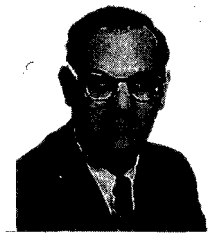
a separate digital expandor, multiplier, and compressor. The conversion algorithms follow directly from the digital expansion and compression algorithms developed previously [6]. Digital attenuators having arbitrary attenuation have been systematically synthesized using simple serial logic. Signal impairment attendant to this operation has been shown.

ACKNOWLEDGMENT

The authors are grateful to P. W. Osborne for discussions that helped to clarify our ideas.

REFERENCES

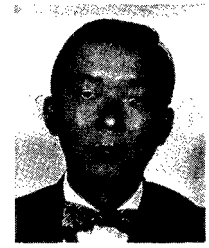
[1] B. Smith, "Instantaneous companding of quantized signals," *Bell Syst. Tech. J.*, vol. 36, no. 3, pp. 653–709, May 1957.
[2] H. H. Henning, "96-channel PCM channel bank," in *Conf. Rec., 1969 IEEE Int. Conf. Communications*, pp. 34.17–34.22.
[3] K. W. Cattermole, *Principles of Pulse Code Modulation*. London: Iliffe Books, 1969.
[4] L. B. Jackson, J. F. Kaiser, and H. S. McDonald, "An approach to the implementation of digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 413–421, Sept. 1968.
[5] A. Kundig, "Digital filtering in PCM telephone systems," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 412–417, Dec. 1970.
[6] H. Kaneko, "A unified formulation of segment companding laws and synthesis of codecs and digital compandors," *Bell Syst. Tech. J.*, vol. 47, no. 7, pp. 1555–1588, Sept. 1970.
[7] H. Kaneko and M. R. Aaron, "Digital conversion between segment companded PCM codes," to be published.
[8] W. L. Montgomery, "Six decibel digital attenuation," *IEEE Trans. Commun. Technol.*, vol.. COM-19, pp. 315–319, June 1971.
[9] Y. Chu, *Digital Computer Design Fundamentals*. New York: McGraw-Hill, 1962.
[10] J. D. Heightley, "Partitioning of digital filters for integrated circuit realization," this issue, pp. 1059–1063.



M. Robert Aaron (S'49–A'52–M'57–F'68) received the B.S.E.E. and M.S.E.E. degrees from the University of Pennsylvania, Philadelphia, Pa., in 1949 and 1951, respectively.

Since 1951 he has been with Bell Telephone Laboratories, Inc., where he has been involved in a variety of development projects on analog and digital communications systems. He is presently Department Head, responsible for exploratory work in digital techniques and associated systems.

Mr. Aaron is a fellow of the American Association for the Advancement of Science.



Hisashi Kaneko (S'60–M'62) was born in Tokyo, Japan, on November 19, 1933. He received the B.S. degree in electrical engineering from the University of Tokyo, Tokyo, in 1956, the M.S. degree in electrical engineering from the University of California, Berkeley, in 1962, and the Dr. Eng. degree from the University of Tokyo in 1967.

He joined the Nippon Electric Company Ltd. in 1956, where he worked at the Central Research Laboratories. From 1960 to 1962 he was a Research Assistant at the University of California under a Fulbright Scholarship. From 1968 to 1970 he worked at Bell Laboratories, Holmdel, N. J., on future digital channel banks. Currently he is responsible for work on PCM systems, digital communications systems, and delta modulation in the Central Research Laboratory of the Nippon Electric Company Ltd.

Dr. Kaneko is a member of the Institute of Electrical Communication Engineers of Japan and the Institute of Electrical Engineers of Japan.

A Survey of Coding Schemes for Transmission or Recording of Digital Data

HISASHI KOBAYASHI, MEMBER, IEEE

***Abstract*—In this survey we shall review coding techniques and results which pertain to such problems as reduction of dc wandering, suppression of intersymbol interference, and inclusion of self-clocking capability. These problems are of engineering interest in the transmission or recording of digital data. The topics to be discussed include: 1) dc free codes such as bipolar signals and feedback balanced codes, 2) correlative level codes and optimal decoding methods, 3) Fibonacci codes and run-length constraint codes, and 4) state-oriented codes.**

Manuscript received June 16, 1971; revised August 3, 1971.
The author is with the IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y. 10598.

I. INTRODUCTION

THE PRESENT paper is intended to review various coding schemes which have been developed and applied to the transmission or recording of digital data. The coding schemes to be discussed are those primarily developed for pulse-code modulation (PCM) systems, high-speed data communication systems, and high density magnetic recording systems in order to reduce dc wandering, suppress intersymbol interference, maintain self-clocking capability, and allow effective

error monitoring. We will show that these various techniques are quite similar in their underlying concepts.

In Section II we review the bipolar signal [2] used in PCM transmission systems and extend this signaling method to the multilevel case in two different forms. The feedback balanced codes of Kaneko and Sawai [5] are discussed in this context.

In section III we discuss the correlative level coding technique developed by Lender [6]–[8], [10], and others (often called the partial-response technique [14]–[16] which is adopted in a large number of high-speed data modems today. We also clarify the important analogy [17]–[20] between a digital magnetic recording system and a partial-response channel. Various processing techniques developed for partial-response modems are equally applicable to magnetic recording systems. Recent developments related to the technique will be brought to the reader's attention; that is, the maximum likelihod decoding (MLD) method [19], [22], [24] and the ambiguity decoding method [34]. Some recent work by Miyakawa and Harashima [37], [38] which extends the correlative level coding concept is also discussed.

Section IV discusses the problem of constructing optimal algebraic block codes subject to constraints on the maximum and minimum separation between successive changes in signal levels. The Fibonacci codes of Kautz [40] and the run-length limited codes of Tang [41], [42], and Tang and Bahl [44] will be reviewed.

Section V discusses a class of codes similar to those of Section IV, but the code generation techniques are based on a finite state machine model of the encoder. The important results due to Freiman and Wyner [46] are revisited, and more recent work by Gabor [52], Tang [43], and Franaszek [49], [53] is discussed.

II. DC Free Codes

For transmission of binary digital information over a line, the simplest code format is unipolar in which the binary symbols 1 and 0 are coded for transmission as presence and absence of pulses, respectively. There are three significant practical problems associated with this unipolar format (Sipress [1]). First, timing information must be extracted from the pulse train by regenerative repeaters. Transmission of long sequences of 0's results in long periods without timing information. Secondly, transmission of long sequences of 1's results in dc wander since the repeaters cannot be dc coupled to the cable medium, and dc restoration circuits are in general expensive. Thirdly, some technique for in-service performance monitoring is desirable. Performance monitoring of the line error rate with the unipolar format is impossible without inclusion of some redundant digits.

One of the simplest approaches is the bipolar code (Aaron [2]) used in the Bell System's T1 carrier PCM system. In bipolar, the binary symbol 0 is represented by no signal on the line, and the binary symbol 1 is represented alternately by positive and negative pulses. This coding method has the advantage of reducing the effects of dc wander, since a pulse of one polarity is certain to be followed eventually by a pulse of the opposite polarity.

The bipolar signal can be generated in various ways. Probably the simplest way is to use the binary input $\{a_k\}$ to drive a binary counter, and to control the polarity of the ternary output $\{c_k\}$, by the present state of the counter. Here we discuss two other methods that have the advantage of being easily generalized for m-ary alphabets. In Fig. 1 the input binary data $\{a_k\}$ is first "integrated" modulo 2. That is, the integrated output $\{b_k\}$ is related to $\{a_k\}$ by

$$\begin{aligned} b_k &= b_{k-1} \oplus a_k \\ &= a_0 \oplus a_1 \oplus \cdots \oplus a_{k-1} \oplus a_k \end{aligned} \tag{1}$$

where $\oplus$ means "modulo 2" addition. It will be clear that the sequence $\{b_k\}$ corresponds to the binary counting of $\{a_k\}$. The sequence $\{b_k\}$ is then passed into the "differential" circuit with a transfer function

$$G(D) = 1 - D \tag{2}$$

where D means a one unit delay. Then the output sequence $\{c_k\}$ given by

$$c_k = b_k - b_{k-1} \tag{3}$$

is a three-level sequence. Equation (1) can be written as

$$b_k - b_{k-1} = a_k \text{ modulo 2.} \tag{4}$$

Therefore, from (3) and (4) we have

$$c_k = a_k \text{ modulo 2.} \tag{5}$$

Thus the original binary signal can be reconstructed simply by rectifying the ternary signal $\{c_k\}$.

Another method of generating the bipolar signal is depicted in Fig. 2. Here $\{s_{k-1}\}$ represents the quantity in the feedback loop and corresponds to the running sum (or integration) of the past output $\{c_n: 1 \leq n \leq k-1\}$:

$$\begin{aligned} s_{k-1} &= s_{k-2} + c_{k-1} \\ &= \sum_{n=1}^{k-1} c_n \end{aligned} \tag{6}$$

$$s_0 = 0. \tag{7}$$

It is not difficult to see, by referring to (3), that the running sum $\{s_k\}$ is equivalent to $\{b_k\}$. The quantity sgn $\{1/2 - s_{k-1}\}$ controls the polarity of the next digit c_k so that $+1$ and -1 alternate in the output sequence $\{c_k\}$.

The power spectrum of the bipolar sequence $\{c_k\}$ is obtained as follows. If $\{a_k\}$ takes on 0's and 1's independently and with equal probability, so does the sequence $\{b_k\}$. Therefore the power spectrums of sequences $\{a_k\}$ and $\{b_k\}$ are flat, i.e.,

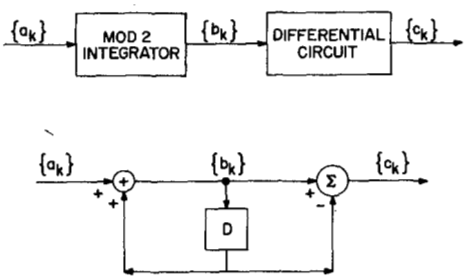


Fig. 1. Generation of bipolar signals (I).

$$P_a(\lambda) = P_b(\lambda) = \frac{1}{4} + \frac{\pi}{2}\,\delta(\lambda), \qquad -\pi \le \lambda \le \pi \tag{8}$$

where $\delta(\cdot)$ is the Dirac delta function, the power spectrum of a given sequence $\{x_k\}$ is defined by [3]

$$P_x(\lambda) = \sum_{k=\infty}^{\infty} R_x(k) \exp(-ik\lambda), \qquad -\pi \le \lambda \le \pi \tag{9}$$

and $R_x(k)$ is the autocorrelation function of the sequence $\{x_k\}$. The inverse transform of (9) is defined by

$$R_x(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_x(\lambda) \exp(+ik\lambda)\, d\lambda,$$

$$k = 0, +1, +2, \cdots. \tag{10}$$

The mapping from $\{b_k\}$ into $\{c_k\}$ is a linear transformation with a transfer function $G(D)$ of (2). Therefore the power spectrum of the sequence $\{c_k\}$ is given by

$$P_c(\lambda) = |G[\exp(-i\lambda)]|^2 P_b(\lambda)$$
$$= \tfrac{1}{4}\,|1 - \exp(-i\lambda)|^2 = \frac{(1 - \cos\lambda)}{2},$$
$$-\pi \le \lambda \le \pi. \tag{11}$$

Note that as expected there is no power at dc.

When the signal-to-noise ratio (SNR) of the channel is sufficiently high, one can use many levels for transmission and consequently increase the data rate with the same symbol rate. Let us assume without loss of generality that the input sequence $\{a_k\}$ takes on values from a set of integers $\{0, 1, \cdots, m-1\}$. If we redefine the summation of (1) as "modulo m" sum, the sequence $\{b_k\}$ is also an m-level sequence. This transformation of $\{a_k\}$ into $\{b_k\}$ is usually referred to as "precoding" and will be further discussed in Section III. The values which $\{c_k\}$ takes on range from $-(m-1)$ to $(m-1)$. For a given c_k, the original data a_k is reconstructed simply by

$$a_k = c_k \text{ modulo } m. \tag{12}$$

The last relationship guarantees that propagation of errors in bit-by-bit detection can be avoided. The first-order distribution of $\{c_k\}$ is given by

$$\Pr\{c_k = i\} = \frac{m - |i|}{m^2},$$
$$-(m-1) \le i \le m-1. \tag{13}$$

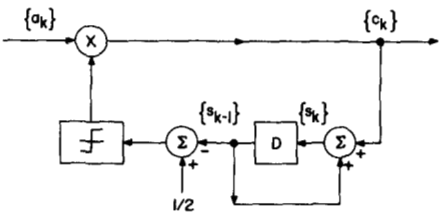


Fig. 2. Generation of bipolar signals (II).

The power spectrum of this multilevel bipolar sequence takes the same form as (11):

$$P_c(\lambda) = \frac{(1 - \cos\lambda)(m^2 - 1)}{6}. \tag{14}$$

The bipolar signal-generating circuit of Fig. 2 is also generalizable to the multilevel case:

$$c_k = a_k \cdot \operatorname{sgn}\left\{\frac{m-1}{2} - s_{k-1}\right\} \tag{15}$$

$$s_k = s_{k-1} + c_k. \tag{16}$$

This encoding procedure is a special case of the feedback balanced codes (FBC) studied by Kaneko and Sawai [5]. The running sum s_k now takes on $2(m-1)$ different values; $\{0, 1, \cdots, 2m-3\}$. From (15) and (16) we can see that the sequence $\{s_k\}$ is a Markov sequence characterized by a regular chain with $2(m-1)$ states. The range of values which $\{c_k\}$ takes on is still $[-(m-1), (m-1)]$, but its distribution is different from (13):

$$\Pr\{c_k = i\} = \begin{cases} \dfrac{1}{m}, & i = 0 \\ \dfrac{1}{2m}, & 1 \le |i| \le m-1. \end{cases} \tag{17}$$

With some manipulation, the autocovariance function of $\{s_k\}$ can be shown, [5] to be given by

$$C_s(k) = E\left\{\left(s_i - \frac{m-1}{2}\right)\left(s_{i+k} - \frac{m-1}{2}\right)\right\}$$
$$= \begin{cases} \frac{1}{12}\{2m(m-1) - 1\}, & k = 0 \\ \frac{1}{12}(m-2)m^{1-|k|}, & |k| \ne 0. \end{cases} \tag{18}$$

Therefore, the spectral density of the sequence $\{s_k\}$ (except for dc component) is given by

$$P_s(\lambda) = \frac{(m^2-1)\{2m^2 - 2m + 1 - 2m\cos\lambda\}}{12(m^2 - 2m\cos\lambda + 1)}. \tag{19}$$

Then the power spectrum of sequence $\{c_k\}$ is obtained from (16) and (19) as

$$P_c(\lambda) = |1 - \exp(-i\lambda)|^2 P_s(\lambda)$$
$$= \frac{(1-\cos\lambda)\{2m^2 - 2m + 1 - 2m\cos\lambda\}(m^2-1)}{6(m^2 - 2m\cos\lambda + 1)}. \tag{20}$$

Clearly when $m = 2$, (20) is equal to (14) the spectrum of multilevel bipolar signal. However, the right-hand expression of (20) approaches twice that of (14) for

large m. This is due to the difference in distribution forms (13) and (17).

Sipress [1] and Franaszek [58] discuss ternary block codes in which the running sum s_k of (16) is defined as the state of the encoder; a code word is chosen so as to maintain the value of s_k close to zero. Some other results which pertain to dc free constraint or spectral shaping are discussed by Wolf [54], Gorog [55], and Croisier [56]. Chien [57] compares various dc free codes in terms of their coding efficiency.

III. Correlative Level Coding and Optimum Decoding Methods

The correlative level coding or partial-response signaling schemes have been developed for applications to digital data modems. As we shall see later, the underlying concepts of these techniques are quite similar to those of the dc free codes discussed in the previous section.

The communication model we assume here is a baseband channel with pulse amplitude modulation (PAM) signal transmission. The results to be obtained later are extendable to other modulation systems. Many authors report applications of the correlative level coding technique to FM [6]–[8], phase-shift keyed (PSK) [8], quadrature amplitude modulation (QAM) [9], [10], vestigial sideband (VSB) or single-sideband (SSB) systems [11], [12].

The system is characterized by a transfer function $H(f)$, which summarizes the overall frequency characteristics of the signal generator, the equivalent baseband channel, and the receiving filter (including an equalizer, if any) as shown schematically in Fig. 3. Let the impulse response function of channel $H(f)$ be given by $h(t)$ (Fig. 4). A conventional digital communication system chooses digit spacing T large enough to avoid intersymbol interference, thus a linear system $H(f)$ combined with the sampler is essentially a "memoryless" digital channel.

If we choose the sampling spacing and phase as shown in Fig. 4(b), the resulting digital channel has a one time-unit memory, and the transfer function from the data source to the sampler output is given by

$$G(D) = 1 + D. \tag{21}$$

Here we assume the values of $h(t')$ at $t' = iT'$ are virtually zero except for $i = 0$ and 1. If these conditions are not met, additional channel shaping is necessary via either an analog filter or a transversal filter. Another way of looking at Fig. 4(b) is that we introduce a full amount of intersymbol interference at $t' = T'$. Lender's duobinary signaling [6]–[8], which we will describe later, is based on this principle.

A binary data sequence $\{a_k\}$ is first precoded into another binary sequence $\{b_k\}$ according to the rule

$$b_k = b_{k-1} \oplus a_k. \tag{22}$$

The precoding allows us to avoid possible propagation of errors, and this transformation is equivalent to the

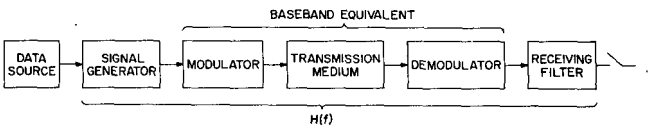


Fig. 3. Digital communication system.

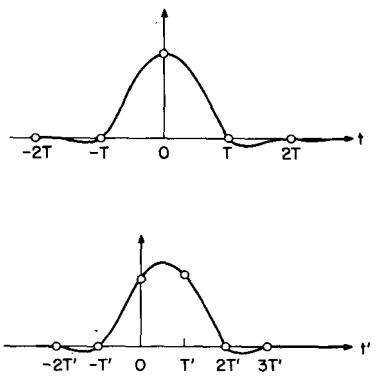


Fig. 4. Impulse response function $h(t)$ and sampling instants. (a) Conventional signal. (b) Duobinary signaling.

modulo 2 integration defined by (1) for bipolar signaling of (2). This is because the precoder is an inverse filter $1/G(D)$ defined over modulo 2 addition [17]. That is, the precoders of the bipolar and duobinary signaling are given by $[1/1 - D]_{\text{mod } 2}$ and $[1/1 + D]_{\text{mod } 2}$, respectively. Clearly these two are equivalent. The accomplishment of this duobinary scheme is to transmit binary data at the Nyquist rate using realizable filters. Furthermore, the system is rather insensitive to the change in data rate [31], [50].

Lender [13] extended the duobinary concept to polybinary signaling

$$G(D) = 1 + D + \cdots + D^N \tag{23}$$

and to polybipolar signaling

$$G(D) = 1 + D + \cdots + D^{N-1} - D^N - \cdots - D^{2N-1}. \tag{24}$$

The cases $N = 1$ in (23) and (24) reduce to the duobinary and bipolar signals, respectively. Since the resulting signal is of multilevel with correlation among successive digits, this class of code transformation is called "correlative level coding." A communication channel with this type of signaling technique is often called a "partial-response channel" [14], [15], since sample points are chosen at points halfway to a full-response [Fig. 4(b)].

A discrete system representation of correlative level coding or partial-response system with a precoder is given in Fig. 5, where $A(D)$ is the polynomial representation of sequence $\{a_k\}$:

$$A(D) = \sum_{k=1}^{\infty} a_k D^k. \tag{25}$$

Among the general class [16] of correlative level coding or partial-response signaling methods the most frequently used is

$$G(D) = 1 - D^2 \tag{26}$$

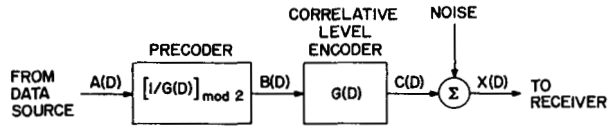


Fig. 5. Discrete system representation of correlative level coding system.

which is obtained by shaping the channel transfer function into $H(f)$ such that

$$T \sum_{m=-\infty}^{\infty} H\left(f - \frac{m}{T}\right) = G[\exp(-i2\pi fT)]$$

$$= 2 \exp\left[-i\left(2\pi fT - \frac{\pi}{2}\right)\right] \sin 2\pi fT \qquad (27)$$

for a chosen data rate $R = 1/T$. The simplest solution to (27) is clearly

$$H(f) = \begin{cases} \frac{1}{T} 2 \exp\left[-i\left(2\pi fT - \frac{\pi}{2}\right)\right] \sin 2\pi fT, & -\frac{1}{2T} \le f \le \frac{1}{2T} \\ 0, & \text{elsewhere} \end{cases} \qquad (28)$$

that is, the phase characteristic of $H(f)$ is linear, and its amplitude characteristic is a half-cycle of sine function, possessing nulls at dc and at the Nyquist frequency $1/2T$. The spectral nulls at both ends are desirable, since they allow insertion of pilot tones which convey the modulating carrier phase and the data clock. It will be clear that the system $G(D) = 1 - D^2$ is mathmatically equivalent to an interleaved form of the bipolar signaling system.

Kobayashi and Tang [17], [18] recently pointed out that a digital magnetic recording system is equivalent to a system with $G(D) = 1 - D$. This is due to the fact that saturating recording at the writing process followed by differentiation at the readback process yields, in effect, a digital transfer function $G(D) = 1 - D$. The so-called nonreturn to zero interleaved (NRZI) recording method is equivalent to the precoding operation of (1). They have proposed a high density recording system, named NRZI, which is essentially equivalent to $G(D) = 1 - D^2$. As far as the processing of readback signal is concerned, the system is linear; hence processing techniques developed for partial-response modems are equally applicable to magnetic recording systems [19], [20].

Because of the many desirable features and the simplicity of implementation, the duobinary signaling and other correlative level coding schemes have been widely used in high-speed data modems. However, the duobinary signal has three levels in the channel output which must be distinguished; thus it seems to require a higher SNR (ranging 2.1 dB ~ 3.0 dB depending on the location of the channel noise source) for equal performance than the ideal binary system. This is the penalty we have paid in exchange for the increase in data rate and the insensitivity to system's perturbation.

Recently, however, Kobayashi [19]–[22] and Forney [23], [24] have shown that this apparent decrease in noise margin is not an inherent drawback of the correlative level coding technique, but is due to nonoptimality of the conventional bit-by-bit detection method. They clarified an analogy between a correlative level coder and a convolutional encoder: both systems are representable by finite state machines. This observation led them to develop a new type of decoding method, namely, the MLD algorithm, which is analogous to the Viterbi algorithm [25], [26] for convolutional codes. Omura [27] has shown that the MLD algorithm is a special case of dynamic programming. He also discusses applications of this algorithm to optimum receivers for a general class of channels with memory [28], [29]. Ungerboeck [59] discusses a sequence decision scheme by applying the maximum likelihood decision rule on bit-by-bit basis.

In the present section we will derive the MLD algorithm in a different way from the earlier publications [19], [22]. We will show that under the Gaussian noise assumption the MLD algorithm can be viewed as a new solution to perform matched filter detection on sampled sequences of infinite length without requiring an unreasonable number of matched filters.

Let us consider the simplest case, i.e., $G(D) = 1 - D$ with a binary input which characterizes a bipolar encoder or a magnetic recording channel. The input $\{a_k\}$, precoded sequence $\{b_k\}$, and the channel input sequence $\{c_k\}$ are related by

$$b_k = a_k \oplus b_{k-1}, \qquad b_0 = 0. \qquad (29)$$

and

$$c_k = b_k - b_{k-1}. \qquad (30)$$

Suppose that the binary information sequence $\{a_k\}$ is of length N, i.e., $1 \le k \le N$. Since the sequence $\{a_k\}$ is mapped into $\{c_k\}$ in one-to-one fashion, there are 2^N different vectors which $\{c_k\}$ can take on.

Let $\{Y_k\}$ be the noisy output from the channel:

$$Y_k = c_k + n_k. \qquad (31)$$

If we assume that the noise sequence $\{n_k\}$ is a Gaussian random variable and is uncorrelated from digit to digit, then the maximum likelihood decision criterion is equivalent to the minimum distance decision rule based on the measure

$$D(\hat{\mathbf{c}}) = \sum_k (Y_k - \hat{c}_k)^2 = ||\mathbf{Y} - \hat{\mathbf{c}}||^2 \qquad (32)$$

where we denote $\hat{\mathbf{c}}$ to represent sequence $\{\hat{c}_k\}$ for brevity. The maximum likelihood solution is that $\hat{\mathbf{a}}$ which minimizes $D(\hat{\mathbf{c}})$, where $\hat{\mathbf{c}}$ and $\hat{\mathbf{a}}$ are related through (29) and (30). Let us rewrite $D(\hat{\mathbf{c}})$ as

$$D(\hat{\mathbf{c}}) = -2 < \mathbf{Y}, \hat{\mathbf{c}} > + ||\hat{\mathbf{c}}||^2 + ||\mathbf{Y}||^2. \qquad (33)$$

Since the last term is independent of $\hat{\mathbf{c}}$, the maximum likelihood solution is that $\hat{\mathbf{a}}$ which maximizes

$$J(\hat{\mathbf{c}}) = \langle \mathbf{Y}, \hat{\mathbf{c}} \rangle - \tfrac{1}{2} ||\hat{\mathbf{c}}||^2. \tag{34}$$

Then a brute-force method would be to compute $J(\mathbf{c})$ of (34), for 2^N different patterns of $\mathbf{c}$ and select the greatest one. Such a decision system could be implemented using 2^N different matched filters in parallel whose impulse response sequences are $\{c_{N-n}\}$, $1 \leq n \leq N$. This receiver structure is, of course, impractical, since the sequence length N is virtually infinite. Now we will show that this dimensionality problem can be overcome by applying the discipline of dynamic programming [30].

Let $\mathbf{c}_k$ denote the first k components of sequence $\mathbf{c}$

$$\mathbf{c}_k = [c_1 \cdots c_k] \tag{35}$$

and let J_k denote $J(\hat{\mathbf{c}}_k)$. Then

$$\begin{aligned} J_k &= \langle \mathbf{Y}_k, \hat{\mathbf{c}}_k \rangle - \tfrac{1}{2} ||\hat{\mathbf{c}}_k||^2 \\ &= \sum_{n=1}^{k} \{(\hat{b}_n - \hat{b}_{n-1}) Y_n - \tfrac{1}{2}(\hat{b}_n - \hat{b}_{n-1})^2\}. \end{aligned} \tag{36}$$

Then the maximum likelihood solution is equivalent to finding the sequence $\{\hat{b}_1, \hat{b}_2, \cdots, \hat{b}_N\}$ which maximizes $J_N = J(\hat{\mathbf{c}})$, where N, the size of data, can be infinite.

Now we write J_k in an iterative form

$$J_k = J_{k-1} + (\hat{b}_k - \hat{b}_{k-1}) Y_k - \tfrac{1}{2}(\hat{b}_k - \hat{b}_{k-1})^2. \tag{37}$$

On defining a function $\mu_k(i)$ as

$$\mu_k(i) = \max_{\{\hat{\mathbf{b}}_{k-1}, \hat{b}_k = i\}} J_k, \qquad i = 0, 1 \tag{38}$$

and substituting (38) into (37), we obtain the recursive relationship for $\{\mu_k(i)\}$:

$$\begin{aligned} \mu_k(i) &= \max_{\{\hat{\mathbf{b}}_{k-1}\}} \{J_{k-1} + (i - \hat{b}_{k-1}) Y_k - \tfrac{1}{2}(i - \hat{b}_{k-1})^2\} \\ &= \max_{j=0,1} \{\mu_{k-1}(j) + (i - j) Y_k - \tfrac{1}{2}(i - j)^2\}, \\ &\qquad i = 0, 1. \end{aligned} \tag{39}$$

Or, equivalently

$$\mu_k(0) = \max \begin{Bmatrix} \mu_{k-1}(0) \\ \mu_{k-1}(1) - Y_k - \tfrac{1}{2} \end{Bmatrix} \tag{40}$$

and

$$\mu_k(1) = \max \begin{Bmatrix} \mu_{k-1}(0) + Y_k - \tfrac{1}{2} \\ \mu_{k-1}(1) \end{Bmatrix}. \tag{41}$$

Note that (40) and (41) are not independent. For example, if $\mu_k(0) = \mu_{k-1}(1) - Y_k - 1/2$, then it follows that $\mu_k(1) = \mu_{k-1}(1)$. Similarly if $\mu_k(1) = \mu_{k-1}(0) + Y_k - 1/2$, then $\mu_k(0) = \mu_{k-1}(0)$.

Given the channel output $\mathbf{Y}_k$, the function $\mu_k(i)$ represents a metric (or the likelihood value) of the most likely sequence among all possible candidates with the constraint $\hat{b}_k = i$. We know the initial condition $b_0 = 0$. Then starting from

$$\mu_0(i) = \begin{cases} 0, & i = 0 \\ -\infty, & i = 1 \end{cases} \tag{42}$$

the repetitive use of (40) and (41), for $k = 1, 2, \cdots, N$, uniquely determines the maximum likelihood solution. An implementation example of the MLD is discussed in [19]. In that paper, several other important problems are addressed: the effect of precoding on the decoding error rate and error pattern, the number of quantization levels required, and the problem of decoder buffer overflows.

The results obtained previously hold with appropriate modification for a class of systems $G(D) = 1 + D^K$ and for m-level signaling $m \geq 2$. The performance of the MLD is analyzed elsewhere [19], [22], [24] in great detail, therefore we simply quote the results. An asymptotic (i.e., for a high SNR) expression for the symbol error rate in the MLD method is given by [22], [24]:

$$P_{\text{MLD}} = 4(m - 1) Q\left(\left(\frac{3R}{m^2 - 1}\right)^{1/2}\right) \tag{43}$$

where

$$Q(x) = \int_x^\infty \frac{1}{2\pi} \exp\left\{-\frac{t^2}{2}\right\} dt \tag{44}$$

and where R is the channel SNR. The symbol error rate of the conventional bit-by-bit detection method is given by [31]

$$P_{\text{bit}} = 2\left(1 - \frac{1}{m^2}\right) Q\left(\left(\frac{3R}{2(m^2 - 1)}\right)^{1/2}\right). \tag{45}$$

The significance of the difference in the arguments of Q function by a factor of $\sqrt{2}$ will be appreciated if one recalls the following approximation formula [32]:

$$Q(x) \cong \frac{1}{x(2\pi)^{1/2}} \exp\left\{-\frac{x^2}{2}\right\}, \qquad x > 3.5. \tag{46}$$

In terms of SNR, the performance gain of the MLD method over the bit-by-bit detection method corresponds to 2.5 ~ 3.0 dB. Therefore we can recover, by use of this new decoding method, the loss in noise margin previously accepted as a penalty in exchange for the desired spectral shaping.

The MLD discussed previously provides the best performance among the reception schemes known so far. We can view this decoding scheme as the one which makes full use of the redundancy inherent in correlative level coded sequences. The conventional bit-by-bit detection method has partially utilized the sequence redundancy, that is, its error detection capability has been used for monitoring purpose (See Lender [8] and Gunn and Lombardi [33]). Kobayashi and Tang [21], [34] have generalized the error detection method to an algebraic form so that a simple circuit can perform error detection for any $G(D)$ and for any m. They have extended the algebraic approach to a more general de-

cision scheme, named the ambiguity zone decoding (AZD) method, in which the quantizer makes a soft decision including ambiguity (or erasure) levels. Most of the digits in the ambiguity levels are replaceable by correct values by using the inherent redundancy of the sequence. The AZD method is an extension of the null-zone detection method studied by Smith [35]. An implementation and the performance of this algebraic decoder are discussed in [34] in great detail. An asymptotic expression for the decoding error rate is given by

$$P_{AZD} = 3\left(1 - \frac{1}{m}\right)Q\left([2(2^{1/2} - 1)]\left(\frac{3R}{m^2 - 1}\right)^{1/2}\right). \quad (47)$$

Although somewhat inferior to the MLD in its performance, the AZD method possesses an advantage over the MLD in its simple implementation. The number of quantization levels is in general much smaller than that required in the MLD [19], [22]. Furthermore, in the MLD method the number of "states," which determines the decoder complexity, is mN for a system $G(D) = 1 + D^N$. Thus the MLD algorithm tends to require a significant amount of computation effort and memory requirement when the number of signal levels is large. The AZD method will be more attractive in that respect.

The discussion we have made thus far in the present section is, in principle, applicable to a general form

$$G(D) = g_0 + g_1 D + \cdots + g_n D^n \quad (48)$$

if the $\{g_i\}$ are a set of integers with their greatest common divisor equal to 1, and g_0 and m are relatively prime [36], [22]. Miyakawa and Harashima [37], [38] recently proposed a scheme which can remove this constraint.

In the correlative level coding system of Fig. 5 we transformed an m-level sequence $A(D)$ into another m-level sequence $B(D)$. However, there is no essential reason why the precoded sequence $B(D)$ must be also an m-level sequence. In the scheme which Miyakawa and Harashima propose (Fig. 6), the information sequence $A(D)$ is transformed into sequence $C(D)$ via transformation T, which is then passed into $G^{-1}(D)$, the inverse of the channel $G(D)$. The output $B(D)$ is then transmitted over the channel $G(D)$ where the $\{g_i\}$ are not necessarily integers. The transformation T must satisfy the following constraints:

1) T is invertible;
2) sequence $B(D)$ is peak limited, i.e., $b_{min} \leq b_k \leq b_{max}$, for some b_{min} and b_{max}.

If, in particular, $G(D)$ represents a conventional correlative level coding system, i.e., g_i are integers and $(g_0, m) = 1$, then by setting

$$T = G(D)[G^{-1}(D)]_{\text{mod } m} \quad (49)$$

$B(D)$ corresponds to the precoded sequence. It should be remarked here that the representation of a precoder in terms of a finite state machine T followed by another finite state machine $G^{-1}(D)$ has been derived by Kanaya [39].

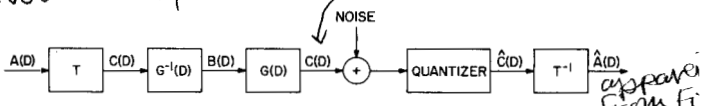


Fig. 6. Generalized correlative level coding system.

Clearly the class of T we allow here is larger than the class represented by (49). We determine T from the following considerations in addition to the two conditions given previously:

3) we should be able to obtain $\hat{A}(D) = T^{-1} \cdot \hat{C}(D)$ without significant error propagation problem;
4) spectrum of sequence $C(D)$ should be desirable from the viewpoint of SNR and possible insertion of pilot tones.

Let us assume here that $G(D)$ is normalized so that $g_0 = 1$. Then the implementation of transformation T followed by $G^{-1}(D)$ is given by Fig. 7, and sequence $I(D)$ in this figure represents the intersymbol interference sequence due to preceding digits

$$I(D) = [G(D) - 1]B(D) \quad (50)$$

or

$$i_k = \sum_{i=1}^{n} g_i b_{k-i}. \quad (51)$$

The intermediate sequence $\{c_k\}$ should be determined so that

$$b_{min} \leq c_k - i_k \leq b_{max}. \quad (52)$$

If the information sequence $\{a_k\}$ is binary and $b_{min} = -b_{max}$ then a possible transformation T can be given graphically by Fig. 8. Note c_k is determined by the present input a_k and precursor interference term i_k which is a function of $a_{-1}, a_{k-2}, \cdots$. Thus c_k is uniquely determined by the past input of $\{a_k\}$.

It will be clear that the MLD and AZD methods developed for correlative level coding systems are, in principle, applicable to this generalized class, since the system is representable as a finite state machine.[1] The practical implementation and performance evaluation are left for further investigation.

IV. Algebraic Block Coding With Run-Length Constraints

In sections II and III we discussed methods of transforming a given digital sequence into a sequence with some desirable properties. But the resulting sequences contained more signal levels than the original one. In other words, redudancy was introduced amplitudewise.

[1] Of course we assume here that the channel is of finite memory, i.e., $g_i = 0, i > n$ [see (48) and (51)].

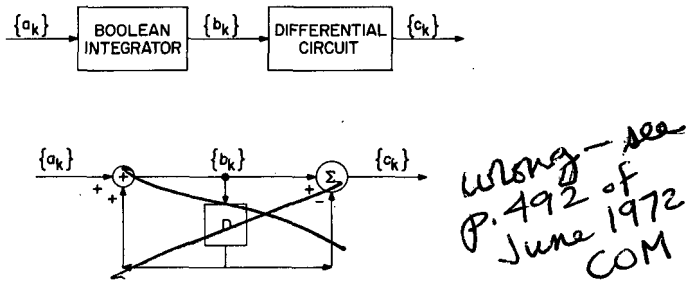


Fig. 7. Implementation of transform T and $G^{-1}(D)$, $(g_0 = 1)$.

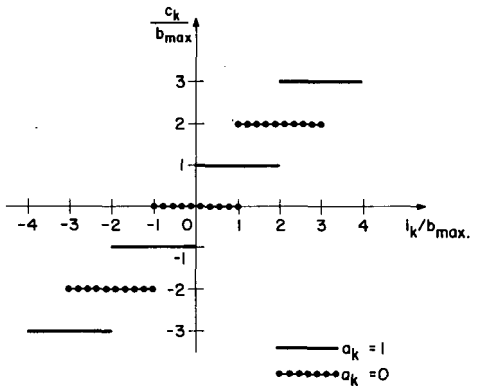


Fig. 8. Transform T for binary input $\{a_k\}$.

There are some applications, however, in which the increase in signal levels is not desirable or allowable. Binary signals are easier to generate and modulate than ternary or multilevel signals. As pointed out earlier, there is a strong analogy between a digital communication system and a digital magnetic recording system. The present recording technology, however, limits the input signal levels to two, since the two saturation levels are the only stable levels. Under these circumstances, therefore, it is necessary to introduce redundancy in a form different from pseudo ternary or correlative level codes.

Kautz [40] introduced a family of codes to represent binary data subject to constraints on the maximum separation between successive changes in signal levels. This constraint is motivated by applications to recording systems in which the clock of reading (receiving) side must be derived from the data itself. As was mentioned in Section III, level transitions yield plus or minus pulses at the readback process; this information can be used as synchronizing information. Naturally the same argument holds for communication channels with bipolar signaling.

Given a sequence $\mathbf{X} = [X_1, X_2, \cdots]$ of 0's and 1's, its "modulo 2 derivative" is defined as a sequence $\mathbf{Y} = [Y_1, Y_2, \cdots]$ with

$$Y_i = X_i \oplus X_{i-1}, \qquad i \geq 1 \tag{53}$$

where X_0 is defined as a reference binary state. By this transformation, each string of 0's and 1's in $\mathbf{X}$ is converted into a string of 0's (but shorter by one) in $\mathbf{Y}$. For example, if $\mathbf{X}$ = [001111000001111000] with $X_0 = 0$, then $\mathbf{Y}$ = [001000100001000 100]. Therefore the problem originally addressed is reduced to finding the class of $\mathbf{Y}$ in which it is required that at most k 0's occur between successive 1's, where k is the constraint parameter $k = 1, 2, 3, \cdots$. We call a sequence (code) subject to this constraint a "k-limited" sequence (code) or "k-constraint" sequence (code), following Tang's terminology [41]–[44].[2]

The system which generates a k-constraint sequence can be conveniently represented by a finite state machine model as in Fig. 9, where the nodes correspond to the states and 0 or 1 is generated at each state transition. A system that generates a sequence under such a constraint is a special case of "discrete noiseless channels" studied by Shannon in his celebrated paper [45], and by Freiman and Wyner [46].

Let $N_k(n)$ denote the number of distinct allowable $\mathbf{Y}$ sequences of length n. Then we have

$$N_k(n) = \begin{cases} 2^n, & n \leq k \\ \sum_{i=1}^{k+1} N_k(n-i), & n \geq k+1. \end{cases} \tag{54}$$

The preceding equation is obtained based on the following observation. For $n \leq k$, any binary sequence does not violate the given constraint. For $n \leq k+1$, if a code sequence starts with $(i-1)$ zeros followed by one, it may be followed by any of N_k $(n-i)$ k-limited sequences of length $(n-i)$. When $k = 1$, the recursive of (54) generates $N_1(0) = 1$, $N_1(1) = 2$, $N_1(3) = 5$, etc., and this is well-known Fibonacci numbers [47], [48].[3] For $k > 1$, Kautz calls the sequence $\{N_k(n)\}$ generalized Fibonacci numbers. As we shall see later, these numbers appear as weighting coefficients in encoders and decoders.

Equation (54) is rewritten as

$$N_k(n) = \begin{cases} 2^n, & 0 \leq n \leq k \\ 2N_k(n-1) - N_k(n-k-2), & n \geq k+1. \end{cases} \tag{55}$$

In order to obtain asymptotic expressions for $N_k(n)$, let us consider the generating function[4] $G_k(z)$ of $\{N_k(n)\}$:

$$G_k(z) = \sum_{n=0}^{\infty} N_k(n) z^n. \tag{56}$$

From (54) and (56) we obtain after some manipulation

[2] There is no particular meaning in choosing the symbol k as a constraint parameter.

[3] The so-called Fibonacci sequence $\{F_n\}$ is defined by $F_0 = 0$, $F_1 = 1$, $F_n = F_{n-1} + F_{n-2}$, $n \geq 2$. Therefore, the following relation holds: $N_1(n) = F_{n+2}$, $n \geq 2$.

[4] If we use z^{-1} instead of z, then $G_k(z)$ represents the z transform of sequence $\{N_k(n)\}$.

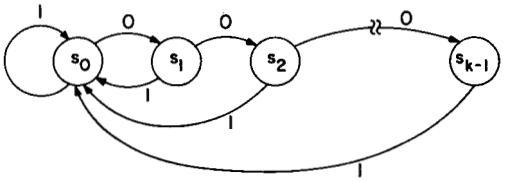


Fig. 9. Finite state machine model of k-constraint sequence.

$$G_k(z) = \frac{1 + z + \cdots + z^k}{1 - \sum_{i=1}^{k+1} z^i}. \tag{57}$$

Consider, for example, the case $k = 1$

$$G_1(z) = \frac{1 + z}{1 - z - z^2} = \frac{1}{\sqrt{5}} \left\{ \frac{\phi^2}{1 - \phi z} - \frac{\hat{\phi}^2}{1 - \hat{\phi} z} \right\} \tag{58}$$

where

$$\phi = \frac{1 + \sqrt{5}}{2}, \qquad \hat{\phi} = 1 - \phi = \frac{1 - \sqrt{5}}{2}. \tag{59}$$

Then we get a closed form expression for $N_1(n)$ as follows:

$$N_1(n) = \frac{1}{\sqrt{5}} (\phi^{n+2} - \hat{\phi}^{n+2}). \tag{60}$$

It can be shown [48] that $N_1(n)$ is bounded as follows:

$$\phi^n \leq N_1(n) \leq \phi^{n+1} \tag{61}$$

which leads to

$$\log_2 \phi \leq \frac{1}{n} \log_2 N_1(n) \leq \frac{n+1}{n} \log_2 \phi. \tag{62}$$

Therefore the information per symbol in an optimal code with the constraint $k = 1$ is given for large n by

$$C = \lim_{n \to \infty} \frac{\log_2 N_1(n)}{n} = \log_2 \phi = 0.6942. \tag{63}$$

This is called the "capacity" of the given "noiseless discrete channel," if we follow Shannon's terminology [45]. For a general $k \geq 1$, the capacity C is given by [45]

$$C = \log_2 z_0 \tag{64}$$

where z_0 is the largest real root of the characteristic

$$1 - \sum_{i=1}^{k+1} z^i = 0 \tag{65}$$

or equivalently

$$1 - 2z + z^{k+2} = 0. \tag{66}$$

Franaszek [49] has found another method to compute the capacity C based on the transition matrix. Define the transition matrix $\mathbf{\bar{D}} = [d_{ij}]$ by

$$d_{ij} = \begin{cases} 1, & \text{if a transition from state } i \text{ to state } j \text{ is allowed} \\ 0, & \text{otherwise} \end{cases} \tag{67}$$

For a finite state machine of Fig. 9, $\mathbf{\bar{D}}$ is given by

$$\mathbf{\bar{D}} = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & & \\ 1 & 0 & 0 & 1 & & \\ \vdots & & & & \ddots & \\ \vdots & & & & & 1 \\ 1 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{68}$$

Then C is bounded by

$$\frac{\log_2 \{ \sum_{ij} [\mathbf{\bar{D}}^n]_{ij} \} - 2 \log_2 k}{n + k - 1} \leq C \leq \frac{\log_2 \{ \sum_{ij} [\mathbf{\bar{D}}^n]_{ij} \}}{n} \tag{69}$$

for any $n \geq 1$. The k is the number of states.

Example 1

Consider the case where $n = 4$ and $k = 1$. Then there are $N_1(4) = 8$ such codewords, $\mathbf{Y}$, as listed in Table I. In Table I $\tilde{\mathbf{Y}}$ lists the complements of code words of $\mathbf{Y}$ and therefore satisfies the following condition:

every $n(=4)$ digit code word in $\mathbf{Y}$ contains no 1 strings longer than $k(=1)$. (70)

The set of all 4-digit binary sequences (with no k constraint) contains 2^4 members, and they can be mapped to integers 0 to $2^4 - 1$, using 2^i $(0 \leq i \leq 3)$ as weights. It will not be difficult to see that eight sequences of Table I, $\mathbf{Y}$, can be mapped into integers 7 to 0 by assigning weights 5, 3, 2, 1 from the leftmost to rightmost digits. Consider, for example, the second member from the top of Table I: $\mathbf{Y} = [0110]$, $\tilde{\mathbf{Y}} = [1001]$. Then

$$\begin{aligned} I(\tilde{\mathbf{Y}}) &= N_1(3) \times 1 + N_1(2) \times 0 \\ &\quad + N_1(0) \times 0 + N_1(0) \times 1 \\ &= 5 \times 1 + 1 \times 1 = 6. \end{aligned} \tag{71}$$

Kautz [40] has generalized this observation and has shown that for general n and k the complement $\tilde{\mathbf{Y}}$ of a code word $\mathbf{Y} = [Y_1 Y_2 \cdots Y_n]$ in k constraint code of n digits can be enumerated by the following formula

$$I(\tilde{\mathbf{Y}}) = \sum_{i=1}^{n} N_k(n - i) \tilde{Y}_i \tag{72}$$

where $N_k(n)$ is a generalized Fibonacci sequence defined by (74). The linear relationship of (72) makes the encoding and decoding operations very simple. Encoding proceeds as follows. Given a binary source sequence, we first chop the sequence into blocks of length ℓ, where $\ell = \lceil \log_2 N_k(n) \rceil$.[5] A binary vector of length ℓ is mapped into an integer I, where $0 \leq I \leq N_k(n) - 1 \leq 2^\ell - 1$. Integer I is then transformed into vector $\tilde{\mathbf{Y}}$ of size n according to the following rule:

$$R_1 = I \tag{73}$$

[5] For a given real number x, we define the integer $\ell = \lceil x \rceil$ by $x \leq \ell < x + 1$.

TABLE I

Y	$\tilde{\mathbf{Y}}$
0101	1010
0110	1001
0111	1000
1010	0101
1011	0100
1101	0010
1110	0001
1111	0000

$$\tilde{Y}_i = \begin{cases} 1, & \text{if } R_i \geq N_k(n-i) \\ 0, & \text{otherwise} \end{cases} \tag{74}$$

$$R_{i+1} = R_i - \tilde{Y}_i \times N_k(n-i), \quad i = 1, 2, \cdots, n. \tag{75}$$

The decoding operation obviously performs the inverse of the encoding procedure.

Tang [41], [42], [44] has extended the results of Kautz to a class of codes in which the minimum as well as the maximum string length is limited. This additional constraint keeps adjacent transitions apart to avoid excessive intersymbol interference. Tang and Bahl [44] further generalize the result to multilevel cases.

Tang defined a dk-limited sequence as a sequence satisfying simultaneously the following conditions:

1) between any pair of adjacent 1's the run-length of 0's is at least d;
2) any run-length of 0's is at most k.

When we "integrate, modulo 2" a dk-limited sequence $\mathbf{Y}$,

$$\begin{aligned} X_i &= X_{i-1} \oplus Y_i \\ &= X_0 + Y_1 \oplus Y_2 \oplus \cdots \oplus Y_i \end{aligned} \tag{76}$$

then the length of any run of 0's and 1's in the resulting sequence $\mathbf{X}$ is at most $k+1$; and except for boundary runs, the run-length is also at least $d+1$.

Example 2

Let $d = 1$, $k = 3$, and $n = 9$. A sequence which satisfies these constraints is given, for example, by $\mathbf{Y} = [100010101]$. By integration we get the $\mathbf{X}$ sequence $\mathbf{X} = [111100110]$ or $[000011001]$ depending on $X_0 = 0$ or $X_0 = 1$. The run-length in $\mathbf{X}$ is at most $k+1 = 4$ and is at least $d+1 = 2$ except at block boundaries.

Let us first consider the case $k = \infty$, i.e., no constraint on the maximum run of 0's. Such a sequence is referred to as d-limited or d-constraint sequence [41], [44]. Let $N_d(n)$ denote the number of distinct d-limited sequences of length n. Then the following recursive equation holds

$$N_d(n) = N_d(n-1) + N_d(n-d-1), \qquad n \geq d+1 \tag{77}$$

with the initial condition

$$N_d(n) = \begin{cases} n+1, & 0 \leq n \leq d \\ 0, & n < 0. \end{cases} \tag{78}$$

The preceding equation is obstained based on the following observation. If a code sequence starts with 0 it may be followed by any of $N_d(n-1)$ d-limited sequences of length $(n-1)$. On the other hand if a code starts with a 1, at least d 0's must follow but then it may be followed by any of $N_d(n-d-1)$ d-limited sequences of length $(n-d-1)$. When $n \leq d$, the sequence must either be all 0's or contain 1 at only one digit. (Recall that d constraint is not required for boundary runs.)

The generating function $G_d(z)$ of $\{N_d(n)\}$ is defined by

$$G_d(z) = \sum_{n=1}^{\infty} N_d(n) z^n. \tag{79}$$

From (77)-(79) we obtain

$$\begin{aligned} G_d(z) &= \frac{1 - z^{d+1}}{(1-z)(1-z-z^{d+1})} \\ &= \frac{(1 + z + \cdots + z^d)}{(1 - z - z^{d+1})}. \end{aligned} \tag{80}$$

Thus the characteristic equation of the d-limited sequence is

$$1 - z - z^{d+1} = 0. \tag{81}$$

Note that $G_d(z)$, for $d = 1$, is equivalent to $G_k(z)$, for $k = 1$, given by (58). In fact, a k-limited code, for $k = 1$, and a d-limited code, for $d = 1$, are complementary to each other, for all n. For example, the eight code words of Table I, $\tilde{\mathbf{Y}}$, form the d-limited code, for $d = 1$, $n = 8$.

Tang and Bahl [44] obtained the recursive equation for $N_{dk}(n)$, the number of dk-sequence of length n:

$$N_{dk}(n) = \begin{cases} n+1, & 1 \leq n \leq d \\ N_{dk}(n-1) + N_{dk}(n-d-1), & d+1 \leq n \leq k \\ (d+k+1-n) + \sum_{i=d}^{k} N_{dk}(n-i-1), & k+1 \leq n \leq d+k \\ \sum_{i=d}^{k} N_{dk}(n-i-1), & d+k+1 \leq n. \end{cases} \tag{82}$$

The characteristic equation of dk-sequence is obtained as

$$z^{k+2} - z^{k+1} - z^{k-d+1} + 1 = 0. \tag{83}$$

It will be clear that (83) reduces to (66) and (81), for $d = 0$, and $k = \infty$, respectively.

The d, k, and dk sequences of finite block length n cannot in general be concatenated without violating the given constraints at the boundaries. Tang and Bahl [44] describe a method of inserting buffering sequences of smallest possible fixed length β between adjacent code sequences so that the given constraints are not violated. Degradation in overall coding efficiency can be made arbitrarily small by choosing n large enough, since β is a function of d and k only.

In the k-limited case ($d = 0$), the insertion of a single 1 between any two k sequences is sufficient to preserve k constraint. Therefore, $\beta = 1$. In the d-limited case ($k = \infty$), the insertion of d zeros is sufficient. Therefore, $\beta = d$. In the general dk-limited case the shortest buffer is dependent on the end of the previous sequence and the beginning of the next sequence [44].

Before closing the present section we discuss some other applications of the k sequences. If the frequency characteristic of a channel contains imperfections at both the low- and high-frequency ends, alternations of 1's and 0's, as well as runs of 1's and 0's tend to build up intersymbol interference [50]. Sequences which avoid these two types of undesirable patterns can be generated based on k sequences. Let a k sequence **Y** be integrated to give **X** sequence, whose transitions are separated by $(k + 1)$ digits. If we further integrate **X** sequence, then the resultant sequence **Z** contains runs of 0's or 1's or alternations of 1's and 0's not more than $(k + 2)$ digit long.

Mine *et al.* [51] discuss a scheme to encode a binary sequence into a three level sequence for asynchronous transmission where they require 1) the signal level +1, 0, or −1 cannot hold its level, i.e., a level transition must occur with every digit; 2) in order to allow dc free transmission, +1 and −1 must alternate. Such requirements are closely related to Kautz's Fibonacci code (or Tang's k sequence) and to the bipolar signaling. First a binary sequence is encoded into a k sequence using $k = 1$. A typical output sequence **Y** is shown in Fig. 10(a) where two successive 1's are separated by at most $k(= 1)$ zeros. The sequence **Y** is then integrated to yield **X**, where runs of 0 and 1 are at most $k + 1$ $(= 2)$. The sequence **X** is then passed into a differential circuit $1 - D$ yielding a desired three-level sequence as shown in Fig. 10(c). It is interesting to note that the modulo 2 integrator is equivalent to precoding for $1 - D$. Thus by taking modulo 2 of the three-level sequence Z, we recover the Fibonacci code output **Y** with no propagation of errors.

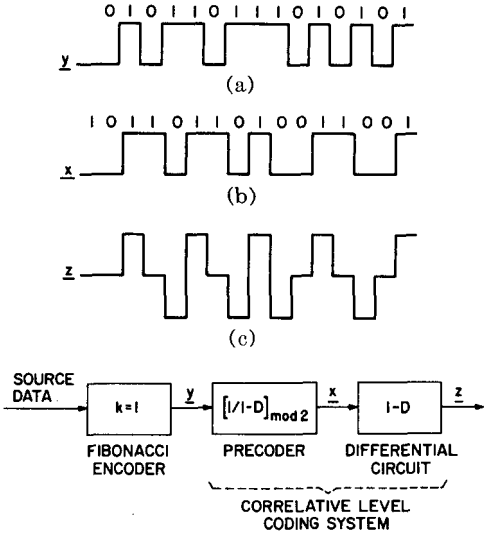


Fig. 10. Combination of Fibonacci coding and correlative level coding.

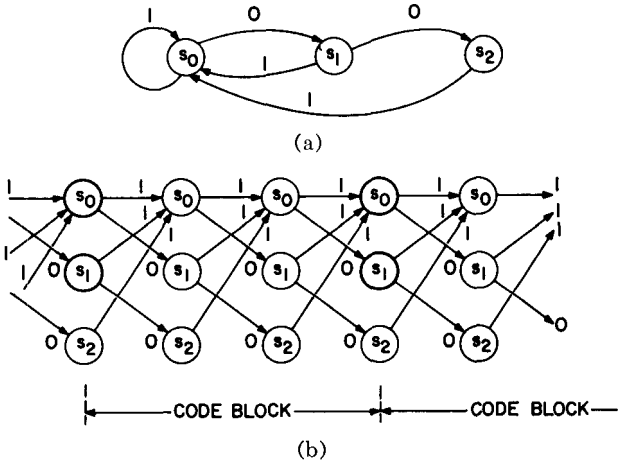


Fig. 11. k-constraint sequence with $k = 2$. (a) State transition. (b) Trellis.

V. State-Oriented Codes

In the generalized Fibonacci codes and the run-length limited codes, we added some fixed symbols to avoid violations of the constraints at the code word boundaries. However, the loss of the efficiency due to these redundant digits is not negligible when the code length is relatively short.

Freiman and Wyner [46] developed a method for determining maximum size block codes with the property that no concatenation of code words violates given code restrictions. Although their results are applicable to any type of constraint characterized by finite state machines, we limit ourselves to the run-length constrained codes.

Let us consider k-constraint codes with $k = 2$. The state transition diagram is given by Fig. 11(a). By introducing time parameter explicitly we obtain what we call the trellis picture of the finite state machine as is shown in Fig. 11(b). We consider the problem of finding an optimal fixed code of length 3. For any chosen code, a sequence of coded messages corresponds to a path on the trellis picture starting at time = 0. Code blocks terminate at times = 3,6,9, $\cdots$, etc. If we restrict ourselves to fixed codes, the selection of a code is equivalent to the selection of a fixed terminal set T. Then the corresponding code consists of those sequences of length 3 which induce transition from every state in T to some state in T. Consider a terminal set $T_0 = \{s_0\}$. Then it will be clear that the code has four members

$$C_0 = \{001, 011, 101, 111\}.$$

Next, we consider a terminal set $T_1 = \{s_0, s_1\}$. The corresponding code C_1 now has five members

$$C_1 = \{010, 011, 101, 110, 111\}.$$

By adding s_1 to the original terminal set T_0, we are forced to eliminate 001 from the code, since 001 is not allowable from the terminal state s_1. However, we now have two additional members 010, 110; thus the net gain is 1 codeword. Consider the largest possible ter-

2/2/79: Mario has seen this for first time (cited by Kotzel Franaszek paper) and says it's important!

minal set $T_2 = \{s_0, s_1, s_2\}$. Since state s_2 does not allow symbol 0 as the emanating symbol, sequences 010, 011 cannot be codewords, and we add one codeword 100. Thus C_2 has four members:

$$C_2 = \{100, 101, 110, 111\}.$$

Therefore, among the three terminal sets considered previously, $T_1 = \{s_0, s_1\}$ gives the code of the largest size. In fact, it can be shown that T_1 is optimal among all possible terminal sets.

Freiman and Wyner [46] have shown that the optimum k-limited block code of length n ($\geq k$) is the one which corresponds to optimal terminal sets $T_{k/2}$ when k is even, $T_{k-1/2}$ or $T_{k+1/2}$ when k is odd, where

$$T_i = \{s_0, s_1, \cdots, s_i\}. \tag{84}$$

The generating function of optimum code size sequence is given by [46]

$$F_k(z) = \frac{1 - \sum_{i=0}^{k=0} \min\{i, k-i\} z^{i+1}}{1 - \sum_{i=1}^{k+1} z^i}. \tag{85}$$

Taking the difference between (57) and (85)

$$G_k(z) - F_k(z) = \frac{\sum_{i=0}^{k-1} [1 + \min\{i, k-i\}] z^{i+1}}{1 - \sum_{i=1}^{k+1} z^i}. \tag{86}$$

The coefficient of the z^n term of (86) represents the number of sequences which must be discarded from Kautz's and Tang's k-limited codes of length n so that code words can be freely concatenated.

Similar results have been obtained for the d-limited case. Fig. 12 show the state transition diagram and trellis diagram for $d = 2$. Consider the code length $n = 4$. Since every sequence allowable from s_i is also allowable from s_j, $0 \leq i \leq j \leq 2$, we need consider only the following three terminal sets: $T_0 = \{s_2\}$, $T_1 = \{s_2, s_1\}$, $T_2 = \{s_2, s_1, s_0\}$. For terminal set T_0, the code consists of three members, i.e., $C_0 = \{0000, 0100, 1000\}$. For terminal set T_1 we have $C_1 = \{0000, 0010, 0100\}$, and for T_2 the corresponding code is given by $C_2 = \{0000, 0001, 0010\}$. Thus an optimal code has size three but the code is not unique. In fact for any d-limited code of any code length $n \geq d$, the following terminal sets are all optimal [46]:

$$T_j = \{s_d, s_{d-1}, \cdots, s_{d-j}\}, \qquad 0 \leq j \leq d. \tag{87}$$

The generating function of optimal code size series is given by

$$F_d(z) = \frac{1}{1 - z - z^{d+1}}. \tag{88}$$

From (86) and (88) we see that the coefficient of z^n in the polynomial

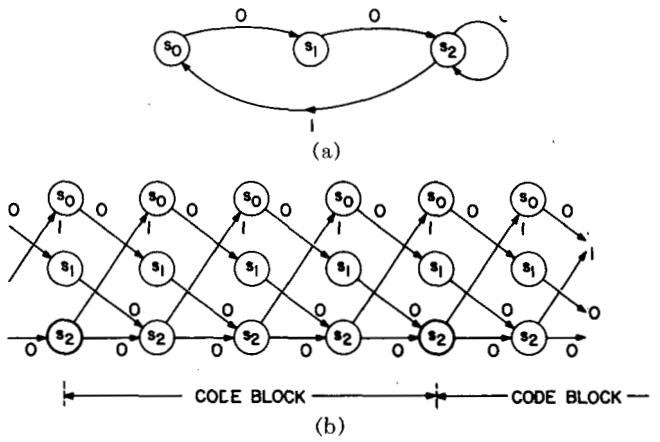


Fig. 12. d-constraint sequence with $d = 2$. (a) State transition. (b) Trellis.

$$G_d(z) - F_d(z) = \frac{z + z^2 + \cdots + z^d}{1 - z - z^{d+1}} \tag{89}$$

gives the number of sequences which must be eliminated from the d-limited code of length n to avoid end effects.

In the class of codes studied previously, code words can be freely concatenated without violating the sequence constraints, since any code word is allowable from every state in the terminal set. Thus we can decode without knowing the present state of the encoder. If we relax the requirement of state independent decoding, we can increase the code efficiency. Gabor [52] introduced a state dependent code for the constraint $k = 1$ and for code length $n = 4$, achieving the rate 2/3 which corresponds to 96 percent of the capacity. Subsequently Tang [43] obtained similar encoding rules for several dk-limited codes of various block lengths. We denote the length of information block and the length of codeword by l and n, respectively. Then a practical encoding scheme has been found [43], for $(d,k;\ l,n) = (1,\infty;\ 2,3)$, $(2,\infty;\ 2,4)$, $(1,5;\ 3,5)$, $(1,10;\ 4,6)$, $(2,5;\ 2,5)$, $(2,10;\ 3,6)$, and $(3,11;\ 2,5)$.

The following example $(d,k;\ l,n) = (2,\infty;\ 2,4)$ is based on Tang's result. We saw that Freiman–Wyner's code for $d = 2$, $n = 4$ contained only $N = 3$ codewords. Therefore, with a fixed code $l = 2$ is not achievable. Consider the largest terminal set $T_2 = \{s_2, s_1, s_0\}$ of Fig. 12. We see from Fig. 12(b) that there are six sequences of length 4 which are allowable from s_2. But there are only three of those from states s_1 and s_0. Therefore if the encoder is at state s_2, the number of sequences available for encoding $l = 2$ bits is more than sufficient; at state s_1 or s_0 the number of sequences is insufficient.

Let the present state of encoder be σ^t. The σ^t is uniquely determined by the last $d(=2)$ symbols of the previous codeword[6]:

$$\sigma^t = g(X_3^{t-1}, X_4^{t-1}). \tag{90}$$

[6] Since the finite state machine of Fig. 12 has memory $d = 2$, σ^t is uniquely determined by X_3^{t-1} and X_4^{t-1}. Specifically: $g(0,0) = s_2$, $g(1,0) = s_1$, and $g(0,1) = s_0$.

Comparable to 3PM!

Let $m(=2)$ bit information data to be encoded be denoted by $\mathbf{a}^t = [a_1{}^t, a_2{}^t]$. Then we consider the following encoding rule

$$\begin{aligned}\mathbf{X}^t &= [X_1{}^t, X_2{}^t, X_3{}^t, X_4{}^t] \\ &= f(\sigma^t; a_1{}^t, a_2{}^t, a_1{}^{t+1}) \\ &= f(g(X_3{}^{t-1}, X_4{}^{t-1}), a_1{}^t, a_2{}^t, a_1{}^{t+1}).\end{aligned} \tag{91}$$

The new state of the encoder is then given by

$$\sigma^{t+1} = g(X_3{}^t, X_4{}^t). \tag{92}$$

One such encoding rule $f(\cdot)$ is given in Table II. Here a blank in the column $a_1{}^{t+1}$ means that it can be either 1 or 0. As is clear from this table and (91), this encoding rule is a "look-ahead" state dependent encoding scheme. The logical equation for encoding and decoding can be derived from the set of truth table of Table II.

With a close observation of Table II, the reader may be able to find the rule which greatly simplifies the encoder. That is, Table II can be replaced by the encoding rule of Table III. This new encoding rule is equivalent to assigning terminal state s_2 at every other time unit in the trellis picture as shown in Fig. 13. Allowable code words start from s_2 and end at s_2 after two or four time units. Franaszek [49] developed a method of constructing this type of code which he calls synchronous variable length codes. Both input and output block lengths are variable but the information bit per symbol ratio is constant over each codeword. This synchronous feature eliminates the need for buffers in encoding the decoding. Franaszek reduces the problem of finding an optimal variable code to that of selecting a set of *principal states*. This notion is analogous to finding an optimal terminal state set in the method of constructing a fixed code. A principal state set T_p has the property that from each state in T_p, there exists a sufficient number of distinct paths to other principal states to maintain the information rate required. A dynamic programming algorithm has been applied to a systematic search of principal states. Optimal variable length codes for various (d,k) constraints have been found [53].

TABLE II

σ^t	a_1^t	a_2^t	a_1^{t+1}	x_1^t	x_2^t	x_3^t	x_4^t	σ^{t+1}
s_2	0	0		0	0	0	0	s_2
	1	0		0	1	0	0	s_2
	1	1		1	0	0	0	s_2
	0	1	0	0	0	0	1	s_0
	0	1	1	0	0	1	0	s_1
s_1	1	0		0	0	0	0	s_2
	1	1	0	0	0	0	1	s_0
	1	1	1	0	0	1	0	s_1
s_0	0	0		0	0	0	0	s_2
	0	1	0	0	0	0	1	s_0
	0	1	1	0	0	1	0	s_1

TABLE III

Input	Output
0	0 0
1 0	0 1 0 0
1 1	1 0 0 0

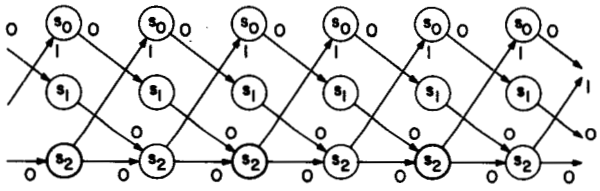


Fig. 13. Assignment of terminal states s_2 at every other time unit.

VI. Concluding Remarks

The review has covered in varying degree of detail several coding techniques for the transmission or recording of digital data. The author hopes that he has indicated some of the major areas of recent development and some underlying mathematical concepts shared by variety of efforts currently engaged in digital communication and recording techniques.

The coding schemes treated in the present paper are different from error detection or correction codes which are intended to combat against the random or burst noise. Rather, these techniques are the means to compensate for undesirable deterministic characteristics of a given channel. Whatever the motivation for code conversion may be, the encoded sequence includes some redundancy which should be exploited as much as possible to improve the reliability against random noise. We have found practical solutions (i.e., the MLD and AZD methods) to a class of correlative level codes. An extension of similar approaches to other types of codes (e.g., run-length limited codes) is left for the future investigation.

Acknowledgment

The author would like to express special thanks to A. Lender, Lenkurt Electric Company Inc., and Prof. H. Miyakawa, and H. Harashima, University of Tokyo, for their cooperation in preparing the present manuscript. He is also indebted to Dr. L. R. Bahl and Dr. C. V. Freiman of IBM Thomas J. Watson Research Center and to anonymous reviewers for their helpful comments given to the original manuscript.

References

[1] J. M. Sipress, "A new class of selected ternary pulse transmission plans for digital transmission lines," *IEEE Trans. Commun. Technol.*, vol. COM-13, pp. 366–372, Sept. 1965.
[2] M. R. Aaron, "PCM transmission in the exchange plant," *Bell Syst. Tech. J.*, vol. 41, pp. 99–141, Jan. 1962.

[3] U. Grenander, *Statistical Analysis of Stationary Time Series.* New York: Wiley, 1957.

[4] W. R. Bennett and J. R. Davey, *Data Transmission.* New York: McGraw-Hill, 1965, p. 115.

[5] H. Kaneko and A. Sawai, "Feedback balanced code for multilevel PCM transmission," *IEEE Trans. Commun. Technol.,* vol. COM-17, pp. 554–563, Oct. 1969.

[6] A. Lender, "The duobinary technique for high-speed data transmission," *AIEE Trans. (Commun. Electron.),* vol. 82, pp. 214–218, May 1963.

[7] ——, "A synchronous signal with dual properties for digital communications," *IEEE Trans. Commun. Technol.,* vol. COM-13, pp. 202–208, June 1965.

[8] ——, "Correlative level coding for binary-data transmission," *IEEE Spectrum,* vol. 3, pp. 104–115, Feb. 1966.

[9] P. J. vanGerwen, "Efficient use of pseudo-ternary codes for data transmission," *IEEE Trans. Commun. Technol.* (Concise Papers), pp. 658–660, Aug. 1967.

[10] A. Lender, "Correlative data transmission with coherent recovery using absolute reference," *IEEE Trans. Commun. Technol.,* vol. COM-16, pp. 108–115, Feb. 1968.

[11] F. K. Becker, E. R. Kretzmer, and J. R. Sheehan, "A new signal format for efficient data transmission," in *1966 NEREM, Conf. Rec.,* pp. 240–241.

[12] A. M. Gerrish and W. J. Lawless, "A new wideband partial-response data set," in *Conf. Rec., 1970 IEEE Int. Conf. Communications,* pp. 4.1–4.7.

[13] A. Lender, "Correlative digital communication techniques," *IEEE Trans. Commun. Technol.,* vol. COM-12, pp. 128–135, Dec. 1964.

[14] E. R. Kretzmer, "Binary data communication by partial-response transmission," in *Conf. Rec., 1965 IEEE Int. Conf. Communications,* pp. 451–455.

[15] F. K. Becker, E. R. Kretzmer, and J. R. Sheehan, "A new signal format for efficient data transmission," *Bell Syst. Tech. J.,* vol. 45, pp. 755–758, May–June, 1966.

[16] E. R. Kretzmer, "Generalization of a technique for binary data communication," *IEEE Trans. Commun. Technol.,* Concise papers, vol. COM-14, pp. 67–68, Feb. 1966.

[17] H. Kobayashi and D. T. Tang, "Application of partial-response channel coding to magnetic recording systems," *IBM J. Res. Develop.,* vol. 14, pp. 368–375, July 1970.

[18] ——, "A new coding approach to digital magnetic recording," in *1971 IEEE Int. Magnetic Conf. Dig.,* p. 6.8.

[19] H. Kobayashi, "Application of probabilistic decoding to digital magnetic recording systems," *IBM J. Res. Develop.,* vol. 15, pp. 64–74, Jan. 1971.

[20] ——, "On optimal processing of digital magnetic recording data," *in 1971 IEEE Int. Magnetic Conf. Dig.,* p. 6.7.

[21] H. Kobayashi and D. T. Tang, "On decoding and error control for correlative level coding system," in *1970 Int. Symp. Information Theory,* Noordwijk, the Netherlands, pp. 43–45.

[22] H. Kobayashi, "Correlative level coding and maximum likelihood decoding," in *1970 Canadian Symp. Communication. Dig.,* pp. 61–62; also *IEEE Trans. Inform. Theory,* vol. IT-17, pp. 586–594, Sept. 1971.

[23] G. D. Forney, Jr., "Error correction for partial-response modems," in *1970 Int. Symp. Information Theory,* Noordwijk, the Netherlands, pp. 34–35.

[24] ——, "Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inform. Theory,* to be published.

[25] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory,* vol. IT-13, pp. 260–269, Apr. 1967.

[26] ——, "Convolutional codes: The state diagram approach to optimal decoding and performance analysis for memoryless channels," Jet Propulsion Lab., Calif. Inst. Tech., Pasadena, Space Program Summary 57–58, vol. 3, pp. 50–55, Aug. 1969.

[27] J. K. Omura, "On the Viterbi decoding algorithm," *IEEE Trans. Inform. Theory,* vol. IT-15, pp. 177–179, Jan. 1969.

[28] ——, "On optimum receivers for channels with intersymbol interference," in *1970 Int. Symp. Information Theory,* Noordwijk, the Netherlands, p. 54.

[29] ——, "Optimal receiver design for convolutional codes and channels with memory via control theoretic concepts," to be published.

[30] R. Bellman, *Dynamic Programming.* Princeton, N. J.: Princeton Univ. Press, 1957.

[31] L. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication.* New York: McGraw-Hill, 1968, p. 98.

[32] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering.* New York: Wiley, 1968, p. 83.

[33] J. F. Gunn and J. A. Lombardi, "Error detection for partial-response systems," *IEEE Trans. Commun. Technol.,* vol. COM-17, pp. 734–737, Dec. 1969.

[34] H. Kobayashi and D. T. Tang, "On decoding of correlative level coding systems with ambiguity zone detection," *IEEE Trans. Commun. Technol.,* vol. COM-19, pp. 467–477, Aug. 1971.

[35] J. W. Smith, "Error control in duobinary systems by means of null zone detection," *IEEE Trans. Commun. Technol.,* vol. COM-16, pp. 825–830, Dec. 1968.

[36] A. M. Gerrish and R. D. Hawson, "Multilevel partial-response signaling," in *Conf. Rec., 1967 IEEE Int. Conf. Communications,* p. 186.

[37] H. Miyakawa and H. Harashima, "A method of code conversion for digital communication channel with intersymbol interference," *Trans. Inst. Electron. Commun. Eng. Jap.,* vol. 52-A, pp. 272–273, 1969.

[38] ——, "Channel coding for digital transmission," *J. Inst. Electron. Commun. Eng. Jap.,* vol. 53, pp. 1494–1497, Nov. 1970.

[39] F. Kanaya, "Automaton model of digital transmission line and its application to logical code conversion," *T. Inst. Electron. Commun. Eng. Jap.,* vol. 52-A, pp. 492–499, Dec. 1969.

[40] W. H. Kautz, "Fibonacci codes for synchronization control," *IEEE Trans. Inform. Theory,* vol. IT-11, pp. 284–292, Apr. 1965.

[41] D. T. Tang, "Run-length limited codes for synchronization and compaction," IBM T. J. Watson Res. Cent., Yorktown Heights, N.Y., IBM Res. Rep. RC-1883, Aug. 1967.

[42] ——, "Run-length limited codes," presented at the 1969 IEEE Int. Symp. Information Theory, Ellenville, N.Y., Jan. 1969.

[43] ——, "Practical coding schemes with run-length constraints," IBM T. J. Watson Res. Cent., Yorktown Heights, N.Y., IBM Res. Rep. RC-2022, Apr. 1968.

[44] D. T. Tang and L. R. Bahl, "Block codes for a class of constrained noiseless channels," *Inform. Contr.,* vol. 17, pp. 436–461, Dec. 1970.

[45] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.,* vol. 27, pp. 379–423, and 623–656, 1948.

[46] C. V. Freiman and A. D. Wyner, "Optimum block codes for noiseless input restricted channels," *Inform. Contr.,* vol. 7, pp. 398–415, 1964.

[47] J. Riordan, *An Introduction to Combinatorial Analysis.* New York: Wiley, 1958, p. 14.

[48] D. E. Knuth, *The Art of Computer Programming: Vol. 1, Fundamental Algorithm.* Reading, Mass.: Addison-Wesley, 1969, pp. 13, 18, and 78–85.

[49] P. A. Franaszek, "On synchronous variable length coding for discrete noiseless channels," *Inform. Contr.,* vol. 15, pp. 155–164, Aug. 1969.

[50] H. Kobayashi, "Coding schemes for reduction of intersymbol interference in data transmission systems," *IBM J. Res. Develop.,* vol. 14, pp. 343–353, July 1970.

[51] H. Mine, T. Hasegawa, and Y. Koga, "Asynchronous transmission schemes for digital information," *IEEE Trans. Commun. Technol.,* vol. COM-18, pp. 562–568, Oct. 1970.

[52] A. Gabor, "Adaptive coding for self-clocking recording," *IEEE Trans. Electron. Comput.* (Short Notes), vol. EC-16, pp. 866–868, Dec. 1967.

[53] P. A. Franaszek, "Sequence-state methods for run-length-limited coding," *IBM J Res. Develop.,* vol. 14, pp. 376–383, July 1970.

[54] J. K. Wolf, "On the application of some digital sequences to communication," *IEEE Trans. Commun. Syst.,* pp. 422–427, vol. CS-11, Dec. 1963.

[55] E. Gorog, "Redundant alphabets with desirable frequency spectrum properties," *IBM J. Res. Develop.,* vol. 12, pp. 234–240, 1968.

[56] A. Croisier, "Introduction to pseudoternary transmission codes," *IBM J. Res. Develop.,* vol. 14, pp. 354–367, July 1970.

[57] T. M. Chien, "Upper bound on the efficiency of dc-constrained codes," *Bell Syst. Tech. J.,* vol. 49, pp. 2267–2287, Nov. 1970.

[58] P. A. Franaszek, "Sequence-state coding for digital transmission," *Bell Syst. Tech. J.,* vol. 47, pp. 143–157, Jan. 1968.

[59] G. Ungerboeck, "Nonlinear equalization of binary bipolar signals in Gaussian noise," in *Conf. Rec., 1971 IEEE Int. Conf. Communications,* pp. 21.31–21.36.

Hisashi Kobayashi (S'66–M'68), for a photograph and biography please see page 280 of the June 1971 issue of this TRANSACTIONS.