall call them the Shannon–Kolmogorov–Pinsker (SKP) water-pouring formulas. They generalize the formulas given by Shannon in 1948 for the important case in which the spectrum of the source or of the channel noise is flat across a band and zero elsewhere. The water-pouring formulas were rediscovered independently by several investigators throughout the 1950's and 1960's.

B. The Water Table

Here is a simple way of obtaining the SKP water-pouring formula for the MSE information rate of a Gaussian source [12]. The spectral representation theorem lets us write any zero-mean stationary random process $\{X(t)\}$ for which $EX(t)^2 < \infty$ in the form

$$X(t) = \int_{-\infty}^{\infty} e^{itf}\, d\xi(f)$$

where $\xi(f)$ is a random process with zero mean, uncorrelated increments. Hence, if A and B are two disjoint sets of frequencies, the zero-mean random processes $\{X_A(t)\}$ and $\{X_B(t)\}$ defined by

$$X_A(t) = \int_A e^{itf} d\xi(f)$$

and

$$X_B(s) = \int_B e^{isg} d\xi(g)$$

satisfy $EX_A(t)X_B(t) = 0$ because $Ed\xi(f)d\xi(g) = 0$ when $f \in A$ and $g \in B$. That is, processes formed by bandlimiting a second-order stationary random processes to nonoverlapping frequency bands are uncorrelated with one another. In the case of a Gaussian process, this uncorrelatedness implies independence. Thus we can decompose a Gaussian process $\{X(t)\}$ with one-sided spectral density $S(f)$ into independent Gaussian processes $\{X_i(t)\}, i = 0, 1, \cdots$ with respective spectral densities $S_i(f)$ given by

$$S_i(f) = \begin{cases} S(f), & \text{if } i\Delta \le f < (i+1)\Delta \\ 0, & \text{otherwise.} \end{cases}$$

Let us now make Δ sufficiently small that $S_i(f)$ becomes effectively constant over the frequency interval in which it is nonzero, $S_i(f) \approx S_i, i\Delta \le f < (i+1)\Delta$.[1] Since the subprocesses $\{X_i(t)\}$ are independent of one another, it is best to approximate each of them independently. Moreover, given any such set of independent approximants, simply summing them yields the best MSE approximation of $\{X(t)\}$ that can be formed from them, the MSE of said sum being the sum of the MSE's of the subprocess approximants. Furthermore, the source-coding rate will be the sum of the rates used to approximate the subprocesses.

Subprocess $\{X_0(t)\}$ is an ideal bandlimited zero-mean Gaussian source with bandwidth Δ and spectral density $S(f) = S_0, 0 \le f < \Delta$. It follows from (4) that the minimum information rate needed to describe it with an MSE of D_0 or less is

$$R_0(D_0) = \max\,[0, \Delta \log\,(S_0\Delta/D_0)].$$

Subprocess $\{X_i(t)\}$ for any $i > 0$ also is a bandlimited zero-mean Gaussian source with bandwidth Δ in positive frequencies, its frequency band being $[i\Delta, (i+1)\Delta)$ instead of $[0, \Delta)$. Consider any coded representation of it with rate R_i bits per second from which one can produce an approximation of it that has an MSE of D_i. Observe that we always can mimic this (R_i, D_i)-performance by by mixing down to baseband $[0, \Delta)$, performing the same coding and reconstruction operations on the result, and then mixing the approximation thus produced back into the band $[i\Delta, (i{+}1)\Delta)$. It follows that the best rate–distortion tradeoff we can achieve for subprocess $\{X_i(t)\}$ is

$$R_i(D_i) = \max\,[0, \Delta \log\,(S_i\Delta/D_i)].$$

By additively combining said approximations to all the subprocesses, we get an approximation to $\{X(t)\}$ that achieves an average distortion of

$$D = \sum_i D_i$$

and requires a total coding rate of

$$R = \sum_i R_i(D_i) = \sum_i \max\,[0, \Delta \log\,(S_i\Delta/D_i)].$$

In order to determine the MSE rate-distortion function of $\{X(t)\}$, it remains only to select those D_i's summing to D which minimize this R. Toward that end we set

$$d(R + \lambda^{-1}D)/dD_i = 0, \qquad i = 0, 1, 2, \cdots$$

where λ is a Lagrange multiplier subsequently selected to achieve a desired value of D or of R. Each D_i of course never exceeds $S_i\Delta$, the value that can be achieved by sending no information about $\{X_i(t)\}$ and then using $\hat{X}_i(t) = 0$ as the approximant. If the solution associated with a particular value

[1] There is some sacrifice of rigor here. Readers desirous of a careful derivation based on the Kac–Murdock–Szego theory of the asymptotic distribution of the eigenvalues of Toeplitz forms may consult Berger [26].

λ of the Lagrange multiplier is such that $D_i < S_i\Delta$, then the preceding equation requires that $-\Delta/D_i + \lambda^{-1} = 0$, or

$$D_i = \lambda\Delta.$$

The value $\lambda = 0$ corresponds to $D_i = 0$ for all i (hence, $D = 0$) and $R = \infty$. This expresses that fact that perfect reconstruction of a continuously distributed source cannot be achieved without infinite data rate, a result that is mathematically satisfying but devoid of physical usefulness. For finite values of λ, we deduce that

$$D_i = \begin{cases} \lambda\Delta, & \text{if } \lambda < S_i \\ S_i\Delta, & \text{if } \lambda \geq S_i. \end{cases}$$

It follows that the D and R values associated with parameter value λ are

$$\begin{aligned} D_\lambda &= \sum_{\{i:S_i>\lambda\}} \lambda\Delta + \sum_{\{i:S_i\leq\lambda\}} S_i\Delta \\ &= \sum_i \Delta \min(\lambda, S_i) \end{aligned}$$

and

$$R_\lambda = \sum_i \max[0, \Delta \log(S_i/\lambda)].$$

We remark that the Lagrange solution tells us that to compute a point $(D, R(D))$ on the MSE rate-distortion function of $\{X(t)\}$, we should combine points on the rate-distortion functions $R_i(\cdot)$ of the subprocesses at points at which the slope $R_i'(\cdot)$ is the same for all i. That is, $R_i'(D_i)$ does not vary with i. This is a recurrent theme in rate-distortion theory. Constant slope means that the same marginal tradeoff is being drawn between rate and distortion for each of the independent components. Indeed, intuition suggests that this must be the case; otherwise it would be possible to lower the overall R for fixed D by devoting more bits to subprocesses being reproduced at points of lower slope and fewer bits to processes being reproduced at points of slope. In this connection the reader should observe that the slope of $R_i(D)$ is continuous everywhere except at $D = S_i\Delta$, where it jumps from $-1/S_i$ to 0. Hence, one can draw a tangent line to $R_i(\cdot)$ at $D = S_i\Delta$ with any slope between $-1/S_i$ and 0. For purposes of combining points in the sense of this paragraph, $R_i(\cdot)$ should be considered to have all slopes between $-1/S_i$ and 0 at $D = S_i\Delta$.

As $\Delta \to 0$ the above sums constituting our parametric representation of $R(D)$ become integrals over frequency, namely,

$$D_\lambda = \int_0^\infty \min[\lambda, S(f)]\, df$$

and

$$R_\lambda = \int_0^\infty \max[0, \log(S(f)/\lambda)]\, df.$$

Two-sided spectral densities with their attendant negative frequencies are less forbidding to engineers and scientists today than they were in the 1940's. Accordingly, the above result now usually is cast in terms of the two-sided spectral density $\Phi(f)$, an even function of frequency satisfying $\Phi(f) = S(f)/2, f \geq 0$. Replacing the parameter λ by $\theta = \lambda/2$, we find that

$$D_\theta = \int_{-\infty}^{\infty} \min[\theta, \Phi(f)]\, df \tag{8}$$

$$R_\theta = \int_{-\infty}^{\infty} \max\left[0, \frac{1}{2}\log(\Phi(f)/\theta)\right] df. \tag{9}$$

Some practitioners prefer to use angular frequency $\omega = 2\pi f$ as the argument of $\Phi(\cdot)$; of course, df then gets replaced in (8) and (9) by $d\omega/(2\pi)$.

The parametric representation (8) of the MSE rate-distortion function of a stationary Gaussian source is the source-coding analog of the SKP "water-pouring" result for the capacity of an input-power-limited channel with additive stationary Gaussian noise. The source-coding result actually is better described in terms of a "water table," though people nonetheless usually refer to it as "water pouring." Specifically, in Fig. 2, the source's spectral density is shown as a heavy "mold" resting atop a reservoir. In those places where there is air between the surface of the water and the mold, the surface is at uniform height θ; elsewhere, the mold presses down to a depth lower than θ. The water height $\min[\theta, \Phi(f)]$ is the MSE distortion as a function of frequency. Equivalently, at each frequency the amount, if any, by which the height of the mold exceeds the water level, namely $\max[\Phi(f) - \theta, 0]$, is the portion of the signal power at that frequency that is preserved by the minimum-rate data stream based from which the source can be reconstructed with average distortion D_θ.

Equations (8) and (9) also specify the MSE rate-distortion function of a time-discrete Gaussian sequence provided we limit the range of integration to $|f| \leq 1/2$ or to $|\omega| \leq \pi$. In such cases, $\Phi(\omega)$ is the discrete-time power spectral density, a periodic function defined by

$$\Phi(\omega) = \sum_{k=-\infty}^{\infty} \phi(k) \exp(j\omega k)$$

where $\phi(k) = EX_jX_{j\pm k}$ is the correlation function of the source data. Note that when the parameter θ assumes a value less than the minimum[2] of $\Phi(\cdot)$, which minimum we shall denote by D^*, (8a) reduces to $D_\theta = \theta$, which eliminates the parameter and yields the explicit expression

$$R(D) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log[\Phi(\omega)/D]\, d\omega, \qquad D \leq D^*.$$

This may be recast in the form

$$R(D) = \frac{1}{2}\log(Q_0/D), \qquad D \leq D^*$$

where

$$Q_0 = \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\Phi(\omega)\, d\omega\right]$$

is known in the information theory literature as the *entropy rate power* of $\{X_k\}$. We shall return to this result when discussing the literature of the 1960's.

[2] More precisely, less than the essential infimum.

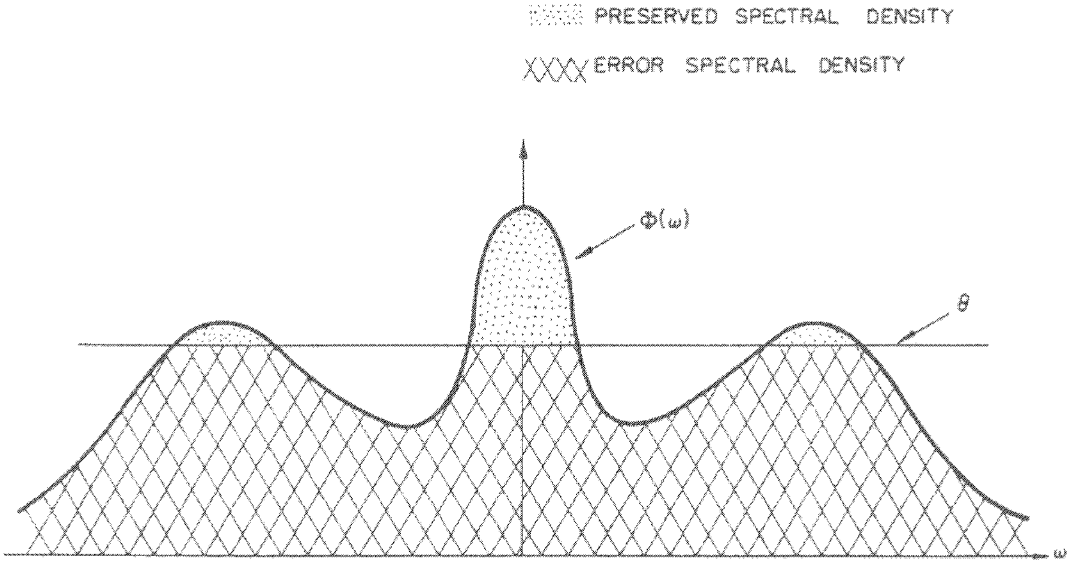

Fig. 2.

IV. Shannon's 1959 Paper

In 1959, Shannon delivered a paper at the IRE Convention in New York City entitled "Coding Theorems for a Discrete Source with a Fidelity Criterion" [2]. This paper not only introduced the term "rate-distortion function" but also put lossy source coding on a firmer mathematical footing. Major contributions of the paper are as follows.

- Definition and properties of the rate-distortion function.
- Calculating and bounding of $R(D)$.
- Coding theorems.
- Insights into source–channel duality.

A. Definition and Properties of the Rate-Distortion Function

A *discrete information source* is a random sequence $\{X_k\}$. Each X_k assumes values in a discrete set $\mathcal{A}$ called the *source alphabet*. The elements of $\mathcal{A}$ are called the *letters* of the alphabet. We shall assume, until further notice, that there are finitely many distinct letters, say M of them, and shall write $\mathcal{A} = \{a(0), a(1), \cdots, a(M-1)\}$. Often we let $a(j) = j$ and hence $\mathcal{A} = \{0, 1, \cdots, M-1\}$; the binary case $\mathcal{A} = \{0, 1\}$ is particularly important.

The simplest case, to which we shall restrict attention for now, is that in which:

1) The X_k are independent and identically distributed (i.i.d.) with distribution $\{p(a), a \in \mathcal{A}\}$.
2) The distortion that results when the source produces the n-vector of letters $\underline{a} = (a_1, \cdots, a_n) \in \mathcal{A}^n$ and the communication system delivers the n-vector of letters $\underline{b} = (b_1, \cdots, b_n) \in \mathcal{B}^n$ to the destination as its representation of $\underline{a}$ is

$$d_n(\underline{a}, \underline{b}) = n^{-1} \sum_{k=1}^{n} d(a_k, b_k). \tag{10}$$

Here, $d(\cdot, \cdot) : \mathcal{A} \times \mathcal{B} \to [0, \infty)$ is called a *single-letter distortion measure*. The alphabet $\mathcal{B}$—variously called the reproduction alphabet, the user alphabet and the destination alphabet—may be but need not be the same as $\mathcal{A}$. We shall write $\mathcal{B} = \{b(0), b(1), \cdots, b(N-1)\}$, where $N < M$, $N = M$, and $N > M$ all are cases of interest. When (10) applies, we say we have a *single-letter fidelity criterion* derived from $d(\cdot, \cdot)$.

Shannon defined the *rate-distortion function* $R(\cdot)$ as follows. First, let $Q = \{Q(b \mid a), a \in \mathcal{A}, b \in \mathcal{B}\}$ be a conditional probability distribution over the letters of the reproduction alphabet given a letter in the source alphabet.[3] Given a source distribution $\{p(j)\}$, we associate with any such Q two nonnegative quantities $d(Q)$ and $I(Q)$ defined by

$$d(Q) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a) Q(b \mid a) d(a, b)$$

and

$$I(Q) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a) Q(b \mid a) \log \left(\frac{Q(b \mid a)}{q(b)} \right)$$

where

$$q(b) = \sum_{a \in \mathcal{A}} p(a) Q(b \mid a).$$

The quantities $d(Q)$ and $I(Q)$ are, respectively, the average distortion and the average Shannon mutual information associated with Q.

The rate-distortion function of the i.i.d. source $\{X_k\}$ with letter distribution $\{p(a) = P[X_k = a]\}$ with respect to the single-letter fidelity criterion generated by $d(\cdot, \cdot)$ is defined by the following minimization problem:

$$R(D) = \min_{Q : d(Q) \le D} I(Q). \tag{11}$$

[3] Such a Q often is referred to as a *test channel*. However, it is preferable to call it a *test system* because it functions to describe a probabilistic transformation from one end of Fig. 1 to the other—from the source all the way to the user—not just across the channel. Indeed, the rate-distortion function has nothing to do with any channel *per se*. It is a descriptor of the combination of an information source and a user's way of measuring the distortion of approximations to that source.

Since the generally accepted object of communication is to maximize mutual information, not to minimize it, many people find the definition of the rate-distortion function counterintuitive.[4] In this regard it often helps to interchange the independent and dependent variables, thus ending up with a *distortion-rate function* defined by

$$D(R) = \min_{Q: I(Q) \le R} d(Q). \tag{12}$$

Everyone considers that minimizing average distortion is desirable, so no one objects to this definition. Precisely the same curve results in the (D, R)-plane, except that now R is the independent variable instead of D. Distortion-rate functions are more convenient for certain purposes, and rate-distortion functions are more convenient for others. One should become comfortable with both.

Properties of the rate-distortion function include:

a) $R(D)$ is well defined for all $D \ge D_{\min}$, where

$$D_{\min} = \sum_{a \in \mathcal{A}} p(a) \min_{b \in \mathcal{B}} d(a,b).$$

The distortion measure can be modified to assure that $D_{\min} = 0$. This is done via the replacement $d(a,b) \leftarrow d(a,b) - \min_b d(a,b)$, whereupon the whole rate-distortion curve simply translates leftward on the D-axis by $D_{\min}$.

b) $R(D) = 0$ for $D \ge D_{\max}$, where

$$D_{\max} = \min_b \sum_a p(a) d(a,b).$$

$D_{\max}$ is the maximum value of D that is of interest, since $R(D) = 0$ for all larger D. It is the value of D associated with the best guess at $\{X_k\}$ in the absence of any information about it other than *a priori* statistical knowledge. For example, $D_{\max} = 1 - \max_a p(a)$ when $\mathcal{A} = \mathcal{B}$ and $d(a,b) = 1$ if $b \ne a$ and 0 if $b = a$.

c) $R(D)$ is nonincreasing in D and is strictly decreasing at every $D \in (D_{\min}, D_{\max})$.

d) $R(D)$ is convex downward. It is strictly convex in the range $(D_{\min}, D_{\max})$ provided $N \le M$, where $N = |\mathcal{B}|$ and $M = |\mathcal{A}|$. In addition to the ever-present straight-line segment $R(D) = 0$, $D \ge D_{\max}$, if $N > M$ then $R(D)$ can possess one or more straight-line segments in the range $D_{\min} < D < D_{\max}$.

e) The slope of $R(D)$ is continuous in $(D_{\min}, D_{\max})$ and tends to $-\infty$ as $D \downarrow D_{\min}$. If there are straight-line segments in $(D_{\min}, D_{\max})$ (see d) above), no two of them share a common endpoint.

f) $R(D_{\min}) \le H$, where

$$H = -\sum_{a \in \mathcal{A}} p(a) \log p(a)$$

is the source entropy. If for each $a \in \mathcal{A}$ there is a unique $b \in \mathcal{B}$ that minimizes $d(a,b)$, and each $b \in \mathcal{B}$ minimizes $d(a,b)$ for at most one $a \in \mathcal{A}$, then $R(D_{\min}) = H$.

Some of these properties were established by Shannon [2], including the essential convexity property d). For proofs of the others see Jelinek [27], Gallager [7], and Berger [26].

B. Calculating and Bounding of $R(D)$

1) Calculating Discrete Rate-Distortion Functions: The domain of variation of Q in the definition of $R(D)$ (see (11)) is contained in the $M(N-1)$-dimensional probability simplex defined by the equality constraints

$$\sum_b Q(b \mid a) = 1, \qquad \text{for every } a \in \mathcal{A}$$

and the inequality constraints

$$Q(b \mid a) \ge 0, \qquad \text{for all } (a,b) \in \mathcal{A} \times \mathcal{B}.$$

In addition, the variation is confined to those Q's that satisfy the constraint on the average distortion, namely,

$$d(Q) := \sum_a \sum_b p(a) Q(b \mid a) d(a,b) \le D.$$

Moreover, the objective function $I(Q)$ is a convex function of Q.[5] Hence, determining $R(D)$ amounts to solving a convex mathematical programming problem. This justifies the following statements.

1) There are no local minima in the search region, just a lone global minimum. Hence, $R(D)$ exists despite the fact that a minimum rather than an infimum appears in its definition because this minimum always is achieved, not just closely approached. The minimum need not necessarily occur at a distinct point; it may be common to a subset of points that constitute a closed, convex subset of the domain.
2) Kuhn–Tucker theory provides necessary and sufficient conditions met by a test system Q that minimizes $I(Q)$ subject to the constraints (i.e., solves the minimization problem that defines $R(D)$).
3) The constraint $D(Q) \le D$ always is satisfied with equality by the minimizing Q. Hence, all the constraints except $Q(b \mid a) \ge 0$ can be handled by Lagrange multiplier theory.

[4] Indeed, Shannon himself seems to have fallen prey to said information-maximizing mindset in the abstract of his 1959 paper, where he wrote (or someone typed):

> In this paper a study is made of the problem of coding a discrete source of information, given a *fidelity criterion* or a *measure of the distortion* of the final recovered message at the receiving point relative to the actual transmitted message. In a particular case there might be a certain tolerable level of distortion as determined by this measure. It is desired to so encode the information that the maximum (sic) possible signaling rate is obtained without exceeding the tolerable distortion level.

The final sentence of this quote should be replaced by, say, "It is desired to minimize the signaling rate devoted to the encoded version of the information subject to the requirement that the tolerable distortion level is not exceeded."

[5] That is,

$$I(\lambda Q_1 + (1-\lambda) Q_2) \le \lambda I(Q_1) + (1-\lambda) I(Q_2)$$

for any $\lambda \in [0,1]$ and any two test systems $\{Q_1(b \mid a)\}$ and $\{Q_2(b \mid a)\}$. See, for example, [7].

Because of the last item on this list, much insight can be gained into the problem of computing $R(D)$ by temporarily ignoring the constraints $Q(b \mid a) \geq 0$ and equating to zero the derivative of the Lagrangian functional $J(Q) = I(Q) + sd(Q)$ with respect to each component $Q(b \mid a)$ of Q. Following this approach, Shannon [2] showed that, for a fixed value s of the Lagrange multiplier, the minimizing Q, call it $\{Q_s(b \mid a)\}$, always is given in terms of a probability distribution $\{q_s(b), b \in \mathcal{B}\}$ by the prescription

$$Q_s(b \mid a) = \lambda_s(a) q_s(b) \exp[sd(a,b)]$$

where

$$\lambda_s^{-1}(a) = \sum_b q_s(b) \exp[sd(a,b)].$$

This reduces the problem of computing the point on the rate-distortion function parameterized by s to that of determining the unknown distribution $\{q_s(b)\}$. The hardest part of that is to determine for which values of b, if any, $q_s(b) = 0$. In certain problems with sufficient symmetry and/or small enough $|\mathcal{B}|$, $q_s(b)$ is strictly positive for all $b \in \mathcal{B}$ (except perhaps at $s = s_{\max}$, the value of s that corresponds to $D_{\max}$.) Shannon [2] used this circumstance to determine $R(D)$ in the special case of an equiprobable M-ary source with $d(a,b) = 1$ if $b \neq a$ and $d(a,b) = 0$ if $b = a$. The result is

$$R(D) = \log_2 M - h(D) - (1-d)\log_2(M-1), \qquad 0 \leq D \leq 1 - M^{-1} = D_{\max} \tag{13}$$

where $h(\cdot)$ is Shannon's binary entropy function

$$h(x) = -x \log_2 x - (1-x)\log_2(1-x).$$

The optimizing Q is

$$Q(b \mid a) = \begin{cases} 1 - D, & \text{if } b = a \\ D/(M-1), & \text{if } b \neq a \end{cases}$$

which says that the whole system should be constructed in such a way that its end-to-end probabilistic transition structure mimics that of an M-ary Hamming channel.

In the special case of a binary equiprobable source ($M = 2$), (13) reduces to

$$R(D) = 1 - h(D) = 1 + D\log_2 D + (1 - D\log_2(1-D), \qquad 0 \leq D \leq 1/2 = D_{\max}.$$

The desired end-to-end system behavior then becomes that of a binary symmetric channel (BSC) with crossover probability D. It follows that, if one seeks to send a Bernoulli($\frac{1}{2}$) source over a BSC that is available once per source letter, then optimum performance with respect to the single-letter fidelity criterion generated by $d(a,b) = 1 - \delta_{a,b}$ can be obtained simply by connecting the source directly to the BSC and using the raw BSC output as the system output. There is need to do any source and/or channel coding. The average distortion will be $D = \epsilon$, where ϵ is the crossover probability of the BSC.

This is another instance of a double coincidence like that of Section II-E. The first coincidence is that a Bernoulli($\frac{1}{2}$) source drives every BSC at capacity, and the second coincidence is that BSC(ϵ) provides precisely the end-to-end system transition probabilities that solve the rate-distortion problem for the Bernoulli($\frac{1}{2}$) source at $D = \epsilon$. Again, this double coincidence represents a precarious saddle point. If the channel were not available precisely once per source symbol, if the Bernoulli source were to have a bias $p \neq \frac{1}{2}$, if the channel were not perfectly symmetric, or if the distortion measure were not perfectly symmetric (i.e., if $d(0,1) \neq d(1,0)$), it would become necessary to employ source and channel codes of long memory and high complexity in order to closely approach performance that is ideal in the sense of achieving equality in the information transmission inequality (5). Shannon illustrated how algebraic codes could be "used backwards" to encode the equiprobable binary source efficiently with respect to the error frequency criterion for cases in which the medium connecting the source to the user is anything other than a BSC. This idea was extended by Goblick [29] who proved that ideally efficient algebraic codes exist for this problem in the limit of large blocklength.

To enhance appreciation for the fragility of the double-coincidence saddle point, let us replace the Bernoulli($\frac{1}{2}$) source with a Bernoulli(p) source, $p \neq \frac{1}{2}$. Calculations (see [26, pp. 46–47]) reveal that the rate-distortion function then becomes

$$R(D) = h(p) - h(D), \qquad 0 \leq D \leq \min(p, 1-p) = D_{\max}.$$

Although the optimum *backward* system transition probabilities $P(a \mid b)$ remain those of BSC(D), the optimum *forward* transition probabilities become those of a binary asymmetric channel. Hence, it is no longer possible to obtain an optimum system simply by connecting the source directly to the BSC and using the raw channel output as the system's reconstruction of the source. Not only does the asymmetric source fail to drive the BSC at capacity, but the BSC fails to provide the asymmetric system transition probabilities required in the $R(D)$ problem for $p \neq 1/2$. For example, suppose $p = 0.25$ so that $R(D) = 0.811 - h(D)$ bits per letter, $0 \leq D \leq 0.25 = D_{\max}$. Further suppose that $\epsilon = 0.15$ so that the channel capacity is $C = 1 - h(0.15) = 0.390$ bits per channel use. Direct connection of the source to the channel yields an error frequency of $D = \epsilon = 0.15$. However, evaluating the distortion-rate function at C in accordance with (5) shows that a substantially smaller error frequency of $R^{-1}(0.390) = 0.0855$ can be achieved using optimum source and channel coding.

The formula that Shannon provided for the rate-distortion function of an M-ary equiprobable source with distortion assessed by the single-letter distortion measure $d(a,b) = 1 - \delta_{a,b}$, namely (13), actually is a special case of a more general result published the preceding year by Erokhin [31], a student in Kolmogorov's seminar. At Kolmogorov's urging Erokhin considered a general i.i.d. discrete source with a finite or countably infinite alphabet and found a formula for what we would now call its rate-distortion function with respect to the error frequency criterion. Erokhin's result is that the rate-distortion function in question is given parametrically by the equations

$$D_\theta = 1 - S_\theta + \theta(N_\theta - 1)$$
$$R_\theta = -\sum_{a:p(a)>\theta} p(a)\log p(a) + (1-D_\theta)\log(1-D_\theta) + (N_\theta - 1)\theta\log\theta,$$

where N_θ is the number of source letters whose probability exceeds θ and S_θ is the sum of the probabilities of these N_θ letters. The parameter θ traverses the range $0 \le \theta \le p(a_2)$ as D varies from 0 to $D_{\max} = 1 - p(a_1)$, where $p(a_1) \ge p(a_2) \ge p(a)$ for all other $a \in \mathcal{A}$.

Moreover, the optimum output probability distribution $\{q(b)\}$ corresponding to parameter value θ is

$$q(b) = \frac{\max[0, p(b) - \theta]}{\sum_b \max[0, p(b) - \theta]}.$$

This, in turn, shows that said $\{q(b)\}$ is supported on a subset of letters assigned high probability by the source. In other words, more and more letters of low source probability are dropped out of use as reproduction letters as θ, and hence D, increases. Once a letter drops out of use, it never reappears for larger values of D, a property that is by no means common to all rate-distortion functions. For cases in which $|\mathcal{A}| < \infty$, the parameter θ can be eliminated when $0 \le \theta \le p_{\min}$, where $p_{\min}$ denotes the smallest of the $p(a)$, $a \in \mathcal{A}$. This results in the explicit formula

$$R(D) = H - h_b(D) - D\log(|\mathcal{A}| - 1), \qquad 0 \le D \le p_{\min}(|\mathcal{A}| - 1).$$

We shall later interpret this as an instance of tightness of a discrete version of the Shannon lower bound, with $p_{\min}(|\mathcal{A}| - 1)$ in the role of the associated critical value of distortion D^*.

2) The Shannon Lower Bound: Shannon then revisited the problem of continuous amplitude sources. Skeptics of Shannon's prowess in rigorous mathematics[6] should note that the paragraph introducing his treatment of "cases where the input and output alphabets are not restricted to finite sets but vary over arbitrary spaces" contains the phraseology "Further, we assume a probability measure P defined over a Borel field of subsets of the A space. Finally, we require that, for each z belonging to B, $d(m, z)$ is a measurable function with finite expectation." [2] For the case of a difference distortion measure $d(a, b) = d(b - a)$ and an i.i.d. time-discrete source producing absolutely continuous random variable (r.v.) with probability density $p(\cdot)$, Shannon used variational principles to derive a lower bound $R_L(D)$ to the rate-distortion function described parametrically as follows:

$$R(D_s) \ge R_L(D_s) := h(p) - h(g_s) \tag{14}$$

where

$$h(p) = -\int_{-\infty}^{\infty} p(x)\log p(x)\,dx$$

is the differential entropy of the instantaneous source density and $h(g_s)$ is the differential entropy of the "tilted" density

$$g_s(z) = \frac{\exp(sd(z))}{\int_{-\infty}^{\infty}\exp(sd(z))\,dz}$$

associated with the parameter s and the difference distortion measure $d(\cdot)$. The distortion coordinate D_s is given by

$$D_s = \int_{-\infty}^{\infty} d(z)\exp(sd(z))\,dz.$$

$R_L(D)$ of (14) has been named the Shannon lower bound [26].

In the case of squared error, $d(a, b) = (b-a)^2$, the parameter s can be eliminated and the Shannon lower bound can be expressed in the compact form

$$R_L(D) = \frac{1}{2}\log(Q_0/D)$$

where Q_0 is the entropy power of the source density. That is, Q_0 is the variance of a Gaussian r.v. that has the same differential entropy as does $p(\cdot)$, namely,

$$Q_0 = (2\pi e)^{-1}\exp(2h(p)).$$

If a typical source r.v. X_k can be expressed as the sum of two independent r.v.'s, one of which is $\mathcal{N}(0, D)$, then $R(D) = R_L(D)$. The largest value of D for which this can be done is called the *critical distortion* and is denoted by D^*. The critical distortion can be as small as 0, in which case the Shannon lower bound to the MSE rate-distortion function is nowhere tight. At the other extreme, if the source variables are themselves $\mathcal{N}(0, \sigma^2)$ r.v., then $Q_0 = D^* = \sigma^2 = D_{\max}$ so that the Shannon lower bound is everywhere tight and

$$R(D) = \max\left[0, \frac{1}{2}\log(\sigma^2/D)\right]. \tag{15}$$

This result is the time-discrete version of (4). It corresponds to taking samples of the ideal bandlimited Gaussian noise process $2W$ times per second and defining $\sigma^2 = S_0W$. Its presence is in keeping with one of Shannon's avowed purposes for writing his 1959 paper, namely, to provide "an expansion and detailed elaboration of ideas presented in [1], with particular reference to the discrete case." (Interpreting "discrete" here to mean discrete amplitude and/or discrete time.)

It is noteworthy that, even when treating situations characterized by abstract reproduction alphabets, Shannon nonetheless meticulously employed discrete output random variables. "Consider a finite selection of points z_i $(i = 1, 2, \cdots, l)$ from the B space, and a measurable assignment of transition probabilities $q(z_i \mid m)$" [2]. Perhaps Shannon did this to insulate the reader from the theory of abstract spaces, but this seems unlikely given his accompanying use of the words "measurable assignment of transition probabilities." Also, providing the reader with such insulation was less a matter for concern in 1959 as it had been in 1948. A better explanation is that Shannon appreciated that the representation of the source

[6] There are none who doubt Shannon's insight and creativity. However, there are those who think that Shannon wrote his papers in a mathematically casual style not to make them more widely accessible but because he was not conversant with the measure-theoretic approach to probability and random processes. Those people are mistaken. That the renowned academician A. N. Kolmogorov referred to Shannon's conception of information coding in terms of the asymptotics of overlapping spheres in n-dimensional finite geometries in the limit as $n \to \infty$ as "incomparably deep" [4] should in itself have been enough to silence such skepticism, but alas it persists.

would always have to be stored digitally; indeed, his major motivation for Section V in 1948 had been to overcome the challenge posed by the fact that continuous-amplitude data has infinite entropy. But, there is an even better explanation. It turns out that the output random variable $\hat{X}$ that results from solving the rate-distortion problem for a continuous-amplitude source usually is discrete! The region, if any, in which the Shannon lower bound is tight for distortions smaller than some positive D^* turns out to be the exception rather than the rule in that $\hat{X}$ is indeed continuous for each D in the range $[0,\, D^*)$. However, for $D \geq D^*$ Rose [158] recently has shown that the optimum $\hat{X}$ is discrete. (See also work of Fix [159] dealing with cases in which X has finite support.) In retrospect, it seems likely that Shannon knew this all along.

C. Source Coding and Information Transmission Theorems

Shannon did not state or prove any lossy source coding theorems in his classic 1948 paper. He did, however, state and sketch the proof of an end-to-end information transmission theorem for the system of Fig. 1, namely, his Theorem 21. Since the notation $R(D)$ did not exist in 1948, Shannon's theorem statement has v_1 in place of D and R_1 in place of $R(D)$. It reads:

> Theorem 21: If a source has a rate R_1 for a valuation v_1 it is possible to encode the output of the source and transmit it over a channel of capacity C with fidelity as near v_1 as desired provided $R_1 \leq C$. This is not possible if $R_1 > C$.

In 1959 Shannon included the word "Theorems" in the title of his article [2] and was true to his word.

He began by generalizing from a single-letter distortion measure to a *local distortion measure of span* g, denoted $d : \mathcal{A}^g \times \mathcal{B}^g \rightarrow [0, \infty)$, and then defining the distortion for blocks of length $m \geq g$ according to the prescription

$$d(\underline{a}, \underline{b}) = \frac{1}{m-g+1} \sum_{k=1}^{m-g+1} d(a_k, a_{k+1}, \cdots, a_{k+g-1}; \\ b_k, b_{k+1}, \cdots, b_{k+g-1}).$$

Local distortion measures represent a significant improvement over single-letter distortion measures in many situations of interest. For example, if one is compressing a text that contains multidigit numbers, such as a company's annual report, a local distortion measure allows one to assign greater penalties to errors made in the more significant digits of such numbers than to errors in the less significant digits. Generalizing to a local distortion measure in no way complicates the proof of source coding theorems, but it significantly complicates the analytical determination of $R(D)$ curves [30].

Next he extended from i.i.d. sources to general ergodic sources.[7] This required generalizing the definition of $R(D)$ to

$$R(D) = \liminf_{m \rightarrow \infty} R_m(d)$$

where $R_m(d)$ is to defined to be the minimum mutual information rate between a vector $\underline{X}$ of m successive source letters and any random vector $\underline{\hat{X}}$ jointly distributed with $\underline{X}$ in such a way that $Ed(\underline{X}, \underline{\hat{X}}) \leq D$, where $d(\cdot, \cdot)$ is the operative local distortion measure of span g. He then stated a "Positive Coding Theorem" and a "Converse Coding Theorem" and sketched their proofs. Both theorems were phrased in terms of what can and what cannot be accomplished when faced with the task of transmitting information about the given source over a given channel of capacity C and then generating a reproduction of the source based on the information available at the channel output. As such, they are examples of what we now call information transmission theorems or joint source–channel coding theorems. We summarize their content by using the first and second sentences of Theorem 21 of Shannon's 1948 paper quoted above, with the terminology appropriately revised to fit the current context.

Positive Theorem: If an ergodic source has a rate-distortion function $R(D)$ with respect to a fidelity criterion generated by a local distortion measure, then it is possible to encode the output of the source and transmit it over a channel of capacity C with fidelity as near D as desired provided $R(D) \leq C$.

Converse Theorem: Let $R(D)$ and C be as in the statement of the Positive Theorem. If $R(D) > C$ then it is not possible to transmit an encoded version of the source data over the channel and then reconstruct the source with fidelity D on the basis of what is received.

It is also possible to state and prove *source coding theorems* that depend only on the source and the distortion measure and have no connection to any channel.

Definition: A block source code of rate R and blocklength n is a collection of $M = \lceil 2^{nR(D)} \rceil$ n-vectors $\mathcal{C} = \{\underline{b}_1, \cdots, \underline{b}_M\}$, where each $\underline{b}_i$ belongs to the nth-power $\mathcal{B}^n$ of the reproduction alphabet.

Definition: Given a block source code $\mathcal{C}$ and any $\underline{x} \in \mathcal{A}^n$, $\underline{b}(\underline{x}) \in \mathcal{C}$ is an *image* of $\underline{x}$ in $\mathcal{C}$ if $d(\underline{x}, \underline{b}(\underline{x})) \leq d(\underline{x}, \underline{b})$ for all $\underline{b} \in \mathcal{C}$; certain vectors $\underline{x}$ may have more than one image in $\mathcal{C}$.

The reader will appreciate that a block source code is simply a collection of vector quantizer "centroids," and that mapping each source word into an image of itself amounts to minimum-distortion vector quantization.

Positive Source Coding Theorem: Let $R(D)$ denote the rate-distortion function of an ergodic source with respect to a local distortion measure d. If $R > (R(D)$ then for sufficiently large n there exists a block source code $\mathcal{C}$ of rate R and blocklength n for which $Ed(\underline{X}, \underline{b}(\underline{X})) \leq D$.

Converse Source Coding Theorem: If $R < R(D)$ then for all n there does not exist a blocklength-n source code of rate R for which $Ed(\underline{X}, \underline{b}(X)) \leq D$.

The proof of the Converse Theorem given by Shannon is adequately rigorous. A corresponding proof of the Converse Source Coding Theorem can be obtained similarly by invoking the readily established facts that $R_n(\cdot)$ is monotonic nonin-

[7] Ergodic sources need not necessarily be stationary. It appears that Shannon intended his discussion to apply to stationary ergodic sources.

creasing and convex downward for every n at appropriate places in the argument.

The situation with respect to the Positive Theorem is more delicate. The nuance is that proving the theorem involves approximating the source by a sequence of sources the nth of which produces successive n-vectors independently of one another according to the n-dimensional marginal of the given stationary source. As $n \to \infty$ intuition suggests that the approximating sources will "converge" to the given source in the sense of mimicking its dependencies ever more closely, except perhaps in relatively narrow intervals near the boundaries of successive blocks. However, there are certain ergodic sources that exhibit extraordinarily long-range statistical dependencies. Initial efforts to prove the Positive Theorem rigorously in the generality stated by Shannon encountered obstacles imposed by the possibility of such long-range dependencies. Over the decades, a succession of increasingly general theorems were proved. First, it was proved only for finite-order Markov sources, then for strongly mixing sources [24], then for block-ergodic sources [25], then for weakly mixing sources, and finally for general stationary ergodic sources [7]. The extent to which Shannon knew, or at least intuited, that the Positive Theorem is true for general ergodic sources shall remain forever unresolved. Later, it was shown that even the ergodic assumption can be removed; stationariness is sufficient [15]. Also, a proof of the source coding theorem via large deviations theory was developed by Bucklew [16].

In 1993 Kieffer wrote an invited survey paper [17] concerning source coding with a fidelity criterion. This comprehensive and well-crafted article focused principally on source coding theorems, recapitulating how they were developed with increasing generality over time, including relatively recent emphases on universality, multiterminal models, and coding for sources modeled as random fields. Kieffer was selected for this task in considerable measure for his several contributions that proved source coding theorems with increasingly relaxed conditions in increasingly general contexts [18], [19], [20], [21], [22], [23]. Kieffer's survey article also contains an invaluable bibliography of 137 items.

It is not our purpose here to enter into the details of proofs of source coding theorems and information transmission theorems. Suffice it to say that at the heart of most proofs of positive theorems lies a random code selection argument, Shannon's hallmark. In the case of sources with memory, the achievability of average distortion D at coding rate $R_n(D)$ is established by choosing long codewords constructed of concatenations of "super-letters" from $\mathcal{B}^n$. Each super-letter is chosen independently of all the others in its own codeword and in the other codewords according to the output marginal $q(\underline{b})$ of the joint distribution $p(\underline{a})Q(\underline{b} \mid \underline{a})$ associated with the solution of the variational problem that defines $R_n(D)$.

D. Insights into Source-Channel Duality

Shannon concluded his 1959 paper on rate-distortion theory with some memorable, provocative remarks on the duality of source theory and channel theory. He mentions that, if costs are assigned to the use of its input letters of a channel, then determining its capacity subject to a bound on expected transmission cost amounts to *maximizing* a mutual information subject to a linear inequality constraint and results in a capacity–cost function for the channel that is *concave* downward. He says, "Solving this problem corresponds, in a sense, to finding a source that is just right for the channel and the desired cost." He then recapitulates that finding a source's rate-distortion function is tantamount to *minimizing* a mutual information subject to a linear inequality constraint and results in a function that is *convex* downward. "Solving this problem," Shannon says, "corresponds to finding a channel that is just right for the source and allowed distortion level." He concludes this landmark paper with the following two provocative sentences:

> This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.

V. The Sixties

With regard to rate distortion, the 1960's were a decade characterized principally by doctoral dissertations, conference presentations, and book sections. Centers of rate-distortion theory research were M.I.T. (to which Shannon had moved from Bell Labs), Yale, Harvard, Cornell, UC Berkeley, and USC. Columbia, Brooklyn Poly, Purdue, Stanford, and Caltech/JPL also were represented.

A. MIT

At M.I.T., Fano and later Gallager supervised doctoral dissertations that addressed aspects of rate distortion. Specifically, Goblick [29] wrote about algebraic source codes, about rate distortion for certain situations involving side-information, and about the rate at which the performance of block source codes could be made to converge to points on the $R(D)$ curve as blocklength increases. Another dissertation, by Pilc [32], [33] also bounded the performance of optimum source codes as a function of their blocklength. Recent research by Yang, Zhang, and Wei corrects the work of Pilc and extends it to sources with unknown statistics that possess memory [34], [35], [36]; see also related work by Linder, Lugosi, and Zeger [37], [38].

Pinkston wrote both a masters thesis [39] and a doctoral dissertation [40] concerning aspects of rate-distortion theory. The former concentrated on computing $R(D)$ and developing codes for situations in which $d(a, b) = \infty$ for certain (a, b)-pairs; this theory parallels analogous in some respects to the theory of the zero-error capacity of discrete channels. The latter also appeared in part as a journal paper [41].

B. Yale

At Yale, Schultheiss supervised a bevy of doctoral students who studied rate distortion. Gerrish [28] dissected the variational problem defining $R(D)$ in considerable detail. Although he did not use Kuhn–Tucker theory, Gerrish derived the necessary and sufficient conditions for optimality that application of that theory would have produced. Specifically, he showed that $Q_s(b \mid a)$, as given above in Section IV-B, is optimum for parameter value s if and only if the output distribution $q_s(b)$ that generates it satisfies the condition

$$c_s(b) := \sum_a p(a)\lambda_s(a)\exp[sd(a,b)]\begin{cases} = 1, & \text{if } q_s(b) > 0 \\ \leq 1, & \text{if } q_s(b) = 0 \end{cases}$$

where

$$\lambda_s^{-1}(a) = \sum_b q_s(b)\exp[sd(a,b)].$$

Using this result Gerrish considerably expanded the class of discrete rate-distortion problems for which $R(D)$ could be determined analytically. He also concocted the famous example

$$\begin{aligned} &\mathcal{A} = \mathcal{B} = \{0, 1, 2\} && d(a,b) = |b-a| \\ &p(1) = p && p(0) = p(2) = (1-p)/2. \end{aligned}$$

This example has the property that, if p is sufficiently small, then $q(1)$ is positive for a range of small D, is zero for slightly larger distortions, and then becomes nonzero for still larger distortions; at $D = D_{\max} = 1-p$, $q(1) = 1$ regardless of the value of p. This example showed that even in a case with small alphabets and considerable symmetry, there is no simple behavior to the set $\{b : q(b) = 0\}$ as a function of distortion, in contrast to what Erokhin had established for the error frequency criterion $d(a,b) = 1 - \delta_{a,b}$. McDonald and Schultheiss [42]–[44] obtained results generalizing the Shannon–Pinsker water table result for Gaussian processes and MSE distortion to different sorts of constraints on the error spectrum. Huang, Spang, and Schultheiss [45], [46] derived enhanced vector quantization schemes with and without feedback by using orthogonal transformations inspired by considerations from rate-distortion theory.

C. Cornell

Research in rate-distortion theory at Cornell was spearheaded by Jelinek and subsequently by Berger. Jelinek analyzed the behavior of rate-distortion functions for small distortion [59]. Also, he used the theory of branching processes to show that performance arbitrarily close to the $R(D)$ curve could be achieved via tree codes [60]. (See also the paper by Davis and Hellman [58] in which a more precise analysis was conducted using branching processes with random environments.) Jelinek and Anderson [61] introduced the M-algorithm, an implementable procedure for encoding tree codes analogous to sequential decoding and stack decoding of tree and trellis channel codes, and documented its performance relative to bounds from rate-distortion theory. Under Berger's direction, information rates of sources modeled as dynamic systems were determined by Toms [64], tree encoding of Gaussian sources with memory was studied by Dick [65], and studies of complete decoding algorithms for triple-error-correcting algebraic source codes were initiated by Vanderhorst [62], [63]. Also, a paper on using Slepian's permutation codes as a mechanism for lossy source coding was written by Berger, Jelinek, and Wolf during a summer visit to Cornell by Wolf [66]. Solo papers by Berger during this period included a rate-distortion study of Wiener processes [67], [68] and a treatment of coding for unknown sources varying over a class either randomly or under the control of an adversary [69]. It was shown, among other things, that the discrete-time Wiener process also exhibits a critical distortion phenomenon, the value of D^* being $\sigma^2/4$, where σ^2 is the variance of the increment between samples. Furthermore, it was established that the rate-distortion function of the Wiener sequence did indeed specify its MSE information rate despite the process being nonstationary. The treatment of unknown sources, like the work of Sakrison on classes of sources cited below, helped pave the way for subsequent studies of universal lossy source coding.

D. Harvard

At Harvard, Tufts supervised an active group of communication theorists including Ramamoorthy, Fine, Kellogg, Trafton, Leiter, Shnidman, and Proakis. Two others of Tufts's students, Berger and Gish, explicitly considered rate-distortion theory as a means for developing absolute performance limits against which to compare the communication and quantization schemes they analyzed [70], [71]. Berger's results showed that, although optimum PAM systems are quite efficient for communicating various types of data sources over filtered channels with additive Gaussian noise when the SNR is low, the gap between optimum PAM and information-theoretically optimum systems widens meaningfully as the SNR increases. This was among the insights that led Price and others to realize that dramatic gains in signaling rate still remained to be reaped in the transmission of digital data over clean telephone channels. Gish's results led to collaboration with Pierce on a theory of asymptotically efficient quantizing [72].

Studying the expression $\frac{1}{2}\log(Q_0/D)$ for the MSE rate-distortion function of a Gaussian sequence for $D \leq D^*$ (cf. Section III-B), Gish and Berger [73] noticed that the formula for the entropy rate power, namely,

$$Q_0 = \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\, \Phi(\omega)\, d\omega\right]$$

is also the formula for the optimum one-step prediction error. That is, the entropy rate power Q_0 equals the variance of the minimum MSE estimate of X_k based on $\{X_j, j < k\}$. This is both intriguing and confounding. A confluence of fundamental quantities always is intriguing. Here is what is confounding. The sequence of successive one-step prediction errors, also called the *innovations process*, is stationary, zero-mean, uncorrelated, and Gaussian. Let us call it $\{I_k\}$. Rate-distortion theory tells us that $\{I_k\}$ can be encoded with an MSE of D using any data rate $R > \frac{1}{2}\log(Q_0/D)$ but no data rate smaller than this. Hence, in the range $0 \leq D \leq D^*$, the MSE rate-distortion function of $\{X_k\}$ is equal to that of $\{I_k\}$. This suggests that perhaps an optimum encoder should compute $\{I_k\}$ from $\{X_k\}$ and then use a code of rate $\frac{1}{2}\log(Q_0/D)$ to convey the memoryless sequence $\{I_k\}$ to the decoder with an MSE of D. However, it is unclear how the receiver could use these lossy one-step prediction errors to generate a D-admissible estimate of $\{X_k\}$. Furthermore, the rate-distortion problem does not impose a restriction to causal estimation procedures the way the one-step prediction problem does, so the apparent connection between them is enigmatic indeed.

E. UC Berkeley

Sakrison conducted and supervised research in rate-distortion at UC Berkeley. His initial papers [74]–[76] treated source coding in the presence of noisy disturbances, gave geometric insights into the source coding of Gaussian data, and treated the effects of frequency weighting in the distortion measure as part of an effort to deal with edge effects and other perceptual considerations in image coding. His paper with Algazi [77] dealt explicitly with two-dimensional coding techniques for images. In this connection, basic formulas for the information rates of Gaussian random fields were being developed contemporaneously at Purdue by Hayes, Habibi, and Wintz [80]. Sakrison also supervised an important dissertation in which Haskell [79] developed a new representation of the rate-distortion variational problem and used it to compute and bound rate-distortion functions in novel ways. Probably the most significant of Sakrison's contributions was his paper dealing with the information rate of a source that is known only to belong to a certain class of sources but is otherwise unspecified [78]. This paper contributed to setting the foundation for the study of universal lossy coding that flourished in succeeding decades.

F. USC

At USC, Gray [81] studied rate-distortion theory under the able tutelage of Scholtz and Welch. His doctoral dissertation contained many interesting results, perhaps the most startling of which was that the binary-symmetric Markov source exhibited a critical distortion phenomenon with respect to the error frequency distortion measure that was similar to that of MSE rate-distortion functions of stationary Gaussian sequences alluded to previously. Specifically, if $P(1 \mid 0) = P(0 \mid 1) = p$ describes the transition matrix of the binary-symmetric source, he showed that there exists a positive D^* such that

$$R(D) = h(p) - h(D), \qquad 0 \le D \le D^*.$$

What's more, using intricate methods involving Kronecker products of matrices and ordinary products of n matrices drawn in all possible ways from a certain pair of matrices, he found the explicit formula for D^* for this problem, namely,

$$D^* = \frac{1}{2}\left[1 - \sqrt{1 - \left(\frac{m}{1-m}\right)^2}\,\right], \qquad m = \min(p, 1-p).$$

He showed that similar behavior is exhibited by the rate-distortion functions of many autoregressive processes over real and finite alphabets, though explicit determination of D^* has proved elusive for any but the binary-symmetric case cited above. This work and extensions thereof were reported in a series of journal papers [82]–[84]. Gray continued research of his own on rate-distortion throughout succeeding decades and supervised many Stanford doctoral students in dissertations of both theoretical and practical importance. Some of these will be dealt with in the portion of the paper dealing with the early 1970's.

G. Feedback Studies: Stanford, Columbia, Caltech/JPL

Schalkwijk and Kailath's celebrated work on capacity-achieving schemes for channels with feedback gave rise to studies of analogous problems for source coding. In this connection, Schalkwijk and Bluestein [48], Omura [49], and Butman [50] studied problems of efficient lossy coding for cases in which there is a feedback link from the user back to the source encoder.

H. The Soviet School

During the 1960's, Soviet scientists continued to contribute to the mathematical underpinnings of information theory in general and rate-distortion theory in particular; see Pinsker [52], Dobrushin [53], [54], and Tsybakov [51]. Also, Dobrushin and Tsybakov [55] wrote a paper extending rate-distortion theory to situations in which the encoder cannot observe the source directly and/or the user cannot observe the decoder output directly; see also Wolf and Ziv [56]. Like Jelinek, Lin'kov [57] provided tight bounds to $R(D)$ curves of memoryless sources for small D.

I. The First Textbooks

In 1968, the first treatments of rate-distortion theory in information theory texts appeared. Jelinek's [27, ch. 11] and Gallager's [7, ch. 9] were devoted exclusively to rate-distortion theory. Gallager's proved therein Shannon's 1959 claim that ergodicity sufficed for validity of the positive theorem for source coding with respect to a fidelity criterion. He also introduced the following dual to the convex mathematical programming problem that defines $R(D)$: Let $\underline{\lambda}$ denote a vector with components $\lambda(a)$ indexed by the letters of the source alphabet $\mathcal{A}$. Given any real s and any $\underline{\lambda} \ge 0$ let $\underline{c}$ denote the vector with components $c(b)$, $b \in \mathcal{B}$ defined by

$$c(b) = \sum_{a \in \mathcal{A}} \lambda(a) p(a) \exp\left[s d(a,b)\right].$$

Let

$$\Lambda_s = \{\underline{\lambda} \ge \underline{0} : \underline{c} \le \underline{1}\}.$$

Gallager proved that

$$R(D) = \max_{s \le 0, \underline{\lambda} \in \Lambda_s} \left[sD + \sum_{a \in \mathcal{A}} p(a) \log \lambda(a) \right].$$

Expressing $R(D)$ as a maximum rather than a minimum allows one to generate lower bounds to $R(D)$ readily. Just pick any $s \le 0$ and any $\underline{\lambda} \ge \underline{0}$. Then evaluate $\underline{c}$. If the largest component of $\underline{c}$ exceeds 1, form a new $\underline{\lambda}$ by dividing the original $\underline{\lambda}$ by this largest $c(b)$. The new $\underline{\lambda}$ then belongs to Λ_s. It follows that the straight line $sD + \sum_a p(a) \log \lambda(a)$ in the (D, R)-plane underbounds $R(D)$. Not only are lower bounds to $R(D)$ produced aplenty this way, but we are assured that the upper envelope of all these lines actually *is* $R(D)$. This dual formulation is inspired by and capitalizes on the fact that a convex downward curve always equals the upper envelope of the family of all its tangent lines. It turns out that all known interesting families of lower bounds to $R(D)$ are special cases of this result. In particular, choosing the components of $\underline{\lambda}$ such that $\lambda(a)p(a)$ is constant yields the Shannon lower bound when the distortion measure is balanced (i.e., every row of the distortion matrix is a permutation of the first row and every column is a permutation of the first column) and yields a generalization of the Shannon lower bound when the distortion measure is not balanced.

VI. The Early Seventies

The period from 1970 to 1973 rounds out the first 25 years of rate-distortion theory. Although it may have appeared to those working in the field at that time that the subject was reaching maturity, it has turned out otherwise indeed. The seemingly "mined" area of computation of rate-distortion functions was thoroughly rejuvenated. Furthermore, foundations were laid that supported dramatic new developments on both the theoretical and practical fronts that have continued apace in the 25 years since.

Gallager's primary interests turned from information theory to computer science and networks during the 1970's. However, rate-distortion theory thrived at Stanford under Gray, at Cornell under Berger, who wrote a text devoted entirely to the subject [26], at JPL under Posner, at UCLA under Omura and Yao, and at Bell Labs under Wyner.[8]

A. Blahut's Algorithm

A Cornell seminar on the mathematics of population genetics and epidemiology somehow inspired Blahut to work on finding a fast numerical algorithm for the computation of rate-distortion functions. He soon thereafter reported that the point on an $R(D)$ curve parameterized by s could be determined by the following iterative procedure [85]:[9]

Step 0: Set $r = 0$. Choose any probability distribution $q_0(\cdot)$ over the destination alphabet that has only positive components, e.g., the uniform distribution $q_0(b) = 1/|\mathcal{B}|$.

Step 1: Compute

$$\lambda_r(a) = \left(\sum_b q_r(b) \exp\left[sd(a,b)\right] \right)^{-1}, \qquad a \in \mathcal{A}.$$

Step 2: Compute

$$c_r(b) = \sum_a \lambda_r(a)p(a) \exp\left[sd(a,b)\right], \qquad b \in \mathcal{B}.$$

If $\max_b c_r(b) < 1 + \epsilon$, halt.

Step 3: Compute $q_{r+1}(b) = c_r(b)q_r(b)$. $r \leftarrow r + 1$. Return to Step 1.

Blahut proved the following facts.

1) The algorithm terminates for any rate-distortion problem for any $\epsilon > 0$.
2) At termination, the distance from the point (D_r, I_r) defined by

$$D_r = \sum_{a,b} p(a)\lambda_r(a)q_r(b) \exp\left[sd(a,b)\right] d(a,b)$$

and

$$I_r = sD_r + \sum_a p(a) \log \lambda_r(a)$$

to the point $(D, R(D))$ parameterized by s (i.e., the point on the $R(D)$-curve at which $R'(D) = s$) goes to zero as $\epsilon \to 0$. Moreover, Blahut provided upper and lower bounds on the terminal value of $I_r - R(D_r)$ that vanish with ϵ.

Perhaps the most astonishing thing about Blahut's algorithm is that it does not explicitly compute the gradient of $R + sD$ during the iterations, nor does it compute the average distortion and average mutual information until after termination. In practice, the iterations proceed rapidly even for large alphabets. Convergence is quick initially but slows for large r; Newton–Raphson methods could be used to close the final gap faster, but practitioners usually have not found this to be necessary. The Blahut algorithm can be used to find points on rate-distortion functions of continuous-amplitude sources, too; one needs to use fine-grained discrete approximations to the source and user alphabets. See, however, the so-called "mapping method" recently introduced by Rose [158], which offers certain advantages especially in cases involving continuous alphabets; Rose uses reasoning from statistical mechanics to capitalize on the fact, alluded to earlier, that the support of the optimum distribution over the reproduction alphabet usually is finite even when $\mathcal{B}$ is continuous.

B. $R(D)$ Under Gray at Stanford

Following his seminal work on autoregressive sources and certain generalizations thereof, Gray joined the Stanford faculty. Since rate distortion is a generalization of the concept of entropy and conditional entropy plays many important roles, Gray sensed the likely fundamentality of a theory of conditional rate-distortion functions and proceeded to develop it [160] in conjunction with his student, Leiner [161], [162]. He defined

$$R_{X|Y}(D) = \min I(X; \hat{X}|Y)$$

where the minimum is over all r.v. $\hat{X}$ jointly distributed with (X, Y) in such a manner that $E_{X,Y,\hat{X}} d(X, \hat{X}) \le D$. This not only proved of use *per se* but also led to new bounding results for classical rate-distortion functions. However, it did not treat what later turned out to be the more challenging problem of how to handle side-information $\{Y_k\}$ that was available to

[8] Centers of excellence in rate distortion emerged in Budapest under Csiszár, in Tokyo under Amari, in Osaka under Arimoto, in Israel under Ziv and his "descendants," in Illinois under Pursley, and at Princeton under Verdú, but those developments belong to the second 25 years of information theory.

[9] Blahut and, independently, Arimoto [86] found an analogous algorithm for computing the capacity of channels. Related algorithms have since been developed for computing other quantities of information-theoretic interest. For a treatment of the general theory of such max-max and min-min alternating optimization algorithms, see Csiszár and Tusnady [87].

$$R_{\mathrm{WZ}}(D) = \begin{cases} h(p) - h(p * D), & \text{if } 0 \le D \le D_c \\ \text{straight line from } (D_c, h(p) - h(p * D_c)) \text{ to } (p, 0), & \text{if } D_c \le D \le p \end{cases} \tag{17}$$

the decoder only and not to the encoder. That had to await ground-breaking research by Wyner and Ziv [94].

Gray also began interactions with the mathematicians Ornstein and Shields during this period. The fruits of those collaborations matured some years later, culminating in a theory of sliding block codes for sources and channels that finally tied information theory and ergodic theory together in mutually beneficial and enlightening ways. Other collaborators of Gray in those efforts included Neuhoff, Omura, and Dobrushin [163]–[165]. The so-called *process definition* of the rate-distortion function was introduced and related to the performance achievable with sliding block codes with infinite window width (codes in the sense of ergodic theory). It was shown that the process definition agreed with Shannon's 1959 definition of the rate-distortion function $\liminf_{n \to \infty} R_n(D)$ for sources and/or distortion measures with memory. More importantly, it was proved that one could "back off" the window width from infinity to a large, finite value with only a negligible degradation in the tradeoff of coding rate versus distortion, thereby making the theory of sliding block codes practically significant.

Seeing that Slepian and Wolf [93] had conducted seminal research on lossless multiterminal source-coding problems analogous to the multiple-access channel models of Ahlswede [90] and Liao [91], Berger and Wyner agreed that research should be done on a lossy source-coding analog of the novel Cover–Bergmans [88], [89] theory of broadcast channels. Gray and Wyner were the first to collaborate successfully on such an endeavor, authoring what proved to be the first of many papers in the burgeoning subject of multiterminal lossy source coding [92].

C. The Wyner–Ziv Rate-Distortion Function

The seminal piece of research in multiterminal lossy source coding was the paper by Wyner and Ziv [94], who considered lossy source coding with side-information at the decoder. Suppose that in addition to the source $\{X_k\}$ that we seek to convey to the user, there is a statistically related source $\{Y_k\}$. If $\{Y_k\}$ can be observed both by the encoder and the decoder, then we get conditional rate-distortion theory *a la* Gray. The case in which neither the encoder nor the decoder sees $\{Y_k\}$, which perhaps is under the control of an adversary, corresponds to Berger's source-coding game [69]. The case in which the encoder sees $\{Y_k\}$ but the decoder does not was long known [29] to be no different from the case in which there is no $\{Y_k\}$. But the case in which the decoder is privy to $\{Y_k\}$ but the encoder is not proved to be both challenging and fascinating. For the case of a single-letter fidelity criterion and (X_k, Y_k)-pairs that are i.i.d. over the index k, Wyner and Ziv showed that the rate-distortion function, now widely denoted by $R_{\mathrm{WZ}}(D)$ in their honor, is given by

$$R_{\mathrm{WZ}}(D) = \min_{Z \in \mathcal{Z}_D} I(X; Z \mid Y) \tag{16}$$

where $\mathcal{Z}_D$ is the set of auxiliary r.v. $Z \in \mathcal{Z}$ jointly distributed with a generic (X, Y) such that:

1) $Y - X - Z$ is a Markov chain; i.e.,

$$p_{Y,X,Z}(y, x, z) = p_Y(y) p_{X|Y}(x \mid y) p_{Z|X}(z \mid x).$$

2) There exists $g : \mathcal{Z} \times \mathcal{Y} \to \hat{\mathcal{X}}$ such that

$$Ed(X, g(Z, Y)) \le D.$$

3) The cardinality of the alphabet $\mathcal{Z}$ may be constrained to satisfy $|\mathcal{Z}| \le |\mathcal{X}| + 1$.

Consider the special case in which $\{X_k\}$ and $\{Y_k\}$ are Bernoulli$(\frac{1}{2})$ and statistically related as if connected by a BSC of crossover probability $p \le 1/2$ and $d(a, b) = 1 - \delta_{a,b}$. $R_{\mathrm{WZ}}(D)$ for this case is shown in (17) at the top of this page, where $p * d = p(1 - D) + (1 - p)D$ and D_c is such that the sraight-line segment for $D \ge D_c$ is tangent to the curved segment for $D \le D_c$. Berger had used Bergmans [89] theory of "satellites and clouds" to show that (17) was an upper bound to $R(D)$ for this binary-symmetric case. The major contribution of Wyner and Ziv's paper resided in proving a converse to the unlikely effect that this performance cannot be improved upon, and then generalizing to (17) for arbitrary (X, Y) and $d(\cdot, \cdot)$.

The advent of Wyner–Ziv theory gave rise to a spate of papers on multiterminal lossy source coding, codified and summarized by Berger in 1977 [95]. Contributions described therein include works by Körner and Marton, [96]–[98], Berger and Tung [99], [100], Chang [101], Shohara [102], Omura and Housewright [103], Wolfowitz [104], and Sgarro [105]. In succeeding decades, further strides have been made on various side-information lossy coding problems [153], [154], [128], [155], [129], [130], and [156]. Furthermore, challenging new multiterminal rate-distortion problems have been tackled with considerable success, including the *multiple descriptions problem* [145], [150], [146]–[149], [151], [152], [157], [132], the *successive refinements problem* [133], and the *CEO problem* [134]–[136]. Applications of multiple descriptions to image, voice, audio, and video coding are currently in development, and practical schemes based on successive refinement theory are emerging that promise application to progressive transmission of images and other media.

D. Rate Distortion in Random Fields

In order for rate-distortion theory to be applied to images, video, and other multidimensional media, it is necessary to extend it from random processes to random fields (i.e., collections of random variables indexed by multidimensional parameters or, more generally, by the nodes of a graph). The work of Hayes, Habibi, and Wintz [80] extending the water-table result for Gaussian sources to Gaussian random fields already has been mentioned. A general theory of the information theory of random fields has been propounded [131], but we are more interested in results specific to rate distortion. Most of these have been concerned with extending the existence of critical distortion to the random field case and then bounding the critical distortion for specific models. The paper of Hajek and Berger [121] founded this subfield. Work inspired thereby included Bassalygo and Dobrushin [122], Newman [123], Newman and Baker [124] in which the critical distortion of the classic Ising model is computed exactly, and several papers by Berger and Ye [125], [126]. For a summary and expansion of all work in this arena, see [127].

E. Universal Lossy Data Compression

Work by Fitingof, Lynch, Davisson, and Ziv in the early 1970's showed that lossless coding could be done efficiently without prior knowledge of the statistics of the source being compressed, so-called *universal lossless coding*. This was followed by development of Lempel–Ziv coding [106], [107], arithmetic coding [108]–[110], and context-tree weighted encoding [111], [112], which have made universal lossless coding practical and, indeed, of great commercial value.

Universal lossy coding has proven more elusive as regards both theory and practice. General theories of universal lossy coding based on ensembles of block codes and tree codes were developed [138]–[144], but these lack sufficient structure and hence require encoder complexity too demanding to be considered as solving the problem in any practical sense. Recent developments are more attractive algorithmically [113]–[120]. The paper by Yang and Kieffer [117] is particularly intriguing; they show that a lossy source code exists that is universal not only with respect to the source statistics but also with respect to the distortion measure. Though Yang–Kieffer codes code can be selected *a priori* in the absence of any knowledge about the fidelity criterion, the way one actually does the encoding does, of course, depend on which fidelity criterion is appropriate to the situation at hand. All universal lossy coding schemes found to date lack the relative simplicity that imbues Lempel–Ziv coders and arithmetic coders with economic viability. Perhaps as a consequence of the fact that approximate matches abound whereas exact matches are unique, it is inherently much faster to look for an exact match than it is to search a plethora of approximate matches looking for the best, or even nearly the best, among them. The right way to trade off search effort in a poorly understood environment against the degree to which the product of the search possesses desired criteria has long been a human enigma. This suggests it is unlikely that the "holy grail" of implementable universal lossy source coding will be discovered soon.

VII. An Impact on Applications

After 25 years, in 1974, the theory of source coding with a fidelity criterion was well-developed, and extensive treatments were available in the literature, including a chapter in the book by Gallager [7] and the comprehensive text by Berger [26]. However, the impact of rate-distortion theory on the practice of lossy source coding, or data compression, was slight. Indeed, Pierce in his 1973 paper states [221], "In general, I am content with the wisdom that information theory has given us, but sometimes I wish that the mathematical machine could provide a few more details."

To assess further the impact of Information Theory on lossy source coding 25 years after Shannon's original paper, we examine textbooks [222] and paper compendia [201], [183] from around that time. It is clearly evident that except for scalar quantization combined with entropy coding, and scalar quantization combined with transform coding for images, there was little in terms of concrete contributions.

Part of the reason for this elegant theory not influencing the practice of data compression can be traced to the observation that the practitioners of information theory and the designers of data compression systems formed mutually exclusive sets. A 1967 special issue of the Proceedings of the IEEE on Redundancy Removal, generally supports this conclusion, although the papers by Pearson [218] and O'Neal [216] directly incorporate some of Shannon's ideas and results. Perhaps a quote from Pearson's paper implies the gulf that existed: "The concept of a rate-distortion function, once grasped, is conceptually a very satisfying one;" the implication being that rate-distortion theory is not simple to comprehend, at least not at first reading.

However, even information theorists were not optimistic concerning the impact of rate-distortion theory on the practice of lossy source coding, but perhaps for much different reasons—they had a full grasp of the theory, its assumptions, and its implementation requirements, and the picture they saw was challenging. For example, rate-distortion theory requires an accurate source model, and such models for important sources were just being explored and were not well-known [7]. Second, fidelity criteria for important sources such as speech and images were not well-developed, although work was in progress [218]. Third, the AEP and random coding arguments used in proving information-theoretic results implied exponential growth in the codebook, and since, as stated by Wozencraft and Jacobs in their classic text, "One cannot trifle with exponential growth" [248, p. 387]; many outstanding researchers felt that implementation complexity might be the dominating issue [227], [228].

Happily, information theory has had a dramatic impact on lossy source coding, or data compression, in the last 25 years, although the three issues, source models, fidelity criteria, and complexity, remain major considerations.

In addition to the results, insights, and tools provided by Shannon's two original papers [1], [2], the legacy of the first 25 years included the results by Huang and Schultheiss [45] and Wintz and Kurtenbach [246] on bit allocation for scalar quantizers, the rate-distortion function for autoregressive (AR) processes and the MSE fidelity criterion as obtained by Gray [82] and Berger [26], and the tree coding theorem for Gaussian AR processes and the MSE fidelity criterion given by Berger [26]. These results served as a springboard to developing lossy coding techniques for sources with memory that explicitly exhibit information-theoretic concepts.

We start with a discussion of memoryless sources and then proceed to examine results for sources with memory. This is followed by developments of the several approaches to compression that have been useful for important sources such as speech, still images, high-quality audio, and video. The goal is to describe the contributions of information-theoretic results on the practice of lossy source coding without producing a voluminous survey of lossy source compression methods for the several sources.

VIII. Memoryless Sources

Uniform and nonuniform scalar quantization was the primary technique for coding memoryless sources in 1974. These quantizers were usually implemented with an adaptive step size or scaling factor to allow the quantizer dynamic range to respond to rapid variations in source variance, and hence, to reduce the number of levels needed to cover this range with the allowable distortion. The adaptation was based upon tracking the input signal variance and was not motivated by any results from rate-distortion theory. The only real connection to information theory was through the idea of entropy coding the quantizer output alphabet. Subsequent work by Farvardin and Modestino [187] investigated the performance of entropy-constrained scalar quantizers for a variety of source input distributions. At the same time, information theorists were studying the encoding of memoryless sources using rate-distortion theory and began specifically drawing upon random coding arguments.

Random coding arguments are a staple in proving positive coding theorems, and hence, the existence of good source codes. However, many researchers and engineers, especially those interested in applications, find rate-distortion theory wanting in that only the existence of good codes is demonstrated and that no method for finding a good code is given. This view is somewhat myopic, though, because each random coding proof of the existence of a good code actually outlines a code construction. For example, the proof of the achievability of the rate-distortion function given in Cover and Thomas [252] begins by generating a codebook of 2^{NR} reproduction sequences and assigning each of them a codeword index. Then, each input sequence of length N is encoded by finding that sequence in the reproduction codebook that falls within the distortion typical set.

If we actually desire to encode i.i.d. Gaussian sequences of length N with average distortion D, we can then mimic this proof and generate a codebook consisting of 2^{NR} reproduction sequences of length N, where the individual components of each sequence are i.i.d. Gaussian random variables with zero mean and a variance of $\sigma^2 - D$. For a given input sequence of length N, the encoding procedure is to find that sequence in the codebook with the smallest distortion. Thus we see that exactly following the proof of achievability yields an explicit encoding procedure. Unfortunately, to accomplish this encoding step requires an exhaustive comparison of the current input sequence of length N with all sequences in the codebook, and subsequently repeating this comparison for all input sequences of length N to be encoded. Since there are 2^{NR} sequences in the codebook and N must be large to approach optimality, the encoding with such codebooks is arbitrarily complex.

An approach to combatting complexity in random codes is to add structure, and researchers did just this by proving coding theorems for tree and trellis codes that approach the rate-distortion bound arbitrarily closely. Results were obtained for tree coding of binary sources and the Hamming distortion measure by Jelinek and Anderson [61] and for tree coding of i.i.d. Gaussian sources and the MSE fidelity criterion by Dick, Berger, and Jelinek [65]. Viterbi and Omura [242] proved a trellis source coding theorem and Davis and Hellman [58] proved a tree coding theorem for source coding with a fidelity criterion, extending the work of Jelinek [60] and Gallager [190]. While this work did not directly impact applications, it did lay the groundwork for later research on coding sources with memory that has found widespread applications.

Likely, the most important lossy source-coding technique that has sprung directly from information theory is vector quantization. Only those who have a grasp of information theory can appreciate the motivation for studying vector quantizers (VQ's) for memoryless sources; additionally, there were many reasons for not pursuing VQ designs, even from an information theorist's viewpoint. Since performance grows asymptotically with vector length N and the number of input points grows proportionally to 2^{NR}, the exponential growth in encoding complexity seemed too daunting to overcome. Furthermore, there was the indication from rate-distortion theory that for Gaussian sources and the MSE distortion measure, only a 0.255-bit/sample reduction in rate, or a 1.53-dB reduction in distortion, with respect to entropy-coded scalar quantization, was available with vector quantization. Some of the best information theorists found this daunting [228]. However, in the late 1970's and the early 1980's, information theorists did turn their attention to vector quantization.

There were three main thrusts at that time. One centered on developing algorithms for the design of nonuniform VQ's, a second thrust examined uniform VQ performance and design, and a third studied the asymptotic performance (in block-length) of VQ's. Uniform VQ's were based upon lattices in N-dimensional space and this work drew upon algebraic structures and space-filling polytopes. Of course, the attraction to lattice (uniform) VQ's was that the regular structure should allow fast encoding methods to be developed and thus avoid the exponential growth in encoding complexity with vector length N. The study of VQ performance included the lattice VQ structures and extended to higher dimensions some of the approaches from scalar quantization. Algorithm development for nonuniform VQ design began with the algorithm by Linde, Buzo, and Gray [204], now called the LBG algorithm. This algorithm was built upon the k-means algorithm from pattern recognition and the scalar quantizer design methods developed by Lloyd [205]. Although it only guaranteed local optimality and the encoding stage was still exponentially complex in the NR product, the possibility of actually using a VQ and testing its performance became possible.

We leave further broad discussion of scalar and vector quantization to the excellent paper in this issue by Gray and Neuhoff [196]. However, later when discussing particular lossy source compression techniques, we will identify the role of vector quantization and the type of VQ employed.

In many applications, it was (and is) necessary to encode several independent memoryless sources subject to an overall rate or distortion constraint. Thus in those applications with a constraint on total rate, it becomes necessary to minimize total distortion by allocating rate across several scalar quantizers. Clearly evident in each of these contributions is the rate-distortion function for independent and identically distributed Gaussian sources and the MSE fidelity criterion as derived by Shannon [1], [2], or the distortion rate version $D = \sigma^2 2^{-2R}$.

In particular, the bit-allocation methods for scalar quantizers used the distortion rate version of Shannon's result with only a multiplicative scale factor on the variance, viz, as a criterion to be minimized by appropriate allocations of bits (rate). By adjusting this multiplicative factor, the rate distortion relationship could be made to approximate that of a distribution other than Gaussian, such as a Laplacian source.

Thus for M independent sources with respective variances σ_i^2, the individual distortions as a function of rate are $D_i = \gamma_i \sigma_i^2 2^{-2R_i}$ and the total distortion to be minimized is $D = \sum_{i=1}^M D_i$ subject to the overall rate constraint $\sum_{i=1}^M R_i \leq R$. The multiplier γ_i accounts for differences in distributions and for different encoding methods. We append the rate constraint using a Lagrange multiplier, so that the functional to be minimized is

$$J(R, \lambda) = D + \lambda R.$$

Letting γ_i be a constant, the resulting rate allocation is

$$R_i = R + \frac{1}{2}\log_2 \frac{\sigma_i^2}{\left[\prod_{j=1}^M \sigma_j^2\right]^{1/M}}.$$

Although this approach often produces noninteger bit allocations for scalar quantizers, and *ad hoc* modifications are required to produce integer allocations and to achieve the desired total bit rate exactly, the coupling of coding independent sources with different scalar quantizers and "optimal" bit allocation was introduced and served as a framework for numerous future lossy coding techniques for both speech and images. Several other approaches to this bit-allocation problem that allow integer bit allocations and other constraints are now common. See Gersho and Gray [191] for a summary.

IX. Sources with Memory

An obvious approach to coding sources with memory when one already has numerous techniques for coding independent sources is to determine a transformation that models the source memory and then use this transformation to decompose the source with memory into several independent (or nearly independent) memoryless sources. Perhaps the most explicit delineation of this approach and the role of rate-distortion theory in coding sources with memory, in general, and transform image compression in particular, is given by Davisson [182]. Davisson decomposes a source with memory into an expansion of orthogonal components and allocates rate to each of these components according to their variance, an approach that was used previously by Gallager in proving a coding theorem for such Gaussian random process sources [7].

More specifically, Davisson [182] shows that the N-block rate-distortion function for a source with covariance matrix Φ_N and eigenvalues λ_i is given by

$$R_N(D) = \frac{1}{2N}[\log |\Phi_N| - \log D]$$

where the distortion is assumed to be small, $D \leq \min \lambda_i$.

These results amplify the work of Kolmogorov [3] and McDonald and Schultheiss [43]. Davisson also evaluates the rate-distortion function for a first-order Gauss–Markov source, a model often used for images, and expresses the result as a difference between the N-block rate-distortion function and the rate-distortion function asymptotic in N

$$R_N(D) - R(D) = \frac{1}{2N}\log(1-\rho^2)^{-1}.$$

Thus for $\rho = 0.95$, the N-block encoding requires $2.4/N$ more bits per sample than the best possible.

Tree and trellis coding theorems for structures involving transform decompositions are proved in [208], [217], and [209].

The rate-distortion function for AR sources, derived by Gray [82] and Berger [26], was a welcome addition since it came at a time when AR processes were finding their initial application to speech coding [170]–[172]. The elucidation of a tree-coding method for Gaussian AR sources and the proof of a tree-coding theorem for these sources, [26], gave impetus to the application of tree-coding techniques in speech-coding applications.

A. Predictive Coding

Predictive coding was a well-known technique for source compression based upon time-domain waveform-following by the time the second 25 years rolled around. In fact, there had been substantive contributions by the early 1950's [180], [186], [215], with the paper by Elias being significantly motivated by information theory ideas—primarily entropy. However, by 1976, predictive coding was an important practical approach to speech coding and also had applications to image coding [201]. The principal motivation behind this work, as well as its success, was the reduction in the dynamic range of the quantizer input and the decorrelation of the quantizer input by the predictor. Rate-distortion theory was just beginning to have an impact on predictive coders in 1976, and doubtless, Jayant [201] is correct in stating that, "... simple DPCM is still one of the classic triumphs of intuitional waveform coding." However, predictive coding was to become extraordinarily important in applications, and rate-distortion theory motivated coders were to play a major role.

1) Speech Compression: Interestingly, multipath-searched versions of differential encoders, such as delta modulation and DPCM, predated or paralleled the development of the tree-coding theorem by Berger, and were motivated by intuition and estimation theory. In particular, Irwin and O'Neal [199] studied multipath searching of a fixed DPCM system to depth 2, but found only modest increases in SNR. Cutler [181] investigated delayed encoding in a delta modulator with the goal of incorporating a more responsive (over-responsive) encoder to track the sudden onset of pitch pulses.

Anderson and Bodie [168] drawing directly on the theoretical results of Berger [26], and previous work on tree/trellis coding of i.i.d. sources, developed tree coders at 2 bits/sample for speech built around fixed DPCM code generators and the MSE distortion measure. Significant increases in SNR were obtained, but the reconstructed speech had a substantial "hissing" sound superimposed on the highly intelligible speech. Becker and Viterbi [175] considered bit rates of 1 bit/sample and took an approach that included a long-term predictor and a finite-state approximation to the AR model. Both the long- and short-term predictors were adaptive. They also reported work on an alternative excitation based upon a trellis. Stewart [236], [237] pursued trellis codes coupled with AR models and pushed the rates down below 1 bit/sample.

The primary result of these studies was an increase in output SNR, but the output speech quality still suffered from audible noise. To improve this speech quality and make tree coding a viable candidate for speech coding required adaptive code generators and perceptually based fidelity criteria. Wilson and Husain [245] examined 1-bit/sample tree coding of speech using a fixed-noise shaping motivated by the classical C noise weighting from telephony. Later work, using innovative adaptive code generators, perceptually weighted distortion measures, and new tree codes, achieved good-quality speech with tree coding at 8 kbits/s [261].

However, the major impetus for code-excited schemes in speech coding came from the paper by Atal and Schroeder [174] that demonstrated that high-quality speech could be generated by a predictive coder with a Gaussian populated codebook with 1024 entries, each of length 40 samples. The rate was estimated at 4 kbits/s, but the predictor coefficients were not quantized and the analysis-by-synthesis codebook search was accomplished by the use of a Cray computer! Thus this was very much a "proof-of-concept" paper, but a principal difference between this work and previous research by information-theoretic researchers on speech coding was that the authors used a perceptually weighted MSE to select the best codebook excitation sequence.

Atal and Schroeder [174] were aware of the earlier work on tree coding, but they were also motivated by the analysis-by-synthesis speech-coding method called multipulse linear predictive coding [173], where the codebook consisted of several impulses (say, 8 per frame of 40 samples or so) with arbitrary location and arbitrary magnitude. Multipulse linear predictive coding (multipulse LPC) produced good-quality highly intelligent speech, but the complexity of searching a relatively unstructured adaptive codebook was prohibitive. From this initial work, the tremendous effort on codebook-excited speech coders was spawned. The keys to producing high-quality highly intelligent speech with these coders are that the code generators, or predictors, are adaptive and the fidelity criterion includes perceptual weighting. The perceptual weighting attempts to keep the noise spectrum below that of the source spectrum at all frequencies of interest.

Complexity is always an issue in tree coding and codebook-excited approaches. In tree coding, complexity is addressed by nonexhaustive searching of the trees using depth-first, breadth-first, or metric-first techniques [169]. Nontree codebooks typically contain many more samples per codeword than tree codes, so the search complexity for these codebooks is related to codebook structure and sparsity. A breakthrough in codebook excited techniques for speech has been the interleaved single pulse permutation (ISPP) codebook that consists of a few sparse impulse sequences that are phase-shifted versions of each other, where all of the pulses have the same magnitude [230]. Prior to this technique, codebooks were often designed off-line by using training mode vector quantization.

The impact of codebook-based approaches on speech coding standards has been dramatic. As shown in [179] and [194], many of the current standards for speech coding are code-excited predictive coding and the quality obtained by these techniques is much higher than might have been expected. For example, G.729 has a Mean Opinion Score (MOS) rating of 4.1, and G.728, a low-delay standard, has a quality rating of 4.0–4.1 [179]. G.728 employs a five-dimensional gain-shape vector quantizer (VQ) for its excitation vectors. Vector quantization for side-information is also commonly used and plays an important role in achieving the lowest possible transmitted data rate. The VQ's used for the coefficient representation are typically split VQ's so that the dimension of the VQ's can be kept as small as possible. These VQ's are designed using the training mode method and training the VQ's provides a substantial improvement in performance over any other VQ design technique.

2) Image Compression: Tree coding was also studied for image compression and some interesting results were obtained [211], [212]. The success of this approach for images has been much less than that for speech since a good image model is difficult to find and time-domain methods have not been able to keep pace with the much lower bit rates achievable in the transform domains.

B. Source Decompositions

Predictive coding is model-based and it works extremely well when the linear prediction, or autoregressive, model can adequately represent a source. Early on, however, speech and image compression researchers were drawn to frequency-domain decompositions to account for source memory. Of course, this is very much an electrical engineering way of thinking, namely, breaking a signal down into its constituent frequency components, and then coding these components separately. Two prominent examples of this approach are subband coding and transform coding.

In subband coding, the source to be compressed is passed through parallel filter banks that consist of bandpass filters, and the outputs of these filters are decimated and lowpass translated. Each of the resulting time-domain signals is coded using PCM (i.e., scalar quantization), DPCM, or some other time-domain compression technique. At the receiver, each signal is decoded and those signals that were not originally baseband are translated back to their appropriate filter band, all signals are interpolated (upsampled), and then all components are summed to yield the overall reconstructed source representation. One of the original challenges in subband coding was designing subband filters that provided good coverage of the desired frequency band without producing aliasing upon the reconstruction step due to the intermediate subsampling. The key advance was the development of quadrature mirror filters that, although they allow aliasing in the downsampling step at the encoder, these filters cancel the aliasing during the reconstruction at the receiver. These ideas continue to be generalized and extended. Allocating bits across the subbands is a critical step as well, and the approach differs depending upon the source and the application.

Transform coders take an M-block of input source samples and perform an M-point discrete transform on them. The principal idea is that a good transform will yield decorrelated or even independent components and will concentrate the signal energy into fewer significant components. Bit-allocation methods then discard unimportant frequency content and code each of the remaining components according to differing accuracies. The source can then be approximately reconstructed from the coded components via an inverse transform. Most transforms that are popular in compression are unitary and separable.

It can be shown that transform methods are a special case of subband techniques where the subband synthesis filters have impulse responses equal to the transform basis functions, the analysis filter impulse responses are the time-reversed basis functions, and the decimation factor in each band is the transform blocklength. Furthermore, wavelet methods allow for nonuniform tiling of the time–frequency plane, and therefore wavelet expansions generalize subband methods. In fact, any wavelet expansion has a corresponding perfect reconstruction filter bank interpretation. However, the differences between subband techniques and transform-domain techniques for coding are the frequency and time resolution, which leads to a preferred quantization approach.

In the following sections, we discuss subband, transform, and wavelet-based compression methods for speech, still images, video, and high-quality audio, with emphasis on information-theoretic influences.

1) Speech Compression: Interestingly, subband coding found its first applications to speech compression and then later to image compression, while transform coding had its first applications to image coding and later to speech/audio compression. The primary motivation for subband coding in speech compression was the ability to code the subbands with differing numbers of bits in order to isolate distortions to their individual bands and to achieve better perceptual coding performance. This turned out to be solid reasoning and subband coding of speech at 12 to 24 kbits/s is very competitive in performance and complexity. The bit allocations across the subbands can use the rate-distortion theory-motivated constrained optimization approach, but the existing subband speech coders employ experimentally determined allocations.

Most of the transform coders for speech have utilized the discrete cosine transform (DCT), although sinusoidal transforms and wavelets are also popular today. Transform-based coders can easily achieve high-quality speech at 16 kbits/s, and with perceptual coding and analysis-by-synthesis methods, they generate good quality speech down to 4.8 kbits/s. Information theory has not had a major impact on these designs and further discussion of these techniques is left to the references [194], [202], [235].

The application of wavelets to speech coding is relatively new and has yet to produce speech coders that are competitive in rate, quality, and complexity with the predictive coding methods.

2) Image and Video Compression: Transform-based methods have been a dominant force in image compression at rates below 2 bits/pixel for over 30 years. The first rate distortion theoretic result to have an impact on image compression was the distortion rate expression for an i.i.d. Gaussian source subject to an MSE fidelity criterion that was used for bit-allocation calculations in transform coding. Typically, the transform coefficients were assumed to be independent and bits were allocated in proportion to the variances of the coefficients subject to an overall constraint on total bit rate. The solution to the resulting constrained optimization problem yields the bit allocation to achieve the minimum average total distortion.

The optimal transform in terms of energy compaction is the Karhunen-Loeve transform [166] which produces uncorrelated transform coefficients but requires the knowledge of the statistics of sources and often involves highly complicated computations. Among many practical transforms, the Discrete Cosine Transform (DCT) [223] is the one used the most, especially for two-dimensional signals. With good energy compactness and the existence of fast algorithms [176], [244], DCT-based transform coders are used in many applications and coding standards, such as H.320, JPEG [219], [243], and MPEG [178], [189].

Whatever transform is used, the transform itself does not compress the source, and the coding step comes after the transform, when transform coefficients are first quantized then entropy-coded under a certain bit budget. Therefore, how to design good quantizers and entropy coders for transform coefficients are a principal focus in transform coder design today.

Wavelets are becoming the decomposition of choice for most applications and new standards for still image and video coding today. Wavelets provide excellent energy compaction and the variable time scales allow the various features of an image to be well-reproduced [258]. Other advantages include easy adaptive coding, as described in the following sections.

3) Bit Allocation: In practical transform-coding schemes, different approaches have been used to achieve the coding limits. Using optimal bit allocation, the number of quantization bits devoted to a component is determined based on the average energy of the component. A simple yet effective way to allocate the available code bits to different components is described in [232], where code bits are assigned to transform components bit by bit in a recursive way. At each stage, one bit is assigned to the component with the highest energy, then this highest energy is reduced by half before going on to the next stage. The procedure ends when all the available code bits are assigned to the components.

A more sophisticated bit-allocation scheme was proposed by Shoham and Gersho in [234]. Based on the reverse water-filling results, all the components should have the same quantization error except for those with energy lower than the quantization error. For an individual component, the slope of its rate-distortion function is just the reciprocal of the quantization level d [26]. Therefore, this allocation scheme tries to find the slope that minimizes the total distortion for the rate-distortion functions of all the components. However, to find this minimum distortion, this scheme becomes computationally intensive since it has to estimate the rate-distortion function for every transform component so that the best slope can be found. This approach sometimes can achieve optimal performance for a given set of coefficients and a fixed set of quantizers. More discussions on bit allocation can be found in [191].

Using fixed bit allocation, the number of bits used for each component is fixed for all sample functions of the source, so the encoder only needs to send out the allocation information once and all other code bits are used to encode the coefficient values. When such schemes are used for two-dimensional signals, they are also called zonal coding schemes [201], [232] because the coded coefficients are in fixed region(s) in the two-dimensional data plane.

Optimal bit allocation is totally dependent on the statistical characteristics of the source; specifically, the variances of transform components are needed and in general, the source has to be stationary. There are drawbacks to such coding schemes. To get accurate estimates of the variances, a reasonably large number of sample functions have to be processed before actual coding starts, which introduces encoding delay. Further, in real-world applications, random sources are rarely truly stationary—the statistics of transform coefficients change either spatially or temporally, whereas estimation over a large number of sample functions can reflect only the average behavior of the source. While producing constant-rate code sequences, coders using fixed bit allocation cannot adapt to the spatial or temporal changes of the source, and thus coding distortion may vary from sample function to sample function due to changes of the source.

To deal with the random changes of a source, adaptive schemes are used, and one very old, yet useful, scheme is the threshold method [177], [222], which is actually the basis of today's JPEG standard. Using a threshold, the coder can determine if a coefficient needs to be coded by comparing its energy with a threshold. If the energy of the coefficient is higher than the threshold, the coefficient will be encoded, otherwise, it will be treated as zero and discarded. As opposed to zonal coding which has to determine the optimal quantization level under a fixed code rate, threshold coding is actually easier to approach: once a threshold is determined, there is no need to do bit allocation. Since a large number of transform coefficients will be quantized to zero, this method can greatly reduce the number of coefficients to be coded and has the ability to adapt itself to changes of the source, since which coefficients are coded can change from sample function to sample function. The drawback is that there is no control over the code rate, since whether a coefficient is coded or not depends only on its own local energy. Such coders usually produce variable rate code sequences.

From sample function to sample function, which coefficients are coded can change due to nonstationarity of the source; therefore, information on which coefficients are coded for each sample function also must be provided to the decoder. Coding thus consists of two steps: one for the location of the coded coefficients, the other for their values.

We refer to the coefficient location information as side-information. For image coding, Ramchandran and Vetterli [224] proposed a thresholding method optimized in an operational rate-distortion sense that can be very effective in reducing the number of coefficients to be coded without sacrificing coding quality. In this method, whether a coefficient is coded or not depends not only on its local coefficient value with respect to a threshold, but also on the total cost of encoding a new coefficient. For each coefficient, the cost of coding is the total bits used for both the coefficient value and the coefficient location, and a decision strategy based on optimizing rate distortion performance for each data block is designed so that the coder can decide if a coefficient higher than the threshold is worth being coded. Therefore, this method is still a threshold-based coding scheme, but the focus is on how to reduce the number of coded coefficients without introducing significant error.

Although this method makes decisions in a rate-distortion sense, the statistical meaning of the rate-distortion function is lost. To calculate the coding cost, all data blocks are treated independently and the rate-distortion function of each data block is obtained as if each data block represented a different source [225]. The problem becomes how to merge all the different sources with rate-distortion optimality, and the basic idea is the same as in the optimal bit-allocation scheme described by Shoham and Gersho in [234], but in [234] the goal is to merge different transform components optimally, while here the goal is to merge different sample functions.

4) Side-Information and the Significance Map: In two-step coding schemes such as threshold coding, after determining which coefficients are to be coded, the encoder has to determine how to encode this information in addition to encoding the values of the chosen coefficients. A significance map is a representation of those transform coefficients with sufficient energy that they must be coded to achieve acceptable reconstructed signal quality. For transform coefficients of a sample function, and a fixed threshold t, a binary bitmap can be built to indicate which coefficients need to be coded. If a coefficient $|c_{ij}| > t$, then it is significant and will be encoded, so in the significance map, $b_{ij} = 1$, otherwise, $b_{ij} = 0$, indicating that the coefficient is not encoded. If a source can be decomposed into M components, then there are a total 2^M different patterns for the bitmap.

To encode the significance map, some practical coders make certain assumptions on the distribution of the significant coefficients. In threshold coding methods such as JPEG, to encode the significance map, a predetermined Huffman coder is used to encode the distance between two consecutive significant coefficients. The Huffman coder is designed based on the distributions of those distances obtained in experiments, such as was done in [177]. Since they are only experimental results, the coder may work very well for some images, but it is also possible that it may perform poorly for images with different statistical characteristics.

Another approach to coding both the significance map and the coefficient values is Shapiro's Embedded Zerotree coding method [233] based on the self-similarity assumption on wavelet transform coefficients. Shapiro's method is also called the EZW algorithm, since the embedded zerotree is used on the coefficients of a Discrete Wavelet Transform (DWT). The self-similarity assumption says that if a coefficient at a coarse scale (i.e., low frequency) is insignificant, then all the coefficients at the same location at finer scales (i.e., higher frequencies) are likely to be insignificant too. This means the significance of higher frequency coefficients can be predicted by the significance of a lower frequency coefficient at the same location. Since DWT coefficients have a natural tree structure, this makes it possible to use a quadtree to encode the significance map and achieve impressive coding performance.

Several related coding schemes have also been used based on analogous ideas, such as Said and Pearlman's set partitioning algorithm [226] which is basically similar to the EZW algorithm, in that they are all based on the self-similarity assumption, thus making these methods limited to certain types of transformations, such as the DWT.

In their three-dimensional (3-D) subband coding scheme, Taubman and Zakhor [238] used a more general approach to encoding the positions of coefficients, or the significance map. They tried to exploit the spatial correlation between coefficients to improve coding efficiency. Other approaches to encoding the significance map have also been attempted [214]. Although no statistical assumption is necessary, like all VQ schemes, this approach needs a training phase before it starts coding.

For image and video compression standards set in the last 10 years, the two-dimensional DCT is almost ubiquitous, appearing in the JPEG, H.261, MPEG1, MPEG2, H.263, and MPEG4 standards [184], [194]. Although bit-allocation methods drawing upon rate distortion theoretic results have been suggestive, many of the bit-allocation methods in the standards are based upon off-line perceptual experiments. The results are striking in that simple, uniform scalar quantizers can generate excellent perceptual results at rates of 0.5 bit/pixel and above. Lossless coding techniques, including Huffman coding and arithmetic coding, are important components of these standards as well.

Evolving standards, such as JPEG-2000 and MPEG4, have wavelet-based decompositions in place of or in addition to the DCT [184], [254].

C. High-Quality Audio Compression

Compression for high-quality audio is most often for playback applications that do not need real-time encoding; hence, relatively complicated techniques can be used for the encoding step. The basic approach has been to separate the input source material into blocks of time-domain samples and then decompose these samples into frequency-domain components for encoding. The importance of this approach is that results from auditory masking experiments in terms of the frequency-domain characteristics of the ear are available and can be incorporated in the distortion measure during the encoding process. Thus this method exhibits the concept of decomposing the source into several independent sources that are to be encoded subject to an overall limitation on rate. The distortion measure to be minimized in this case is very much a perceptual one and the achievement of the desired rate with the smallest audible distortion is done by iterative bit allocations until certain masking criteria are satisfied. Lossless coding techniques are also routinely employed.

Note that the approach for compression of high-quality audio is to devise a structure such that transparent perceptual quality is obtained, and whatever bit rate is necessary to achieve that goal is accepted (up to a point). Thus this compression problem is very much a rate-distortion problem—that is, minimize the rate for a specified distortion (perceptually transparent)—as opposed to a distortion-rate problem, as in many speech-coding applications [194], [255].

X. Recurring Themes

The influence of rate-distortion theory on lossy source coding can be seen in a few recurring themes for the optimization of specific source coders. The most common is to develop the operational rate distortion or distortion rate function for a particular source, source coder, and distortion measure, and then consider the constrained optimization problem that results by appending the appropriate rate or distortion constraint. The basis for this approach lies in considering Nth-order rate-distortion theory.

A. Nth-Order Rate Distortion Theory and Constrained Optimization

Let X^N denote the input source vector $(X_1, X_2, \cdots, X_N)$ and let its reconstruction be denoted by $\hat{X}^N$. The distortion between X^N and $\hat{X}^N$ is $d_N(X^N, \hat{X}^N)$ so that the average distortion over all source vectors and reproductions is given by

$$D_N = \frac{1}{N} E\{d_N(X^N, \hat{X}^N)\}.$$

The Nth-order distortion rate function can then be written as

$$D_N(R) = \inf_{p(\hat{X}^N \mid X^N)} \left\{ \frac{1}{N}[d_N(X^N, \hat{X}^N)] \mid \frac{1}{N} I(X^N; \hat{X}^N) \le R \right\}$$

and asymptotically in blocklength

$$D(R) = \lim_{N \to \infty} D_N(R).$$

To find $D_N(R)$, we append the rate constraint with a Lagrange multiplier and minimize the functional

$$J(p(\hat{X}^N \mid X^N)) = E[d_N(X^N, \hat{X}^N)] + \lambda I(X^N; \hat{X}^N).$$

Let us define the length of the codeword that represents X^N to be $\ell_N(X^N)$ and so the average rate in bits per source symbol is

$$R_N = \frac{1}{N} E\{\ell_N(X^N)\}.$$

Then, for a given source encoder $\alpha_N(X^N)$, and decoder $\beta_N(X^N)$ that yields rate R_N and reconstruction $\hat{X}^N$, we can write the operational distortion rate function as

$$\hat{D}_N(R) = \inf_{\alpha_N, \beta_N} \left\{ \frac{1}{N} E[d_N(X^N, \hat{X}^N)] \mid \frac{1}{N} E[\ell_N(X^N)] \leq R \right\}.$$

$D_N(R)$ lower-bounds $\hat{D}_N(R)$ and the bound becomes tight as $N \to \infty$. We can pose a constrained optimization problem using the operational distortion rate function as

$$J(\alpha_N, \beta_N) = E[d_N(X_N, \hat{X}_N)] + \lambda E[\ell_N(X^N)]. \qquad (18)$$

Since the operational distortion rate function is not necessarily convex or continuous, Lagrangian methods will not find $\hat{D}_N(R)$, however, we can use the Lagrangian formulation to find the convex hull.

Thus the approach is to iteratively minimize the functional in (18) using an algorithm similar to the generalized Lloyd method used for VQ design [191], [196].

B. Duality

An underutilized concept in obtaining lossy source compression methods is that of duality. Error-control coding and source coding are dual problems in the following sense: Decoding error-control codes consists of finding the best match to a received sequence given certain assumptions, a distortion criterion, and models. Alternatively, encoding for source compression entails the same steps. Further, decoding in source compression consists of receiving a particular transmitted sequence and mapping it into a unique output. Similarly, encoding for error-control coding maps a presented input directly into a particular transmitted sequence.

The development of trellis-coded quantization (TCQ) was spurred by this duality observation based upon results on trellis-coded modulation. In addition to providing good performance for speech coding at 16 kbits/s [253], TCQ is part of the vertification model of JPEG-2000 at the time of this writing. In fact, TCQ combined with wavelets was the top-ranked coder in both objective and subjective performance at 0.125 and 0.25 bit/pixel (bpp) during the JPEG-2000 evaluations [254].

XI. Research Challenges

A. Joint Source/Channel Coding

A fundamental result of information theory is that, assuming stationarity, optimal source coding and channel coding can be considered separately without loss of optimality. There are two caveats to this statement: First, separating source and channel coding may be more complex than a combined design [167], [207]; and second, both source and channel coding must be performing optimally, because if one is suboptimal, the other may be aided by incorporating the knowledge of this suboptimality.

Practitioners of lossy source coding for communications applications have always implemented coders that are robust to channel errors to some degree, with some attributable loss in source compression performance in the error-free case. This robustness is often obtained in waveform coding of speech by simply fading the memory of the encoder and decoder to "forget" channel errors and thus resynchronize the encoder and decoder adaptation. Another common approach to resychronizing the source encoder and decoder in video-compression applications is to transmit an intracoded frame (no motion compensation) at some specified interval. For example, this happens every 132nd frame in the H.320 video conferencing standard and is accomplished in the MPEG1 and MPEG2 standards with I-frames. However, the I-frames in MPEG were inserted primarily for search-motivated applications more than error resilence.

Another way to achieve error robustness without implementing error-correction codes is to use natural source redundancy and/or models of the channel to detect and correct errors. For example, Sayood *et al.* [231] exploits known Markov properties of the source in an MAP search for the best match to a received sequence. Phamdo and Farvardin [220] take a similar approach.

For a given transmitted bit rate, splitting bits between source coding and channel coding has the expected result—namely, if bits are allocated to channel coding and the channel is ideal, there is a loss in performance compared to source coding alone. Similarly, if there are no bits allocated to channel coding and the channel is very noisy, there will be a loss in performance compared to using some error-protection coding. Numerous studies for speech, image, and video coding have investigated joint source–channel coding. These solutions specify the allocation of transmitted rate between source coding and channel coding for chosen sources, source compression methods, and channel models to achieve the best source reconstruction.

For many applications today, of which wireless communications is a prime example, channels are far from ideal and it is best to combine source and channel coding. The most common way this is evident in standards is by the use of unequal error protection (UEP). That is, some compressed source bits can have a much more profound effect on reconstructed source quality than others, so these bits must be error-protected. Thus the source and channel coding is joint in the sense that the channel coding uses knowledge of the source bit sensitivity as well as the channel, and that the source compression frees up a portion of the bit rate for the error protection function.

It is a recent trend in wireline and wireless applications to sense the quality of the channel or the channel SNR versus frequency by sending known sequences or tones and then using the channel quality information at the transmitter to optimize digital communications system performance. Examples of this method are precoding in V.34 modems, DMT-based ADSL modems, and SNR estimation in the IS-127 mobile standard. This same technique can be extended to joint source/channel coding where we could use channel quality measurements to determine how to partition the available transmitted bit rate between source and channel coding.

B. Background Impairments

One of the principal challenges to mobile speech compression today is the presence of unwanted sounds or other speakers at the handset or microphone input. In order for the speech coders to achieve the desired reconstructed quality at the low rates needed, the speech coders have incorporated source-specific and sink-specific models in the encoder. These models are based on the assumption that what is present at the source coder input is the source to be encoded, and the source alone. When other sounds are present, the source coder forces these assumptions on the input signal during the encoding process with sometimes disastrous results.

More specifically, users of voice communications devices are somewhat forgiving of naturally occurring sounds, but when the speech coder attempts to use its assumed models on signals that are not speech, the results of coding natural sounds may be unnatural sounding artifacts upon reconstruction. The usual approaches today are either to filter the incoming signal or to attempt to cancel unwanted signals at the input. Under appropriate assumptions, the filtering approach may be optimal.

Dobrushin and Tsybakov [55], Wolf and Ziv [56], and Berger [26] have investigated the mean-squared error encoding of a source with additive distortion. The general result is that, asympotically in blocklength, the optimal encoder/decoder pair consists of an optimal estimator followed by optimal encoding of the resulting estimate. An application and extension of this work is reported by Fischer, Gibson, and Koo [188], where results are presented for training mode vector quantization and speech sources. Gibson, Koo, and Gray [193] develop optimal filtering algorithms for additive colored noise with applications to speech coding. One of their algorithms is the optional noise canceller in the Japanese half-rate digital cellular standard. Neither filtering nor cancellation is entirely effective.

C. Error Concealment

When channel errors cannot be corrected, lossy source compression techniques depend on robustness properties of the source decoder to reconstruct an approximation of the source without catastrophic distortions. However, if entire frames or packets are lost, special modifications are required. Twenty-five years ago, when such modifications were first considered, they were labeled with the perhaps misleading term, soft-decision demodulation. Today, these modifications are called error-concealment techniques.

Error-concealment methods generally consist of estimation or interpolation techniques using decoded signals that had been received previously. In speech coding for mobile radio applications, when a frame is lost, the lost frame is often compensated for by repeating the data from the preceding frame along with some muting of the reconstructed speech.

In many image- and video-compression applications, the need for error concealment arises due to the loss of a block of data, such as the coded coefficients representing a block of pixels as in transform coding. For these situations, error concealment can be performed in the transform domain or in the pixel domain, using adjacent blocks.

Video applications that have low transmission rates can have the data for an entire frame in one packet. A lost packet in these situations requires temporal interpolation.

D. Variable-Rate Coding

In order to respond to the changing characteristics of the input source and hence be efficient in the utilization of the available bandwidth, there is a trend toward variable-rate coding of speech and video. The challenges here are to sense the changes in the source and adapt the source coder to these changes, and to make the variable-rate stream interoperate with the possibly fixed-rate transmission channel. Of course, the use of buffering to interface fixed-to-variable length lossless source codes to the channel is common; however, rate variations in these new lossy schemes can have a wide swing and hence amplify the challenges.

Variable-rate coders for speech and images have been studied for 25 years [206], [239]–[241], but key rate indicators are still difficult to determine. Rate indicators that have been used range from simple input energy calculations to measuring correlation or other spectral properties of the source, such as estimates of source spectral entropy [210]. It is expected that variable-rate coders will be the rule rather than the exception in future applications and standards.

E. Layered Coding or Scalability

To respond to changing network conditions, such as available bit rate or channel congestion, there is another clear trend toward layered or scalable compression schemes. The principal concept in scalability is that an improvement in source reproduction, namely, reduced distortion, can be achieved by sending only an incremental increase in rate over the current transmitted rate that is achieving a coarser reproduction of the source. SNR, spatial, and temporal scalability are all important in applications. It is evident that a source-compression method designed to operate at several lower rates cannot outperform the compression method designed for the overall total rate, so the question is when do optimal or near-optimal scalable compression methods exist?

SNR scalability has been addressed from the rate-distortion theory viewpoint by Koshelev [262]–[264] who called it divisibility, and by Equitz and Cover [133] under the heading of successive refinement of information. Equitz and Cover address the problem of starting out with a low rate but optimal source coder, that is, one that operates exactly on the rate-distortion bound, and then finding those conditions under which an incremental addition in rate also yields rate-distortion optimal encoding. It is shown that successive refinement in the rate-distortion optimal sense is not always possible and that a neessary and sufficient condition for successive refinement is that the individual encodings be expressible as a Markov chain. Rimoldi [265] generalizes these results and provides an insightful interpretation in terms of typical sequences.

Spatial and temporal scalability is nonstandard in terms of classical discrete-time rate-distortion theory since both involve changes in the underlying sampling rate. To address spatial and temporal scalability or layered coding, many researchers pose the operational rate distortion problem for their particular coder and optimize with respect to the convex hull of the performance curves.

Progressive coding has become important in image coding, since in a network environment, different users may have different access capability, such as different bandwidths, CPU power, etc., and may want to access the source at different levels of quality. In such circumstances, a coder that can provide a coded sequence in a progressive way has an advantage.

In transform coding, progressive coding can be accomplished in two basic ways: spectral selection and successive approximation. For example, in DCT-based image coders, an encoder using a spectral selection strategy can first encode all the dc coefficients, then the ac coefficients in the low-frequency region, and finally the high-frequency ac coefficients. Since for many common images, most activity is concentrated in the low-frequency area, if only limited code bits can be received, the decoder can still reconstruct the image at a lower quality using all the dc coefficients and some low-frequency ac coefficients. This is useful in browsing applications when a user only wants to get a rough picture of an image to decide if the selected image is the one needed.

The prioritized DCT method [197], [198] is a more advanced approach based on the same idea. In a prioritized DCT coder, the transmission order is determined by the coefficient energy, that is, coefficients with higher energy, i.e., containing more information, are transmitted first. This is intuitively quite straightforward, since the idea of transmitting the dc coefficients first in the above mentioned scheme is based on the observation that most of the time dc coefficients have the highest energy among all the transform coefficients. The prioritized DCT method adds some flexibility to the same strategy in the sense that the coder can decide which coefficients are to be transmitted first based on the actual values of the coefficients, instead of assuming that the dc coefficients and low-frequency ac coefficients will have higher energy. This is also an adaptive-coding scheme.

Another powerful progressive coding scheme is successive approximation. Instead of transmitting the low-frequency coefficients to their highest accuracy, the successive approximation method first sends only the most significant bits for all of the coefficients, then sends the next most significant bits, and so on. In contrast to spectral selection, which generates minimum distortion for selected coefficients but discards all of the other coefficients, successive approximation produces relatively constant distortion for all the coefficients, which is closer to the rate-distortion result.

Examples of coders that use successive approximation are the Embedded Zerotree algorithm (EZW) by Shapiro [233], and the modified version of the EZW algorithm proposed by Said and Pearlman [226]. In both methods, the Discrete Wavelet Transform (DWT) coefficients are encoded. One of the novelties of the two coders is the way the coder arranges the order of the DWT coefficients that enables the coder to efficiently encode the side-information as well as the coefficient values, as already discussed in Section IX-B2. A similar approach was also studied by Xiong *et al.* in a DCT-based image coder [249]. A modified version of the prioritized DCT scheme is proposed by Efstratiadis and Strintzis [185], in which DWT coefficients are considered and instead of using a spectral selection strategy, this coder uses successive approximation to implement a hierarchical image coder.

Directly encoding DCT coefficients by layers can also be found in the literature [203]. Bit-plane encoding offers such easy functionality for progressive coding that it is widely adopted in new applications [254].

F. *Multiterminal Source Coding*

We have already noted the results on successive refinement of information (or divisibility) by Equitz and Cover [133], Koshelev [262]–[264], and Rimoldi [265] in Section XI-E, and their relationship to SNR scalability. Another multiterminal rate-distortion theoretic result that is finding applications in lossy source coding is the multiple descriptions problem [148], [150]. In this problem, the total available bit rate is split between (say) two channels and either channel may be subject to failure. It is desired to allocate rate and coded representations between the two channels, such that if one channel fails, an adequate reconstruction of the source is possible, but if both channels are available, an improved reconstruction over the single-channel reception results. Practical interest in this problem stems from packet-switched networks where the two channels can be realized by sequences of separately marked packets, and from diversity implementations in wireless applications. For recent results, see [266] and [267].

XII. Standards

Standards-setting for compression of speech, high-quality audio, still images, and video has been a dominant force in compression research since the mid-1980's. Although some might criticize these standards activities as inhibiting research and stifling innovation, most would agree that these efforts have generated an incredible interest in lossy compression and have lead to extraordinary advances in performance. The principal effect on lossy compression research is to make the research problem multifaceted in that not only must compression rate versus distortion performance be evaluated, but background impairments, channel errors, implementation complexity, and functionality (such as scalable coding, searching, and backwards compatibility) also become important considerations for many applications.

A challenge for researchers is to define the problem well and to fold as many of these other constraints into the problem as necessary to address the application of interest. Because of the tremendous emphasis on standards, it is perhaps most important for those involved in basic research to avoid being limited by current trends and the constraints of the many standards in order to generate the new results and directions needed for substantial advances in performance.

For more details on lossy compression techniques and standards, the reader is referred to [179], [184], [194], and [254]–[257].

XIII. Epilogue

Rate-distortion theory and the practice of lossy source coding have become much more closely connected today than they were in the past. There is every reason to anticipate that a much tighter fusion of theory and practice will prevail in 2009 when we celebrate the fiftieth anniversary of Shannon's 1959 paper.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423; 623–656, July and Oct. 1948. (Also in *Claude Elwood Shannon: Collected Papers*, N. J. A. Sloane and A. D. Wyner, Eds. Piscataway, NJ: IEEE Press, 1993, pp. 5–83.)

[2] ———, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Conv. Rec.*, vol. 7, 1959, pp. 142–163. (Also in *Information and Decision Processes*, R. E. Machol, Ed. New York: McGraw-Hill, 1960, pp. 93–126, and in *Claude Elwood Shannon: Collected Papers*, N. J. A. Sloane and A. D. Wyner, Eds. Piscataway, NJ: IEEE Press, 1993, pp. 325–350.)

[3] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 102–108, 1956.

[4] ———, "The theory of transmission of information, plenary session of the Academy of Sciences of the USSR on the automization of production" (Moscow, USSR, 1956), *Izv. Akad. Nauk SSSR*, pp. 66–99, 1957.

[5] ———, "A new metric invariant of transitive dynamic systems and automorphisms in Lebesgue spaces," *Dokl. Akad. Nauk. SSSR*, vol. 119, pp. 861–864, 1958.

[6] J. L. Holsinger, "Digital communication over fixed time-continuous channels with memory—With special application to telephone channels," Sc.D. dissertation, Dept. Elec. Eng., MIT, Cambridge, MA (Tech. Rep. TR 366, Lincoln Labs., Lexington, MA), 1968.

[7] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[8] Y. G. Sinai, "On the concept of entropy of a dynamical system," *Dokl. Akad. Nauk. SSSR*, vol. 124, pp. 768–771, 1959.

[9] D. S. Ornstein, "Bernoulli shifts with the same entropy are isomorphic," *Adv. Math.*, vol. 4, pp. 337–352, 1970.

[10] E. C. Posner and E. R. Rodemich, "Epsilon entropy and data compression," *Ann. Math. Statist.*, vol. 42, pp. 2079–2125, 1971.

[11] R. J. McEliece and E. C. Posner, "Hiding and covering in a compact metric space," *Ann. Statist.*, vol. 1, pp. 729–739, 1973.

[12] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, 1949.

[13] M. S. Pinsker, "Mutual information between a pair of stationary Gaussian random processes," *Dokl. Akad. Nauk. USSR*, vol. 99, no. 2, pp. 213–216, 1954.

[14] ———, "Computation of the message rate of a stationary random process and the capacity of a stationary channel," *Dokl. Akad. Nauk. USSR*, vol. 111, no. 4, pp. 753–756, 1956.

[15] R. M. Gray and L. D. Davisson, "Source coding theorems without the ergodic assumption," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 625–636, 1974.

[16] J. A. Bucklew, "The source coding theorem via Sanov's theorem," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 907–909, 1987.

[17] J. C. Kieffer, "A survey of the theory of source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1473–1490, 1993.

[18] ———, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 674–682, 1978.

[19] ———, "On the minimum rate for strong universal block coding of an class of ergodic sources," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 693–702, 1980.

[20] ———, "A method for proving multiterminal source coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 565–570, 1981.

[21] ———, "Fixed-rate encoding of nonstationary information sources," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 651–655, 1987.

[22] ———, "Strong converses in source coding relative to a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 37, pp. 257–262, 1991.

[23] ———, "Sample converses in source coding theory," *IEEE Trans. Inform. Theory*, vol. 37, pp. 263–268, 1991.

[24] T. J. Goblick Jr., "A coding theorem for time-discrete analog data sources," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 401–407, May 1969.

[25] T. Berger, "Rate-distortion theory for sources with abstract alphabets and memory," *Inform. Contr.*, vol. 13, pp. 254–273, 1968.

[26] ———, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[27] F. Jelinek, *Probabilistic Information Theory*. New York: McGraw-Hill, 1968.

[28] A. M. Gerrish, "Estimation of information rates," Ph.D. dissertation, Dept. Elec. Eng., Yale Univ., New Haven, CT, 1963.

[29] T. J. Goblick Jr., "Coding for a discrete information source with a distortion measure," Ph.D. dissertation, Dept. Elec. Eng., MIT, Cambridge, MA, 1962.

[30] T. Berger and W. C. Yu, "Rate-distortion theory for context-dependent fidelity criteria," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 378–384, May 1972.

[31] V. Erokhin, "ϵ-entropy of a discrete random variable," *Theory Probab. Its Applic.*, vol. 3, pp. 97–101, 1958.

[32] R. Pilc, "Coding theorems for discrete source-channel pairs," Ph.D. dissertation, Dept. Elec. Eng., MIT, Cambridge, MA, 1967.

[33] ———, "The transmission distortion of a discrete source as a function of the encoding block length," *Bell Syst. Tech. J.*, vol. 47, pp. 827–885, 1968.

[34] Z. Zhang, E. H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion—Part I: Known statistics," *IEEE Trans. Inform. Theory*, vol. 43, pp. 71–91, Jan. 1997.

[35] E.-h. Yang and Z. Zhang, "The redundancy of universal fixed rate source coding," presented at the 1998 IEEE Int. Symp. Information Theory, MIT, Cambridge, MA, Aug. 16–21, 1998.

[36] ———, "Abstract alphabet source coding theorem revisited: Redundancy analysis," presented at the 1998 IEEE Int. Symp. Information Theory, MIT, Cambridge, MA, Aug. 16–21, 1998.

[37] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1728–1740, 1994.

[38] ———, "Fixed-rate universal lossy source coding and rates of convergence for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 41, pp. 665–676, 1995.

[39] J. T. Pinkston, "Information rates of independent sample sources," M.S. thesis, Dept. Elec. Eng., MIT, Cambridge, MA, 1966.

[40] ———, "Encoding independent sample sources," Ph.D. dissertation, Dept. Elec. Eng., MIT, Cambridge, MA, 1967.

[41] ———, "An application of rate-distortion theory to a converse to the coding theorem," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 66–71, 1969.

[42] R. A. McDonald, "Information rates of Gaussian signals under criteria constraining the error spectrum," D. Eng. dissertation, Yale Univ. School Elec. Eng., New Haven, CT, 1961.

[43] R. A. McDonald and P. M. Schultheiss, "Information rates of Gaussian signals under criteria constraining the error spectrum," *Proc. IEEE*, vol. 52, pp. 415–416, 1964.

[44] ———, "Effective bandlimits of Gaussian processes under a mean square error criterion," *Proc. IEEE*, vol. 52, p. 517, 1964.

[45] J. J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian variables," *IRE Trans. Commun.*, vol. CS-11, pp. 289–296, 1963.

[46] H. A. Spang and P. M. Schultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun.*, vol. CS-12, pp. 373–380, 1964.

[47] B. Bunin, "Rate-distortion functions for correlated Gaussian sources," Ph.D. dissertation, Dept. Elec. Eng., Polytech. Inst. Brooklyn, Brooklyn, NY, 1969.

[48] J. P. M. Schalkwijk and L. I. Bluestein, "Transmission of analog waveforms through channels with feedback," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 617–619, 1967.

[49] J. K. Omura, "Optimum linear transmission of analog data for channels with feedback," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 38–43, 1968.

[50] S. Butman, "Rate-distortion over bandlimited feedback channels," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 110–112, 1971.

[51] B. S. Tsybakov, "ϵ-entropy of a vector message," presented at the IEEE Int. Symp. Information Theory, Ellenville, NY, 1969.

[52] M. S. Pinsker, "Sources of messages," *Probl. Pered. Inform.*, vol. 14, pp. 5–20, 1963.

[53] R. L. Dobrushin, "Individual methods for the transmission of information for discrete channels without memory and messages with independent components," *Sov. Math.*, vol. 4, pp. 253–256, 1963.

[54] ———, "Unified methods for transmission of information: The general case, *Sov. Math.*, vol. 4, pp. 289–292, 1963.

[55] R. L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 293–304, 1962.

[56] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 406–411, 1970.

[57] Y. N. Lin'kov, "Evaluation of ϵ-entropy of random variables for small ϵ," *Probl. Inform. Transm.*, vol. 1, pp. 12–18 (Russian pp. 12–18), 1965.

[58] C. R. Davis and M. E. Hellman, “On tree coding with a fidelity criterion,” *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 373–378, 1975.

[59] F. Jelinek, “Evaluation of distortion rate functions for low distortions,” *Proc. IEEE*, vol. 55, pp. 2067–2068, 1967.

[60] ———, “Tree encoding of memoryless time-discrete sources with a fidelity criterion,” *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 584–590, 1969.

[61] F. Jelinek and J. B. Anderson, “Instrumentable tree encoding of information sources,” *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 118–119, 1971.

[62] J. A. Vanderhorst, “The error locator polynomial for binary and primitive triple error correcting BCH codes,” M.S. thesis, School Elec. Eng., Cornell Univ., Ithaca, NY, Sept. 1971.

[63] ———, “Complete decoding of some binary BCH codes,” Ph.D. dissertation, School Elec. Eng., Cornell Univ., Ithaca, NY, Aug. 1972.

[64] W. E. Toms and T. Berger, “Information rates of stochastically driven dynamic systems,” *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 113–114, 1971.

[65] R. J. Dick, “Tree coding for Gaussian sources,” Ph.D. dissertation, School Elec. Eng., Cornell Univ., Ithaca, NY, May 1973. (See also R. J. Dick, T. Berger, and F. Jelinek, *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 332–336, 1974.)

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

[67] T. Berger, “Information rates of wiener sequences,” presented at the IEEE Int. Symp. Information Theory, Ellenville, NY, Jan. 28–31, 1969.

[68] ———, “Information rates of Wiener processes,” *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 134–139, 1970.

[69] ———, “The source coding game,” *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 71–76, 1971.

[70] ———, “Nyquist’s problem in data transmission,” Ph.D. dissertation, Harvard Univ., Div. Eng. App. Phys., Harvard Univ., Cambridge, MA, 1966.

[71] H. Gish, “Optimum quantization of random sequences,” Ph.D. dissertation, Harvard Univ., Div. Eng. Appl. Phys., Harvard Univ., Cambridge, MA, 1966.

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

[73] Private communication, 1967.

[74] D. J. Sakrison, “Source encoding in the presence of a random disturbance,” *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 165–167, 1968.

[75] ———, “A geometric treatment of the source encoding of a Gaussian random variable,” *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 481–486, 1968.

[76] ———, “The rate-distortion function of a Gaussian process with a weighted square error criterion,” *IEEE Trans. Inform. Theory*, Addendum 610–611, vol. IT-14, pp. 506–508, 1968.

[77] D. J. Sakrison and V. R. Algazi, “Comparison of line-by-line and two-dimensional coding of random images,” *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 386–398, 1971.

[78] D. J. Sakrison, “The rate of a class of random processes,” *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 10–16, 1970.

[79] B. G. Haskell, “The computation and bounding of rate-distortion functions,” *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 525–531, 1969.

[80] J. F. Hayes, A. Habibi, and P. A. Wintz, “Rate-distortion function for a Gaussian source model of images,” *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 507–509, 1970.

[81] R. M. Gray, “Information rates of autoregressive sources,” Ph.D. dissertation, Elec. Eng., Univ. South. Calif., Los Angeles, CA, 1969.

[82] ———, “Information rates of autoregressive processes,” *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 412–421, 1970.

[83] ———, “Rate-distortion functions for finite-state finite-alphabet Markov sources,” *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 127–134, 1971.

[84] ———, “Information rates of stationary ergodic finite-alphabet sources,” *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 516–523, 1971; Correction, vol. IT-19, p. 573, 1973.

[85] R. E. Blahut, “Computation of channel capacity and rate-distortion functions,” *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, 1972.

[86] S. Arimoto, “An algorithm for calculating the capacity of an arbitrary discrete memoryless channel,” *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 14–20, 1972.

[87] I. Csiszár and G. Tusnady, “Information geometry and alternating minimization procedures,” in *Statistics and Decisions/Supplement Issue*, no. 1, E. J. Dudewicz, D. Plachky, and P. K. Sen, Eds. Munich, Germany: R. Oldenbourg Verlag, 1984, pp. 205–237. (Formerly entitled *On Alternating Minimization Procedures*, preprint of the Math. Inst. Hungarian Acad. Sci., no. 35/1981, 1981.)

[88] T. M. Cover, “Broadcast channels,” *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 2–14, 1972.

[89] P. Bergmans, “Random coding theorem for broadcast channels with degraded components,” *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 197–207, 1973.

[90] R. Ahlswede, “Multi-way communication channels,” in *Proc. 2nd. Int. Symp. Information Theory* (Tsahkadsor, Armenian SSR), 1971, pp. 23–52.

[91] H. Liao, “Multiple access channels,” Ph.D. dissertation, Dept. Elec. Eng., Univ. Hawaii, Honolulu, HI, 1972.

[92] R. M. Gray and A. D. Wyner, “Source coding for a simple network,” *Bell Syst. Tech. J.*, vol. 58, pp. 1681–1721, 1974.

[93] D. Slepian and J. K. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, 1973.

[94] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side-information at the receiver,” *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 1–11, 1976.

[95] T. Berger, “Multiterminal source coding,” in *The Information Theory Approach to Communications* (CISM Courses and Lectures, no. 229). Wien, New York: Springer-Verlag, 1977, pp. 171–231.

[96] J. Korner and K. Marton, “The comparison of two noisy channels,” in *Trans. Keszthely Colloq. Information Theory* (Hungarian National Academy of Sciences, Keszthely, Hungary, Aug. 8–12), pp. 411–423.

[97] ———, “Images of a set via two channels and their role in multi-user communications,” *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 751–761, 1977.

[98] ———, “How to encode the modulo-two sum of binary sources,” *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 219–221, 1979.

[99] T. Berger and S. Y. Tung, “Encoding of correlated analog sources,” in *Proc. 1975 IEEE–USSR Joint Work. Information Theory*. Piscataway, NJ: IEEE Press, Dec. 1975, pp. 7–10.

[100] S. Y. Tung, “Multiterminal rate-distortion theory,” Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1977.

[101] M. U. Chang, “Rate-distortion with a fully informed decoder and a partially informed encoder,” Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1978.

[102] A. Shohara, “Source coding theorems for information networks,” Ph.D. dissertation, Univ. Calif. Los Angeles, Tech. Rep. UCLA-ENG-7445, 1974.

[103] J. K. Omura and K. B. Housewright, “Source coding studies for information networks,” in *Proc. IEEE 1977 Int. Conf. Communications* (Chicago, Ill., June 13–15, 1977). New York: IEEE Press, 1997, pp. 237–240.

[104] T. Berger, K. B. Housewright, J. K. Omura, S. Y. Tung, and J. Wolfowitz, “An upper bound on the rate-distortion function for source coding with partial side information at the decoder,” *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 664–666, 1979.

[105] A. Sgarro, “Source coding with side information at several decoders,” *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 179–182, 1977.

[106] J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 337–343, 1977.

[107] ———, “Compression of individual sequences via variable-rate coding,” *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 337–343, 1978.

[108] R. Pasco, “Source coding algorithms for fast data compression,” Ph.D. dissertation, Stanford Univ., Stanford, CA, 1976.

[109] J. Rissanen, “Generalized Kraft inequality and arithmetic coding,” *IBM J. Res. Develop.*, vol. 20, p. 198–, 1976.

[110] ———, “Universal coding, information, prediction and estimation,” *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, 1984.

[111] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, “The context-tree weighting method: Basic properties,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, 1995.

[112] ———, “Context weighting for general finite-context sources,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 1514–1520, 1996.

[113] Y. Steinberg and M. Gutman, “An algorithm for source coding subject to a fidelity criterion based on string matching,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 877–886, 1993.

[114] Z. Zhang and V. K. Wei, “An on-line universal lossy data compression algorithm by continuous codebook refinement,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 803–821, 1996.

[115] ———, “An on-line universal lossy data compression algorithm by continuous codebook refinement, Part II: Optimality for ϕ-mixing source models,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 822–836, 1996.

[116] I. Sadeh, “Universal compression algorithms based on approximate string matching,” in *Proc. 1995 IEEE Int. Symp. Information Theory* (Whistler, BC, Canada, Sept. 17–22, 1995), p. 84.

[117] E. H. Yang and J. Kieffer, “Simple universal lossy data compression schemes derived from the Lempel–Ziv algorithm,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 239–245, 1996.

[118] E. H. Yang, Z. Zhang, and T. Berger, "Fixed-slope universal lossy data compression," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1465–1476, 1997.

[119] I. Kontoyiannis, "An implementable lossy version of the Lempel-Ziv algorithm—Part I: Optimality for memoryless sources," NSF Tech. Rep. 99, Dept. Statist., Stanford Univ., Stanford, CA, Apr. 1998.

[120] ——, "Asymptotically optimal lossy Lempel-Ziv coding," presented at IEEE Int. Symp. Information Theory, MIT, Cambridge, MA, Aug. 16–21, 1998.

[121] "A decomposition theorem for binary Markov random fields," *Ann. Probab.*, vol. 15, pp. 1112–1125, 1987.

[122] L. A. Bassalygo and R. L. Dobrushin, "ϵ-entropy of the random field," *Probl. Pered. Inform.*, vol. 23, pp. 3–15, 1987.

[123] C. M. Newman, "Decomposition of binary random fields and zeros of partition functions," *Ann. Probab.*, vol. 15, pp. 1126–1130, 1978.

[124] C. M. Newman and G. A. Baker, "Decomposition of ising model and the mayer expansion," in *Ideas and Methods in Mathematics and Physics,—In Memory of Raphael Hoegh-Krohn (1938–1988)*, S. Albeverio *et al*., Eds. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[125] T. Berger and Z. Ye, "ϵ-entropy and critical distortion of random fields," *IEEE Trans. Inform. Theory*, vol. 36, pp. 717–725, 1990.

[126] Z. Ye and T. Berger, "A new method to estimate the critical distortion of random fields," *IEEE Trans. Inform. Theory*, vol. 38, pp. 152–157, 1992.

[127] ——, *Information Measures for Discrete Random Fields*. Beijing, China: Chinese Acad. Sci., 1998.

[128] A. H. Kaspi and T. Berger, "Rate-distortion for correlated sources with partially separated encoders," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 828–840, 1982.

[129] T. Berger and R. W. Yueng, "Multiterminal source encoding with one distortion criterion," *IEEE Trans. Inform. Theory*, vol. 35, pp. 228–236, Mar. 1989.

[130] ——, "Multiterminal source encoding with encoder breakdown," *IEEE Trans. Inform. Theory*, vol. 35, pp. 237–244, 1989.

[131] T. Berger, S. Y. Shen, and Z. Ye, "Some communication problems of random fields," *Int. J. Math. Statist. Sci.*, vol. 1, pp. 47–77, 1992.

[132] Z. Zhang and T. Berger, "Multiple description source coding with no excess marginal rate," *IEEE Trans. Inform. Theory*, vol. 41, pp. 349–357, 1995.

[133] W. E. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, pp. 269–275, 1991. (See also W. E. Equitz and T. M. Cover, "Addendum to 'successive refinement of information'," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1465–1466, 1993.)

[134] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem," *IEEE Trans. Inform. Theory*, vol. 42, pp. 887–903, May 1996.

[135] H. Viswanathan and T. Berger, "The quadratic Gaussian CEO problem," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1549–1561, 1997.

[136] Y. Oohama, "The rate distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1057–1070, May 1998.

[137] E.-h. Yang, Z. Zhang, and T. Berger, "Fixed-slope universal lossy data compression," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1465–1476, Sept. 1997.

[138] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 511–523, 1975.

[139] K. M. Mackenthum Jr. and M. B. Pursley, "Strongly and weakly universal source coding," in *Proc. 1977 Conf. Information Science and Systems* (The Johns Hopkins University, Baltimore, MD, 1977), pp. 286–291.

[140] M. B. Pursley and K. M. Mackenthum Jr., "Variable-rate source coding for classes of sources with generalized alphabets," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 592–597, 1977.

[141] K. M. Mackenthum Jr. and M. B. Pursley, "Variable-rate universal block source coding subject to a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 349–360, 1978.

[142] H. H. Tan, "Tree coding of discrete-time abstract alphabet stationary block-ergodic sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 671–681, 1976.

[143] J. Ziv, "Coding of sources with unknown statistics—Part II: Distortion relative to a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 389–394, 1972.

[144] T. Hashimoto, "Tree coding of sources and channels," Ph.D. dissertation, Dept. Mech. Eng., Osaka Univ., Toyonaka, Osaka, Japan, 1981.

[145] A. Gersho and A. D. Wyner, "The multiple descriptions problem," presented by A. D. Wyner at the IEEE Information Theory Work., Seven Springs Conf. Ctr., Mt. Kisco, NY, Sept. 1979.

[146] H. S. Witsenhausen, "On source networks with minimal breakdown degradation," *Bell Syst. Tech. J.*, vol. 59, pp. 1083–1087, 1980.

[147] J. K. Wolf, A. D. Wyner, and J. Ziv, "Source coding for multiple descriptions," *Bell Syst. Tech. J.*, vol. 59, pp. 1417–1426, 1980.

[148] L. H. Ozarow, "On the source coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, pp. 1909–1922, 1980.

[149] H. A. Witsenhausen and A. D. Wyner, "Source coding for multiple descriptions II: A binary source," *Bell Syst. Tech. J.*, vol. 60, pp. 2281–2292, 1981.

[150] A. E. Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 851–857, 1982.

[151] T. Berger and Z. Zhang, "Minimum breakdown degradation in binary source encoding," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 807–814, 1983.

[152] R. Ahlswede, "The rate-distortion region for multiple descriptions without excess rate," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 721–726, 1985.

[153] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder—II: General sources," *Inform. Contr.*, vol. 38, pp. 60–80, 1978.

[154] H. Yamamoto, "Source coding theory for cascade and branching communication systems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 299–308, 1981.

[155] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 727–734, 1985.

[156] H. Yamamoto, "Source coding theory for a triangular communication system," *IEEE Trans. Inform. Theory*, vol. 42, pp. 848–853, 1996.

[157] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 502–521, July 1987.

[158] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1939–1952, 1996.

[159] S. L. Fix, "Rate-distortion functions for squared error distortion measures," in *Proc. 16th Annu. Allerton Conf. Communication, Control and Computers* (Monticello, IL, 1978).

[160] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 480–489, 1973.

[161] B. M. Leiner, "Rate-distortion theory for sources with side information," Ph.D. dissertation, Stanford Univ., Stanford, CA, Aug. 1973.

[162] B. M. Leiner and R. M. Gray, "Rate-distortion for ergodic sources with side information," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 672–675, 1974.

[163] R. M. Gray, D. L. Neuhoff, and J. K. Omura, "Process definitions of distortion rate functions and source coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 524–532, 1975.

[164] R. M. Gray, D. L. Neuhoff, and D. S. Ornstein, "Nonblock source coding with a fidelity criterion," *Ann. Probab.*, vol. 3, pp. 478–491, 1975.

[165] R. M. Gray, D. S. Ornstein, and R. L. Dobrushin, "Block synchronization, sliding-block coding, invulnerable sources and zero error codes for discrete noisy channels," *Ann. Probab.*, vol. 8, pp. 639–674, 1975.

[166] N. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Signal Processing*. New York: Springer-Verlag, 1975.

[167] T. C. Ancheta Jr., "Joint source channel coding," Ph.D. dissertation, Dept. Elec. Eng., Univ. Notre Dame, Notre Dame, IN, Aug. 1977.

[168] J. B. Anderson and J. B. Bodie, "Tree encoding of speech," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 379–387, July 1975.

[169] J. B. Anderson, "Recent advances in sequential encoding of analog waveforms," in *Conf. Rec., IEEE Nat. Telecommunications Conf.*, 1978, pp. 19.4.1–19.4.5.

[170] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *WESCON Tech. Papers*, 1968, paper 8/2.

[171] ——, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, vol. 49, pp. 1973–1986, Oct. 1970.

[172] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[173] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1982, pp. 614–617.

[174] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech at very low bit rates," in *Proc. IEEE Int. Conf. Communications*, 1984, pp. 1610–1613.

[175] D. W. Becker and A. J. Viterbi, "Speech digitization and compression by adaptive predictive coding with delayed decision," in *Conf. Rec. Nat. Telecommunications Conf.*, 1975, pp. 46-18–46-23.

[176] W. Chen, C. H. Smith, and S. Fralick, "A fast computational algorithm for the discrete cosine transform," *IEEE Trans.Commun.*, vol. COM-25, pp. 1004–1009, Sept. 1977.

[177] W.-H. Chen and W. K. Pratt, "Scene adaptive coder," *IEEE Trans. Commun.*, vol. COM-32, pp. 225–232, Mar. 1984.

[178] L. Chiariglione, “MPEG and multimedia communications,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 5–18, Feb. 1997.

[179] R. V. Cox, “Three new speech coders from the ITU cover a range of applications,” *IEEE Commun. Mag.*, vol. 35, pp. 40–47, Sept. 1997.

[180] C. C. Cutler, “Differential quantization of communications,” U.S. Patent 2 605 361, July 29, 1952.

[181] ——, “Delayed encoding: Stabilizer for adaptive coders,” *IEEE Trans. Commun.*, vol. COM-19, pp. 898–907, Dec. 1971.

[182] L. D. Davisson, “Rate-distortion theory and application,” *Proc. IEEE* , vol. 60, pp. 800–808, July 1972.

[183] L. D. Davisson and R. M. Gray, *Data Compression*. Dowden: Hutchinson & Ross, 1976.

[184] T. Ebrahimi and M. Kunt, “Visual data compression for multimedia applications,” *Proc. IEEE* , vol. 86, pp. 1109–1125, June 1998.

[185] S. N. Efstratiadis and M. G. Strintizis, “Hierarchical prioritized predictive image coding,” in *Proc. Int. Conf. Image Processing (ICIP ’96)* (Switzerland, Sept. 1996), pp. 189–192.

[186] P. Elias, “Predictive coding—Parts I and II,” *IRE Trans. Inform. Theory*, vol. IT-1, pp. 16–33, Mar. 1955.

[187] N. Farvardin and J. W. Modestino, “Optimal quantizer performance for a class of non-Gaussian memoryless sources,” *IEEE Trans. Inform. Theory*, vol. 30, pp. 485–497, 1984.

[188] T. R. Fischer, J. D. Gibson, and B. Koo, “Estimation and noisy source coding,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 23–34, Jan. 1990.

[189] R. D. LeGall, “MPEG: A video compression standard for multimedia applications,” *Commun. Assoc. Comput. Mach.*, vol. 34, pp. 60–64, Apr. 1991.

[190] R. G. Gallager, “Tree encoding for symmetric sources with a distortion measure,” *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 65–76, Jan. 1974.

[191] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992.

[192] I. A. Gerson and M. A. Jasiuk, “Vector sum excited linear prediction (VSELP) speech coding at 8 kbps,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Apr. 1994, pp. 461–464.

[193] J. D. Gibson, B. Koo, and S. D. Gray, “Filtering of colored noise for speech enhancement and coding,” *IEEE Trans. Signal Processing*, vol. 39, pp. 1732–1742, Aug. 1991.

[194] J. D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, and R. L. Baker, *Digital Compression for Multimedia: Principles and Standards*. San Francisco, CA: Morgan-Kaufmann, 1998.

[195] R. M. Gray, *Source Coding Theory*. Norwell, MA: Kluwer, 1990.

[196] R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE Trans. Inform. Theory*, this issue, pp. 2325–2383.

[197] Y. Huang, N. P. Galatsanos, and H. M. Dreizen, “Priority DCT coding for image sequences,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (Toronto, Ont., Canada, 1991), pp. 2629–2632.

[198] Y. Huang, H. M. Dreizen, and N. P. Galatsanos, “Prioritized DCT for compression and progressive transmission of images,” *IEEE Trans. Image Processing*, vol. 1, pp. 477–487, Oct. 1992.

[199] J. D. Irwin and J. B. O’Neal, “The design of optimum DPCM (Differential Pulse Code Modulation) encoding systems via the Kalman predictor,” *1968 JACC Preprints*, pp. 130–136, 1968.

[200] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[201] N. S. Jayant, Ed., *Waveform Quantization and Coding*. New York: IEEE Press, 1976.

[202] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communications Systems*. Chichester, U.K.: Wiley, 1994.

[203] J. Li and C.-C. J. Kuo, “An embedded DCT approach to progressive image compression,” in *Proc. Int. Conf. Image Processing (ICIP ’96)* (Switzerland, Sept. 1996), pp. 201–204.

[204] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[205] S. P. Lloyd, “Least squares quantization in PCM,” unpublished, Bell Lab. Tech. Note, 1957.

[206] D. T. Magill, “Adaptive speech compression for packet communication systems,” in *Conf. Rec. Nat. Telecommunications Conf.*, Nov. 1973, pp. 29D-1–29D-5.

[207] J. L. Massey, “Joint source channel coding,” in *Communication Systems and Random Process Theory*, J. K. Skwirzynski, Ed. Amsterdam, The Netherlands: Sijthoff and Nordhoff, 1978, pp. 279–293.

[208] B. Mazor and W. A. Pearlman, “A trellis code construction and coding theorem for stationary Gaussian sources,” *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 924–930, Nov. 1983.

[209] ——, “A tree coding theorem for stationary Gaussian sources and the squared-error distortion measure,” *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 156–165, Mar. 1986.

[210] S. McClellan and J. D. Gibson, “Variable-rate CELP based on subband flatness,” *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 120–130, Mar. 1997.

[211] J. W. Modestino and V. Bhaskaran, “Robust two-dimensional tree encoding of images,” *IEEE Trans. Commun.*, vol. COM-29, pp. 1786–1798, Dec. 1981.

[212] J. W. Modestino, V. Bhaskaran, and J. B. Anderson, “Tree encoding of images in the presence of channel errors,” *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 667–697, Nov. 1981.

[213] P. Noll, “MPEG digital audio coding,” *IEEE Signal Processing Mag.*, vol. 14, pp. 59–81, Sept. 1997.

[214] L. V. Oliveira and A. Alcaim, “Identification of dominant coefficients in DCT image coders using weighted vector quantization,” in *Proc. 1st Int. Conf. Image Processing (ICIP ’94)* (Austin, TX, Nov. 1994), vol. 1, pp. 110–113.

[215] B. M. Oliver, “Efficient coding,” *Bell Syst. Tech. J.*, vol. 31, pp. 724–750, July 1952.

[216] J. B. O’Neal Jr., “A bound on signal-to-quantizing noise ratios for digital encoding systems,” *Proc. IEEE* , vol. 55, pp. 287–292, Mar. 1967.

[217] W. A. Pearlman and P. Jakatdar, “A transform tree code for stationary Gaussian sources,” *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 761–768, 1985.

[218] D. E. Pearson, “A realistic model for visual communication systems,” *Proc. IEEE*, vol. 55, pp. 380–389, Mar. 1967.

[219] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*. New York: Van Nostrand Reinhold, 1988.

[220] N. Phamdo and N. Farvardin, “Optimal detection of discrete Markov sources over discrete memoryless channels—Applications to combined source-channel coding,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 186–193, Jan. 1994.

[221] J. R. Pierce, “The early days of information theory,” *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 3–8, Jan. 1973.

[222] W. K. Pratt, *Digital Image Processing*. New York: Wiley, 1978.

[223] K. Rao and P. Yip, *Discrete Cosine Transform*. Boston, MA: Academic, 1990.

[224] K. Ramchandran and M. Vetterli, “Rate-distortion optimal fast thresholding with complete JPEG/MPEG decoder compatibility,” *IEEE Trans. Image Processing*, vol. 3, pp. 700–704, Sept. 1994.

[225] ——, “Best wavelet packet bases in a rate-distortion sense,” *IEEE Trans. Image Processing*, vol. 2, pp. 160–175, Apr. 1993.

[226] A. Said and W. A. Pearlman, “A new, fast, and efficient image codec based on set partitioning in hierarchical trees,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.

[227] D. J. Sakrison, “A geometric treatment of the source encoding of a Gaussian random variable,” *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 481–486, 1968.

[228] ——, “Image coding applications of vision models,” in *Image Transmission Techniques*, W. K. Pratt, Ed. New York: Academic, 1979, pp. 21–51.

[229] R. Salami *et al.*, “Design and description of CS-ACELP: A toll quality 8 kb/s speech coder,” *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 116–130, Mar. 1998.

[230] R. Salami *et al.*, “ITU-T G.729 Annex A: Reduced complexity 8 kb/s CS-ACELP codec for digital simultaneous voice and data,” *IEEE Commun. Mag.*, vol. 35, pp. 56–63, Sept. 1997.

[231] K. Sayood *et al.*, “A constrained joint source/channel coder design,” *IEEE J. Select. Areas Commun.*, vol. 12, pp. 1584–1593, Dec. 1994.

[232] K. Sayood, *Introduction to Data Compression*. San Francisco, CA: Morgan-Kaufmann, 1996.

[233] J. M. Shapiro, “Embedded image coding using zerotrees of wavelet coefficients,” *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[234] Y. Shoham and A. Gersho, “Efficient bit allocation for an arbitrary set of quantizers,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1445–1453, Sept. 1988.

[235] A. S. Spanias, “Speech coding: A tutorial review,” *Proc. IEEE* , vol. 82, pp. 1541–1582, Oct. 1994.

[236] L. C. Stewart, “Trellis data compression,” Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, June 1981.

[237] L. C. Stewart, R. M. Gray, and Y. Linde, “The design of trellis waveform coders,” *IEEE Trans. Commun.*, vol. COM-30, pp. 702–710, Apr. 1982.

[238] D. Taubman and A. Zakhor, “Multirate 3-D subband coding of video,” *IEEE Trans. Image Processing*, vol. 3, pp. 572–588, Sept. 1994.

[239] A. G. Tescher and R. V. Cox, “Image coding: Variable rate differential pulse modulation through fixed rate channel,” *Proc. Soc. Photo-Optical Instr. Eng.*, vol. 119, pp. 147–154, Aug. 1977.

[240] A. G. Tescher, “Transform image coding,” in *Advances in Electronics and Electron Physics*. New York: Academic, 1979, pp. 113–155.

[241] V. R. Viswanathan *et al.*, “Variable frame rate transmission: A review of methodology and application to narrowband LPC speech coding,” *IEEE*

Trans. Commun., vol. COM-30, pp. 674–686, Apr. 1982.

[242] A. Viterbi and J. Omura, "Trellis encoding of memoryless discrete-time sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 325–332, May 1974.

[243] G. K. Wallace, "The JPEG still picture compression standard," *Commun. Assoc. Comput. Mach.*, vol. 34, pp. 30–44, Apr. 1991.

[244] Z. Wang, "Reconsideration of a fast computational algorithm for the discrete cosine transform," *IEEE Trans. Commun.*, vol. COM-31, pp. 121–123, Jan. 1983.

[245] S. G. Wilson and S. Husain, "Adaptive tree encoding of speech at 8000 bps with a frequency-weighted error criterion," *IEEE Trans. Commun.*, vol. COM-27, pp. 165–170, Jan. 1979.

[246] P. A. Wintz and A. J. Kurtenbach, "Waveform error control in PCM telemetry," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 640–661, Sept. 1968.

[247] J. W. Woods, Ed., *Subband Image Coding*. New York: Kluwer, 1991.

[248] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

[249] Z. Xiong, O. G. Guleryuz, and M. T. Orchard, "A DCT-based embedded image coder," *IEEE Signal Processing Lett.*, vol. 3, pp. 289–290, Nov. 1996.

[250] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 299–309, Aug. 1977.

[251] ——, "Approaches to adaptive transform speech coding at low bit rate," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 89–95, Feb. 1979.

[252] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[253] M. W. Marcellin, T. R. Fischer, and J. D. Gibson, "Predictive trellis coded quantization of speech," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 38, pp. 46–55, Jan. 1990.

[254] P. J. Sementilli, A. Bilgin, J. H. Kasner, and M. W. Marcellin, "Wavelet TCQ: Submission to JPEG-2000 (keynote address)," in *Proc. SPIE* (San Diego, CA, July 1998), vol. 3460.

[255] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Mag.*, vol. 14, pp. 59–81, Sept. 1997.

[256] T. Sikora, "MPEG digital video-coding standards," *IEEE Signal Processing Mag.*, vol. 14, pp. 82–100, Sept. 1997.

[257] R. V. Cox, B. G. Haskell, Y. LeCun, B. Shahraray, and L. Rabiner, "On the applications of multimedia processing to communications," *Proc. IEEE*, vol. 86, pp. 755–824, May 1998.

[258] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

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

[260] H. Dudley, "The vocoder," *Bell Lab. Rec.*, vol. 17, pp. 1221–1226, 1939.

[261] H. C. Woo and J. D. Gibson, "Low delay tree coding of speech at 8 kbits/s," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 361–370, July 1994.

[262] V. Koshelev, "Multilevel source coding and data-transmission theorem," in *Proc. VII All-Union Conf. Theory of Coding and Data Transmission* (Vilnius, USSR, 1978), pt. 1, pp. 85–92.

[263] ——, "Hierarchical coding of discrete sources," *Probl. Pered. Inform.*, vol. 16, pp. 31–49, 1980.

[264] ——, "An evaluation of the average distortion for discrete scheme of sequential approximation," *Probl. Pered. Inform.*, vol. 17, pp. 20–33, 1981.

[265] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. 40, pp. 253–259, Jan. 1994.

[266] V. Vaishampayan, J.-C. Batallo, and A. R. Calderbank, "On reducing granular distortion in multiple description quantization," presented at the 1998 IEEE Int. Symp. Information Theory, MIT, Cambridge, MA, Aug. 16–21, 1998.

[267] V. K. Goyal, J. Kovacevic, and M. Vetterli, "Multiple description transform coding: Robustness to erasures using tight frame expansions," presented at the 1998 IEEE Int. Symp. Information Theory, MIT, Cambridge, MA, Aug. 16–21, 1998.