A. Croisier

Introduction to Pseudoternary Transmission Codes

Abstract: This paper describes many of the pseudoternary (PT) codes (twinned, bipolar, partial response, etc.) used in data transmission. In these, binary information is transmitted through three-level rather than binary pulse codes for controlling the power distribution in the frequency spectrum, improving clock recovery, allowing error detection or for just increasing the binary data rate. Linear and nonlinear PT codes are considered, the latter being divided into alphabetic and nonalphabetic codes. Among the nonalphabetic codes, emphasis is given to the modified bipolar codes used in pulse code modulation systems. Two recently developed codes of this type are described: High Density Bipolar (HDB) and Compatible High Density Bipolar (CHDB). They are particularly suitable for PCM transmission on repeatered lines.

Another nonalphabetic code, the Transparent Interleaved Bipolar (TIB) is presented for the first time. This code features all the advantages of partial-response (or Interleaved Bipolar) signalling and, in addition, guarantees a minimum density of pulses, regardless of the data.

Introduction

Most of the digital information transmitted today is either pure binary data or data that have been encoded in binary digits. It thus seems advantageous to construct line signals with 2^n levels, each signal element representing n binary digits. Although this approach is usually used in the design of high-efficiency modems (modulator-demodulators), we find an increasing number of designs based on three-level pulse-amplitude modulation. A set of rules assigning a three-level signal to a binary message is called a *pseudoternary code* (or sometimes pseudoternary transmission plan).

One of the best known PT codes, *the bipolar code*[1] was introduced by AT & T for the multiplex transmission of telephone signals by pulse code modulation (PCM) on cables. This code is now an almost universal standard to which cable repeaters are produced by the hundreds of thousands.

Another much discussed code is A. Lender's *duobinary*,[2,3] generated by binary signal elements which interfere with each other. The ability of duobinary to reduce the bandwidth requirement of baseband transmission has been the subject of much controversy and has thus created much interest for this class of codes (including the bipolar code), which we term *linear pseudoternary codes.*

Ternary *partial response*[4] is another important example of a linear PT code that is used in many recent data sets.

On the other hand, there is an infinite variety of pseudoternary codes that are not created by interfering binary elements. These *nonlinear* PT *codes* are less tractable mathematically and more difficult to derive systematically, but they often present unique advantages and cannot be ignored. Some of them increase the information rate of a given channel, while others guarantee a minimum number of transitions from which bit timing can be retrieved regardless of the data sequence. A good example is the B6ZS code chosen by AT & T for the 6.3-Mbps digital repeatered lines of the T2 system.[5]

The present paper is mainly a compilation of many known PT codes and an attempt to present them systematically in such a way that the reader may easily check their relative merits with regard to his own needs. The section on nonlinear PT codes ends, however, with the description of some of our own developments (HDB, CHDB, TIB), one of which is published here for the first time (TIB).

Generally speaking, there are four reasons one might prefer pseudoternary to binary transmission. They are listed here in order of decreasing frequency of practical usage:

1) Improving the frequency spectrum (in particular through dc removal).
2) Making clock recovery easier.
3) Introducing redundancy for error checking.
4) Increasing the data rate.

The most important of these properties is undoubtedly the frequency spectrum shaping; therefore, this paper will be mostly concerned with spectral (or, more generally, statistical) properties. The first Section will briefly recall

The author is at the IBM Centre d'Etudes et de Recherches, La Gaude, France.

some spectral properties of digital codes in general. We shall then present the principal pseudoternary codes and examine their construction—linear codes first and then nonlinear codes. The latter will include the new material on TIB codes.

Important properties of pulse amplitude modulation (PAM)

• *Definitions, Nyquist band*

A multilevel pulse-amplitude-modulated (PAM) signal can be described by the formula

$$S(t) = \left[\sum_n a_n\, \delta(t - nT)\right] * S_0(t), \tag{1}$$

where * denotes the operation of convolution in the time domain, a_n is the amplitude assigned to pulse n, δ is Dirac's distribution and $S_0(t)$ the shape of the elementary pulse. In our case, a_n will be assumed to take the values -1, 0, $+1$ exclusively.

The Fourier transform of $S(t)$ is

$$S(\omega) = \left[\sum_n a_n e^{-j\omega nT}\right] \cdot S_0(\omega). \tag{2}$$

We see that

1) The effect of $S_0(t)$ amounts to a simple filtering in the frequency domain, which is independent of the code transmitted. The spectral properties of pulse codes are thus best described by assuming that Dirac's impulses are transmitted. The choice of another pulse shape will affect all codes in the same way.

2) As the reader may easily verify,

$$\frac{S}{S_0}\left(\omega + k\,\frac{2\pi}{T}\right) = \frac{S}{S_0}(\omega) \text{ (where } k \text{ is any integer)} \tag{3}$$

and

$$\frac{S}{S_0}\left(\frac{\pi}{T} + \omega\right) = \frac{S^*}{S_0}\left(\frac{\pi}{T} - \omega\right). \tag{4}$$

In other words, as Nyquist discovered in 1928,[6] the frequency spectrum of any modulated impulse train is entirely described by its segment $0 < \omega < \pi/T$; all other segments can be deduced by symmetry or repetition. Hence the "Nyquist bandwidth" $\Delta f = 1/2T$, the minimum bandwidth for the error-free transmission of the amplitudes a_n. (In particular, any code free of dc components will have no frequency component at the signalling rate $1/T$.) This result is used in the design of PAM signalling.

If we consider an ideal low-pass filter of bandwidth $F_{max} = 1/2T$, its impulse response is

$$S_0(t) = \frac{\sin\left(\frac{\pi t}{T}\right)}{\frac{\pi t}{T}}. \tag{5}$$

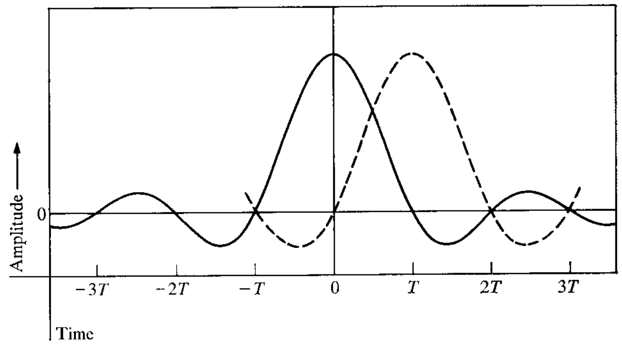
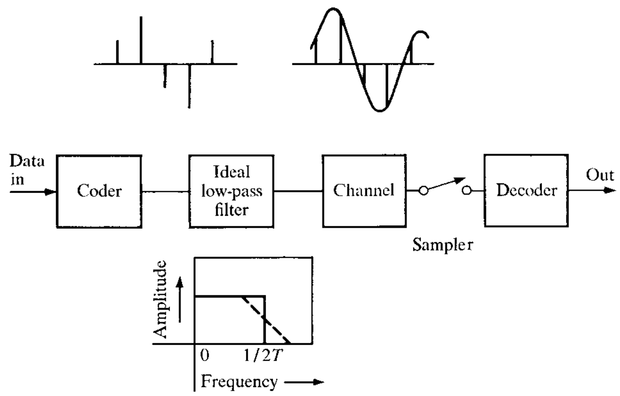
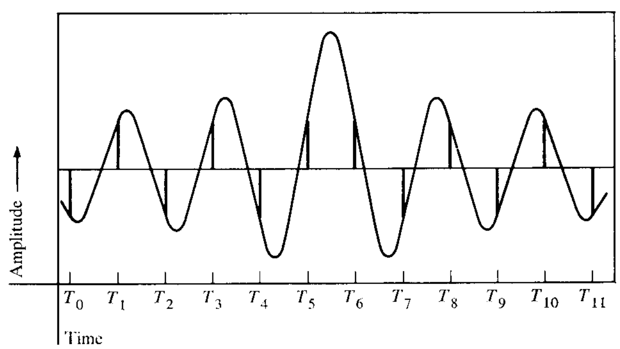


Figure 1 Ideal low-pass filter impulse response ($F_{max} = 1/2T$).

Figure 2 (a) Minimum bandwidth signalling method. (b) Unfavorable data sequence and filtered waveform.



(a)



(b)

This well-known type of pulse, represented in Fig. 1, has zero crossings at all multiples of T except $T = 0$, so that pulses transmitted at the rate $1/T$ do not interfere with each other.

The ideal PAM signalling method is depicted in Fig. 2(a). A coder delivers amplitude coded impulses at the rate $1/T$ to a low-pass filter of bandwidth $1/2T$. The output of this filter is a smooth waveform, the amplitudes of

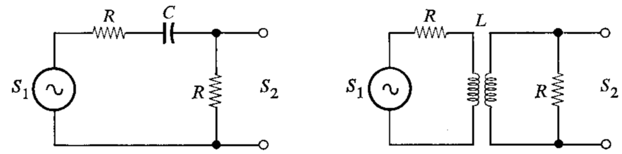


Figure 3 Coupling networks.

which, at the times $t = nT$, are proportional to the amplitudes of the coded impulses. The impulse train may thus be reconstructed by sampling the received signal. (In this ideal case, channel distortions are assumed to be properly compensated.)

A basic difficulty with this signalling method is that the sum $\sum_k |S_0(t + kT)|$ does not converge if $t \neq nT$. It follows that the PAM signal may take unbounded amplitudes *between* the sampling points as illustrated in Fig. 2(b). Obviously, a signal of this type requires a very accurate definition of its sampling instants if it is to be demodulated correctly.

This problem is usually solved by giving some *roll-off* to the low-pass filter as shown by the dotted line in Fig. 2(a). If the roll-off is antisymmetric, the signal element will still have the desired zero crossings and it will decrease more rapidly in the time domain than t^{-1}. We shall discuss this point again in the chapter on linear PT codes. The reader interested in a more thorough treatment of PAM is referred to Lucky, Salz and Weldon's excellent textbook[7] or Gibby and Smith's paper on the Nyquist telegraph theory.[8]

• *Frequency spectrum*

The average *spectral power density* generated by a given code may be obtained by computing the coded signal's autocorrelation function and taking its Fourier transform.

The method of calculating the autocorrelation function is to represent the code by a Markov diagram and compute the successive powers of the transition probability matrix.[9–11] It is not our purpose here to describe this known process in any detail but we shall give a few results for the codes studied. These results will be based on the assumption that the binary data digits are independent and that zeros and ones are equiprobable.

For the practical design of transmission systems, it may also be useful to find the *spectral envelope*, defined as the absolute maximum signal power at any frequency. In some simple cases, such a function can actually be defined, but it is not generally true. "Worst case" low-frequency components in particular are not easily related to the signal distortion produced by an ac coupling network. The concept of *digital sum* explained below has proven much more tractable.

• *Digital sum variation*

Let us consider the simple ac coupling networks shown in Fig. 3. Such networks, which do not transmit direct current, are necessary in most transmission systems. In either case of Fig. 3, we have:

$$S_2(t) = \tfrac{1}{2} S_1(t) - \frac{1}{\tau} \int S_2(t)\, dt, \tag{6}$$

where $\tau = 2RC$ or $\tau = 2L/R$.

The second term on the right-hand side of (6) is the *distortion* introduced by the coupling network; if this distortion is small (or $S_1 \approx S_2$), we see that it is proportional to a primitive of the applied signal.

Since we are interested in the properties of codes independently of the pulse shapes used, we prefer to say that the ac coupling distortion suffered by a coded signal is proportional to the digital sum[12] of the code. By definition, the digital sum is

$$\sigma_N = \sum_{n=M}^{N} a_n, \tag{7}$$

where M is arbitrary, but fixed, and N numbers the last transmitted impulse at the time considered. As we limit our study to codes with stationary properties, σ_n may be either bounded or allowed to grow at the most linearly with N.

In the first case, the average of the signal decreases to zero as the averaging interval increases and the code spectrum has no dc component. This property is shared by most pseudoternary codes and very often is the basic reason for the choice of such a code. When studying a code without a dc component, it is very important to know the maximum *digital sum variation* (DSV) of this code. The DSV, defined as the difference between the upper and lower bounds of σ_N, limits the distortion of any coded signal. In the design of transmission codes, we shall thus try to have the smallest possible DSV.

Linear pseudoternary (PT) codes

• *Basic linear PT codes*

We call a pseudoternary code "linear" if it can be linearly derived from the binary message:

$$a_n = \sum_{k=0}^{K} b_{n-k}\alpha_k, \tag{8}$$

where the coefficients α_k define the code, while $a_n = -1, 0, +1$ is the ternary code level and $b_n = -1, +1$ is the binary encoded data. In this case, the coded sequence can be written:

$$S(t) = \sum_{n=-\infty}^{+\infty} \sum_{k=0}^{K} b_{n-k}\alpha_k S_0(t - nT) \tag{9}$$

or $$S(t) = \sum_{m=-\infty}^{+\infty} b_m \sum_{k=0}^{K} \alpha_k S_0[t - (m + k)T]. \qquad (9a)$$

Thus, a linear pseudoternary code is actually a particular case of a binary code in which the signal element $S_0(t)$ has been replaced by

$$S_0'(t) = S_1(t) * S_0(t), \qquad (10)$$

where $S_1(t)$ is the sequence of impulses

$$S_1(t) = \sum_{k=0}^{K} \alpha_k\, \delta(t - kT). \qquad (11)$$

A linear PT encoding is thus equivalent to a filtering operation. The frequency response of the equivalent filter is given by the Fourier transform of S_1

$$S_1(\omega) = \sum_{k=0}^{K} \alpha_k e^{-i\omega kT} \qquad (12)$$

In order for the a_n to have only three possible values for any sequence, it is necessary that there be only two $\alpha_k \neq 0$ and that they be either equal or opposite. We thus have two basic linear PT codes, each containing two pulses in its signal element:

a) *Twinned binary*

$$\alpha_0 = -\tfrac{1}{2}, \quad \alpha_1 = +\tfrac{1}{2} \qquad (13)$$

b) *Duobinary*

$$\alpha_0 = +\tfrac{1}{2}, \quad \alpha_1 = +\tfrac{1}{2}. \qquad (14)$$

Another example would be three-level partial-response, in which $\alpha_0 = -1/2$, $\alpha_1 = 0$, $\alpha_2 = +1/2$, but this code can be deduced from twinned binary by interleaving, a general method for modifying pulse code properties, which will be explained later.

Twinned binary[1] is free of dc since each binary signal element $S_1(t)$ is composed of a positive and a negative pulse of equal amplitudes. If one constructs coded sequences, it becomes obvious that the zero level can be maintained indefinitely, but the first pulse that follows a positive pulse is always negative (see Fig. 4). The DSV of twinned binary equals one unit, the smallest possible value for any three-level code; therefore, twinned binary is very insensitive to low-frequency removal. This is one of the reasons for which the bipolar code, a modified twinned binary, is so widely used for PCM transmission on cables. The other reason is the ease of clock recovery from the full-wave rectified and clipped signal. In addition, a single transmission error will always create a "bipolar violation," i.e., a pulse of the same polarity as the preceding one. This property does not give a good error-detection scheme, but it allows one to monitor the average error rate without having any information on the transmitted data.

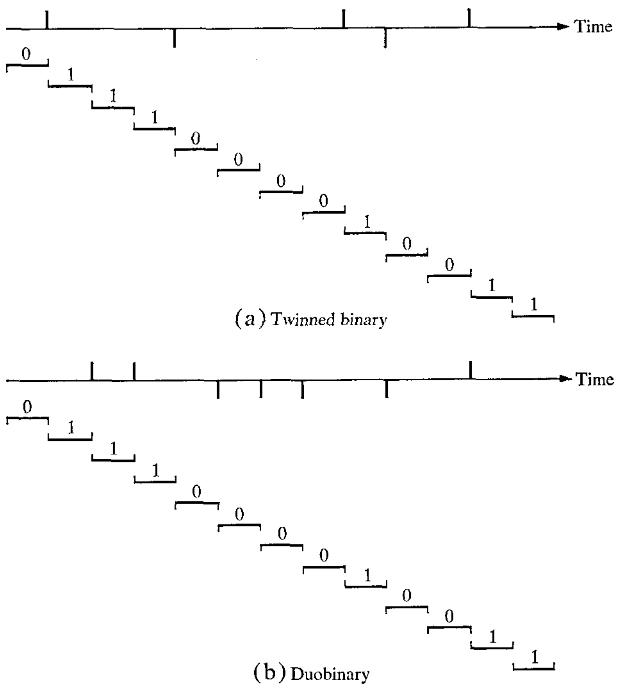


Figure 4 Construction of linear PT-coded sequences from the composite signal elements.

Duobinary,[2] on the contrary, has a dc component that is the same as the corresponding binary code. It is characterized by the absence of direct transitions between levels $+1$ and -1.

The spectral density functions of linear PT codes are easy to compute since we know that the frequency response of the equivalent filter is given by (12). The spectral density is then simply $|S_1(\omega)|^2$.

These well-known results are shown in Fig. 5. Both functions are periodic in the frequency domain, but for any transmitted sequence only the cross-hatched area in Fig. 5 is not redundant. Hence only this area needs to be transmitted.

Note that the nonredundant frequency range depends only upon the signalling frequency and is the same for duobinary, twinned binary, straight binary or ternary coding. The claim once made[3] that duobinary reduces by a factor of two the minimum bandwidth requirements may have resulted from a confusion between the contributions of the code and those of the signal element to the shape of the frequency spectrum. By choosing Dirac impulses as signal elements, we isolate the code properties and the confusion disappears. We shall, however, briefly go back to bandlimited signal elements in order to show which new kind of tradeoff is introduced by the linear pseudoternary codes.

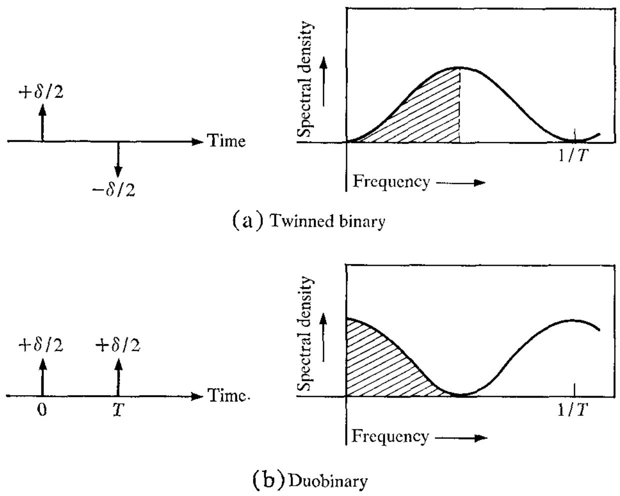


Figure 5 Signal elements and spectral density functions of linear PT codes.

If we consider the ideal system of Fig. 2 in the case of binary, twinned binary and duobinary codes we have the bandlimited signal elements:

Binary

$$S_B(t) = \frac{\sin\left(\frac{\pi t}{T}\right)}{\frac{\pi t}{T}}, \tag{14}$$

Twinned binary

$$S_T(t) = \frac{\sin\left(\frac{\pi t}{T}\right)}{\frac{\pi t}{T}} - \frac{\sin\left(\frac{\pi t}{T} - \pi\right)}{\frac{\pi t}{T} - \pi}$$

$$= \frac{\sin\left(\frac{\pi t}{T}\right)}{\pi}\left[\frac{1}{\frac{t}{T}} + \frac{1}{\frac{t}{T} - 1}\right], \tag{16}$$

Duobinary

$$S_D(t) = \frac{\sin\left(\frac{\pi t}{T}\right)}{\frac{\pi t}{T}} + \frac{\sin\left(\frac{\pi t}{T} - \pi\right)}{\frac{\pi t}{T} - \pi}$$

$$= \frac{\sin\left(\frac{\pi t}{T}\right)}{\pi}\left[\frac{1}{\frac{t}{T} - \left(\frac{t}{T}\right)^2}\right]. \tag{17}$$

$|S_B|$ and $|S_T(t)|$ both decrease asymptotically as $1/t$, while $|S_D(t)|$ decreases as $1/t^2$. The integrals of $|S_B|$ and $|S_T|$ do not converge. Practically, this means that with unfavorable data sequences, the binary and twinned binary signals may take very large values between sampling points, which is not the case for duobinary.

It is thus possible to transmit any duobinary sequence in *exactly* the Nyquist bandwidth for binary symbols while twinned binary, like straight binary or ternary, signalling always needs *a little more* than the Nyquist bandwidth. How much more depends on the precision with which we can design the circuits, particularly the output filter. In practice, an excess bandwidth of about 25 percent has been achieved, for example, with digital echo modulation techniques.[13,14] Since duobinary requires three-level signalling, a fair comparison of bandwidths must be made with ternary rather than binary transmission. With two ternary digits, one can represent three bits. The ratio of the bandwidths of duobinary and straight signalling is thus

$$\frac{3}{2 \times 1.25} = 1.2 \tag{18}$$

in favor of the straight ternary system.*

We see that the use of duobinary (or partial response) actually decreases the absolute performance limits of three-level pulse transmission. However, some designers may rightfully choose these schemes for their simplicity of realization.

- *Precoded linear PT codes*

The basic linear codes described in the previous Section suffer a common drawback: in either case the erroneous reception of one ternary element may result in a chain of errors when it is decoded back into the binary form. This effect can be avoided by precoding the binary data.

In the case of twinned binary, we see that pulses are generated whenever the coded signal switches from a logical ZERO to a logical ONE condition and vice versa. Let us transform the input data in such a way that there is a transition in the transformed data for each ONE in the input. After encoding in *twinned binary* we get a pulse (either positive or negative) for each data ONE and an empty slot for each data ZERO. This construction defines the *bipolar code* widely used for PCM transmission on cables.

The bipolar code can be directly generated from the input data with the following rules:

1) Each data ZERO is represented by an empty slot.
2) Each data ONE is represented by a pulse.
3) Positive and negative pulses alternate regardless of the number of empty slots between them.

* The same conclusion applies to interleaved bipolar or ternary partial response as described in the section on *interleaved linear PT codes.*

Decoding of the bipolar signal is easily achieved by full-wave rectification and clipping at one-half the peak amplitude. There is no error propagation. The same precoding operation applies to duobinary, which also allows decoding by full-wave rectification.

Encoders for all four cases are shown in Fig. 6. As we can see from these circuit diagrams, precoding is a very inexpensive operation and is entirely justified by the simplification it brings in the decoding process.

Interleaved linear PT codes

Another interesting modification of transmission codes in general is *interleaving*.[1] This operation is described by Fig. 7. We see from this diagram that odd bits and even bits are coded separately and interleaved. Usually both coders are identical.

The purpose of interleaving is to modify the signal frequency spectrum. In random binary data there is no correlation between odd and even bits; consequently, the code autocorrelation function is just twice that of a single coder. On the other hand, in interleaving, each coder operates at one-half of the total rate. Consequently, each spectral component of the original code has its frequency reduced by a factor of two when interleaving is used.

Interleaved linear PT codes are most easily generated by the circuits of Fig. 6 in which the delay element duration is increased from T to $2T$. Figure 8 shows the effect of interleaving on the frequency spectrum of the bipolar code. The advantage of interleaving in this case is obviously to create a spectral zero at the Nyquist rate in order to make filtering easier and to make it possible to transmit at *exactly* the Nyquist rate. The interleaved bipolar code thus combines the practical advantages of both bipolar and duobinary.

Interleaving can be pursued to a higher degree in order to create spectral "holes" *within* the Nyquist bandwidth, but this possibility does not yet seem to have found a practical application.

The most publicized application of interleaving is Becker, Kretzmer and Sheehan's three-level *partial response* technique,[4] which is simply a form of interleaved twinned binary. This code, which has been described several times by unrelated authors,[15,16] has of course the same spectrum as interleaved bipolar.*

Partial response is attractive because of its simplicity, but it cannot perform as well as straight ternary signalling combined with highly efficient transmission (such as digital echo modulation) for the same reasons as were invoked for duobinary.

* The name "partial response" does not apply only to the three-level variety with which we are concerned here. A more general description will be found in Ref. 17 and 18.

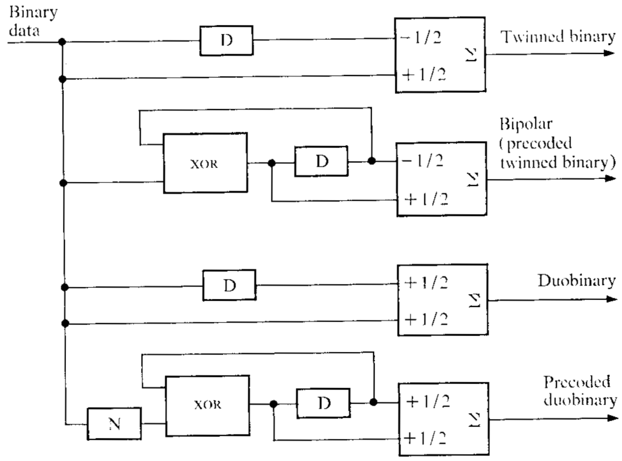


Figure 6 Encoders for linear pseudoternary codes. D = one bit delay; N = inverter; XOR = exclusive or; Σ = summing amplifier.

Figure 7 Arrangement for interleaved coding.

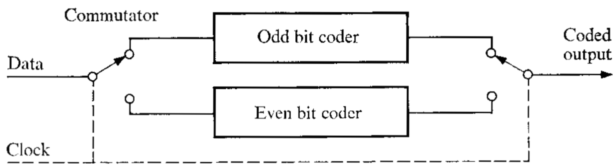
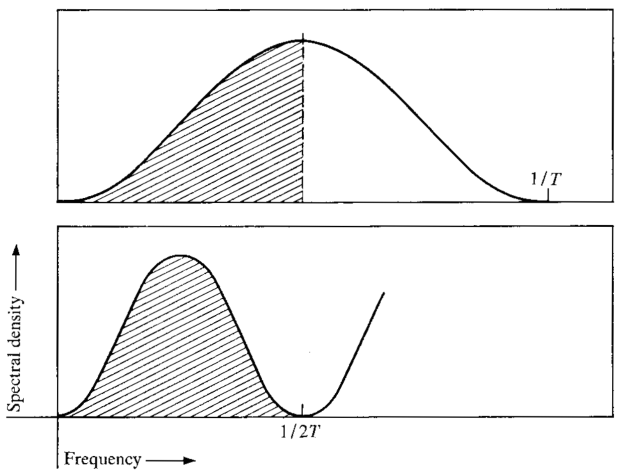


Figure 8 Effect of interleaving on the bipolar code spectrum.



A drawback of interleaving is the doubling of the digital sum variation (DSV). In other words, interleaved signals suffer twice as much distortion as their noninterleaved counterparts when transmitted through coupling transformers. The DSV of partial response is two, which

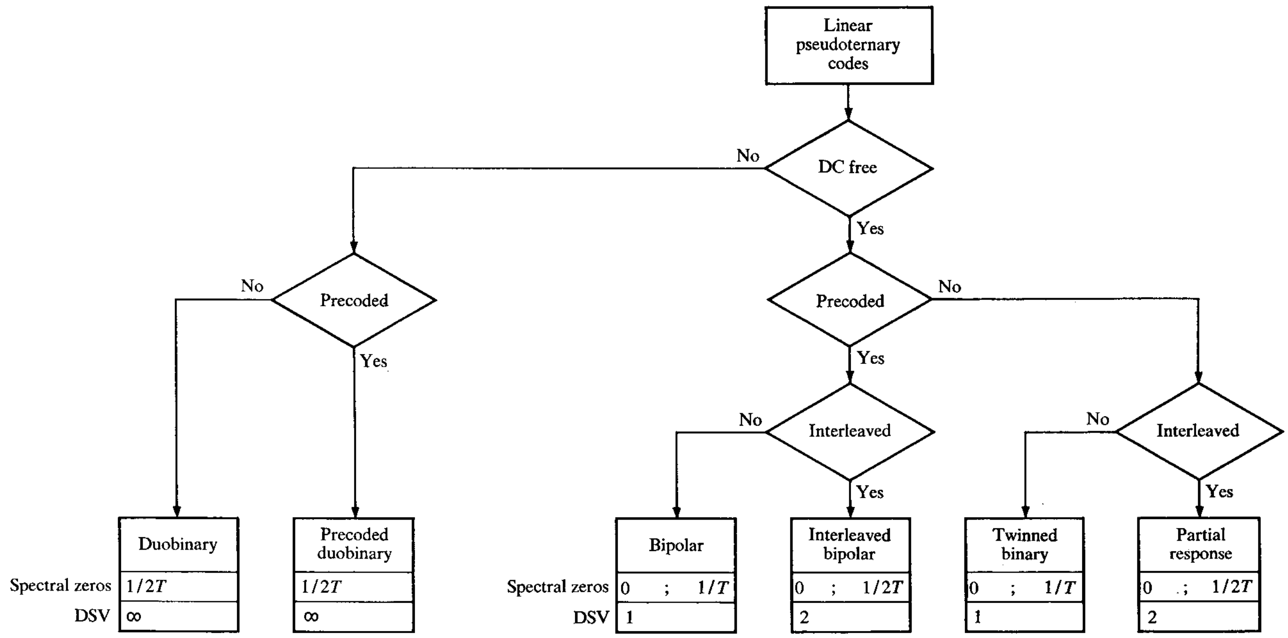


Figure 9 Relations among linear pseudoternary codes.

is still reasonable, but interleaving codes with a higher DSV might be quite difficult.

- *Summary of linear PT codes*

The codes just reviewed are displayed in Fig. 9. This graph is self-explanatory. It recalls the various options that lead to a particular code. The position of the spectral zeros and the DSV appear under each code name.

Nonlinear pseudoternary codes

The domain of nonlinear pseudoternary codes is much wider than that of the preceding linear codes; we shall therefore limit ourselves to codes with practical significance in data communications. The first distinction to make is between alphabetic and nonalphabetic codes.

- *Alphabetic PT codes*

By the designation "alphabetic PT codes" we mean those pseudoternary codes whose digits are grouped in "characters" of n consecutive digits, encoding being done one character at a time. If a ternary character is n digits long, there are 3^n possible characters. For coding, the binary data are divided into binary characters of m bits each; to each binary character corresponds one or several possible ternary characters. Obviously, $2^m \leq 3^n$, or $m \leq 1.58n$.

We see that alphabetic pseudoternary codes may potentially increase the data rate up to 58 percent over straight binary. On the other hand, alphabetic coding requires character synchronization and thus synchronization recovery circuits, which can be complex and may not work on every coded sequence.

- *Alphabetic codes without rate increase*

The most publicized of these codes is Paired Selected Ternary[19] (PST), in which a group of two bits is represented by a group of two ternary digits. The purpose of this type of encoding is to simultaneously eliminate dc and the ternary character 00. The characters to use in each case are defined in the following table:

Binary	*Ternary*
0 0	− +
0 1	0 + or 0 −
1 0	+ 0 or − 0
1 1	+ −

We see that there are two possible ternary characters encoding the binary combinations 01 and 10. In either case one of the characters gives a positive contribution to the digital sum, while the other gives a negative contribution. The dc is balanced by alternating the two kinds of optional characters.

PST has found practical use in very-high-speed PCM systems.[20] The drawbacks of PST relative to bipolar transmission are a factor-of-three increase in the DSV and a factor-of-1.5 increase in the average transmitted power. We shall see that the design goals of PST are better reached with nonalphabetic codes.

There are some variations of the PST code, but we do not know of a fundamentally different alphabetic PT code that is not designed to provide a data rate increase.

• *Alphabetic PT codes with rate increase*

If we wish to approach the maximum possible increase in speed, the best (yet comparatively simple) method is to use two ternary digits to encode three bits of data. The transmission rate is then 1.5 that of binary, which is about 95 percent of the theoretical limit. In fact, only one of the nine combinations of two ternary digits is not used. Depending on the type of transmission hardware, the choice of the unused character may not be arbitrary; very often the character 00 may be undesirable. This 3B-2T code is the most efficient three-level transmission scheme when dc is not objectionable.

But we may also be more ambitious and attempt to eliminate dc while transmitting at an increased data rate. This idea has brought the development of two very interesting codes which represent four binary digits with three ternary digits. Both are based on the following decomposition of the 27 possible three-digit ternary characters.

1) Character 000 is eliminated from the code.
2) The following six characters have a balanced digital sum and can be used without restriction to represent six binary characters:

$$
\begin{array}{lll}
0+- & +0- & +-0 \\
0-+ & -0+ & -+0.
\end{array}
$$

3) The remaining 20 ternary characters have a nonzero digital sum (ten positive and ten negative). They can be paired to represent ten different binary characters. With each binary character, one will associate either the positive or the negative ternary character so as to best compensate the current digital sum.

We see that a total of $6 + 10 = 16$ binary characters can be represented, which corresponds exactly to the number of combinations of four binary digits

The first such code is the British-developed 4B-3T (Ref. 21) in which the characters in each pair are inverses of each other; e.g.,

$$
\begin{array}{llll}
+++ & \text{and} & --- & \\
+\,0\,0 & \text{and} & -\,0\,0\,, & \text{etc.}
\end{array}
$$

This 4B-3T code fulfills its purpose of balancing dc and increasing the transmission rate; however, its digital sum can vary over seven units (DSV = 7), which means a dc removal distortion seven times that of bipolar. A very similar code, described by Franaszek, uses a more sophisticated attribution of ternary code words and is able to reduce the DSV to 5 units, which is a little better. Franaszek's MS43 code is generated under constant monitoring of the digital sum, which is brought back to only one of four possible levels at the end of each code word. Table 1, which is taken from Ref. 12, gives the choice of the code words in each case of binary input and digital sum level at the beginning of the code word. Franaszek's paper gives a method for the derivation of both these codes and variable-length codes (see below).

Table 1 MS43 coding table. (From Ref. 12).

Binary Equivalent	*Digital sum*		
	1	*2 or 3*	*4*
0000	+ + +	− + −	− + −
0001	+ + 0	0 0 −	00 −
0010	+ 0 +	0 − 0	0 − 0
0100	0 + +	− 0 0	− 0 0
1000	+ − +	+ − +	− − −
0011	0 − +	0 − +	0 − +
0101	− 0 +	− 0 +	− 0 +
1001	0 0 +	0 0 +	− − 0
1010	0 + 0	0 + 0	− 0 −
1100	+ 0 0	+ 0 0	0 − −
0110	− + 0	− + 0	− + 0
1110	+ − 0	+ − 0	+ − 0
1101	+ 0 −	+ 0 −	+ 0 −
1011	0 + −	0 + −	0 + −
0111	− + +	− + +	− − +
1111	+ + −	+ − −	+ − −

As stated before, the digital sum can take only four different "terminal" levels at the end of a code word, but two additional levels may be reached within the code words with such characters as $+-0$ or $+--$ when starting from extreme terminal levels. This gives a total of six levels; hence, a DSV of 5.

Another class of alphabetic codes has been developed by Gorog[22] for the transmission of information in *n*-ary (in particular ternary) systems with constraints in the frequency spectrum.

• *Variable-length alphabetic codes*

In these sophisticated coding schemes, the code words may have several different lengths. To be practical, variable-length codes must transmit data at a constant rate; i.e., each code word must carry a number of data bits proportional to its length

In Franaszek's VL43 code,[12] for instance, data is encoded in blocks of 4 or 8 bits represented with code words of 3 or 6 ternary digits, respectively.

The algorithm for VL43 encoding can be summarized as follows:

1) Compute the running digital sum.
2) Examine the next 4 bits.
3) Look in the table for a 3-digit code word corresponding to these 4 bits and the digital sum.

4) If no code word is found, increase the number of bits examined to 8 and look for a suitable 6-digit code word.
5) Record the selected word and go back to 1).

The VL43 code is claimed to have two advantages over MS43: a digital sum variation of 4 instead of 5 and a higher density of transitions from which timing may be recovered.[12]

It must be emphasized here that many codes which have a state dependent encoding like 4B-3T, MS43 and VL43 can actually be decoded without this state dependence. This is a very important point since, otherwise, errors would tend to propagate.

• *Filled bipolar codes* (*nonalphabetic*)

Many designs for nonalphabetic, nonlinear PT codes have been proposed.[5,12,23–27] The only ones to have found important applications seem to be of the "filled bipolar" type. They have the same purpose as paired selected ternary (PST), namely to eliminate a basic deficiency of the bipolar code, which is to represent a binary ZERO by an empty time slot.* Usually, these codes are not designed for sharp filtering before transmission and do not have a spectral ZERO at $f = 1/2T$. This property may, however, be obtained by interleaving as with linear PT codes.

If a long sequence of ZEROS is encoded in bipolar form, it will result in an equally long time interval without any transmitted energy, and clock recovery from the signal will become impossible. Therefore, the bipolar code is said not to be "data transparent." With practically realized receivers, the maximum "empty interval" allowed for consecutive ZEROS is about 15, an unacceptable restriction in many cases.

The immediate answer to this problem is to fill the long, empty intervals with a special ternary sequence that will maintain the receiving clock in operation. The "filling" sequence must be recognized and eliminated by the receiver. For this reason filling sequences must contain a bipolar violation (i.e., a pulse whose polarity is the same as that of the last previous pulse).

In order to simplify the presentation of filled bipolar codes, we shall adopt the following conventions:

The symbol B represents a normal bipolar coded pulse.
The symbol V represents a bipolar violation pulse.
The symbol 0 represents a bipolar ZERO.
1 and ϕ represent binary ONES and ZEROS.

The normal bipolar encoding law can thus be written:

$$1 \rightarrow B; \qquad \phi \rightarrow 0.$$

* See Section entitled *Precoded linear PT codes*, p. 358.

Johannes et al.[25] consider *modal* and *nonmodal* codes. With our convention, a nonmodal code is one in which the filling sequence is the same regardless of the previously transmitted data. Johannes et al. demonstrate that a nonmodal filling sequence must contain at least four pulses. As an example, they describe the bipolar-with-six-zero-substitution (B6ZS) code whose 6-bit filling sequence is B0VB0V. B6ZS was also disclosed as part of the T2 PCM multiplex system,[5] the reason for choosing it over a modal code is not indicated.

In our opinion a modal scheme would have brought a greater efficiency. Before entering this discussion, let us briefly note that the ZEROS in the B0VB0V sequence are there to prevent two adjacent pulses of the same polarity from occurring. This is because many PCM cable repeaters could not generate such signals.

• *Some recent developments in filled bipolar codes* (*HDB, CHDB*)

The digital sum variation for B6ZS is 3, against 1 for pure bipolar. We have asked ourselves whether a filled bipolar code with a DSV of 2 was possible, and this has led us to design the *High Density Bipolar*,[26] the *Compatible High Density Bipolar*[27] and the *Transparent Interleaved Bipolar** codes (HDB, CHDB and TIB).

The last is not exactly a filled bipolar code and will be discussed separately. HDB and CHDB are characterized by their filling sequences whose length $n + 1$ (which may be chosen arbitrarily) defines the maximum number n of consecutive ZEROS in the coded message. (We call this last number the order of the HDB or CHDB code, and the code of order n is called HDBn or CHDBn).

These sequences are:

HDB	$B\,0\,0\,,\cdots,V$
	or $0\,0\,0\,,\cdots,V$
CHDB	$0\,0\,,\cdots,B\,0\,V$
	or $0\,0\,,\cdots,0\,0\,V.$

As we see, these codes are modal, i.e., they each have more than one possible filling sequence.

The rule for choosing the proper sequence is:

The number of B pulses between two consecutive V pulses must always be odd.

The reader may easily verify that this rule produces violation pulses of alternate polarities and that, consequently, the DSV is only 2. At the same time, HDB and CHDB have fewer pulses (about 1.5, on the average) in the filling sequence and thus require less transmitter power and produce less crosstalk than nonmodal codes.

* Unpublished.

HDB and CHDB are very similar; CHDB may be considered as an improvement of HDB (which was developed first). CHDB coding and decoding hardware is somewhat simpler. In particular the decoder is always the same, regardless of the order of the CHDB code.

Another strong point for HDB and CHDB codes is that single transmission errors are easily detected by checking whether the polarities of the violation pulses effectively alternate.

It is interesting to note that filled bipolar codes can also be described as variable-length alphabetic codes, but this type of description does not highlight their interesting features. This point is illustrated by Table 2, which presents HDB2 as a variable-length alphabetic codes.

- *A new development: the transparent interleaved bipolar codes (TIB)*

We have seen in the second section of this paper that any code without dc is also free of spectral components at the signalling rate $1/T$. We have also seen that interleaving two independently coded sequences would modify the frequency spectrum homotetically, reducing the frequency of each component by a factor of two. Consequently, *any interleaved dc free code has a spectral zero at the Nyquist rate* ($f = 1/2T$).

It would seem attractive to use this property and interleave, for instance, two CHDB codes in order to combine the filtering facility of partial response with the data transparency of CHDB.

The scheme obviously works, but produces a digital sum variation (DSV) of 4, against 2 for either CHDB or partial response.

The next idea would be to interleave a CHDB and a bipolar message. The DSV would be only 3 and the decoder would not have to know which is the CHDB message; it would just assume *both* to be CHDB.

This scheme would already be an improvement, but we intend to show that a family of codes can be constructed which has the following properties:

1) No dc component.
2) No component at the Nyquist rate $f = 1/2T$.
3) A digital sum variation = 2.
4) A bound on the number of consecutive zeros.

In other words, the advantages of partial response and filled bipolar can be combined without increasing the DSV.

The new codes, which we name Transparent Interleaved Bipolar of order n (TIBn) are defined as follows.

Each code of the family is composed of two interleaved pseudoternary sequences or "subchannels."

a) Each sequence normally follows the bipolar encoding law (successive pulses have opposite polarities).

Table 2 A variable-length alphabetic coding table for HDB 2

	Last violation			
	+		−	
	Digital sum		*Digital sum*	
Binary	+1	0	0	−1
1	−	+	−	+
01	0−	0+	0−	0+
001	00−	00+	00−	00+
000	−0−	00−	00+	+0+

Table 3 TIB filling sequences

Polarity of last violation in first subchannel	*Polarity of last transmitted pulse in 1st subchannel*	*Polarity of last transmitted pulse in 2nd subchannel*	*Filling sequence*
+	+	+	0 0, ⋯ , − 0 − +
	+	−	− + − +
	−	+	0 0 − +
	−	−	0 + − +
−	+	+	0 − + −
	+	−	0 0 + −
	−	+	+ − + −
	−	−	+ 0 + −

b) When a number $n + 1$ of consecutive ZEROS is found in the composite pulse train, they are replaced by a special "filling" sequence terminated with a bipolar violation in each subchannel. This filling sequence is:

$0\,0, \cdots, 0\,0\,X\,X\,V\,V,$

where V represents a pulse in bipolar violation (in its respective subchannel) and X is either a ZERO or a normal bipolar pulse chosen in such a way that the number of B pulses between two successive V pulses is odd in each subchannel.

c) The two V pulses in the same filling sequence have opposite polarities.

Depending on the polarity of the last violation and the last pulse in each subchannel, we find the eight possible filling sequences listed in Table 3.

The operation of the TIB codes will be best understood from Fig. 10, which represents a particular message encoded in TIB5 (i.e., the TIB code which generates at

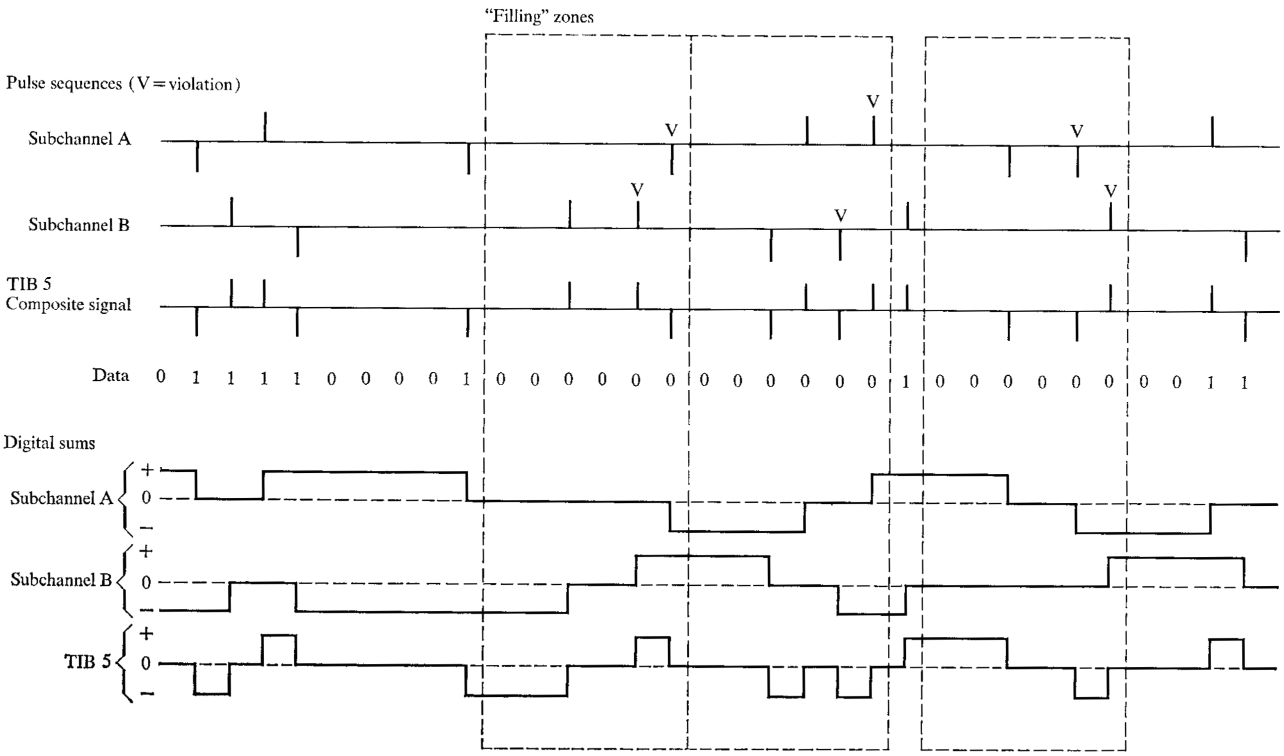
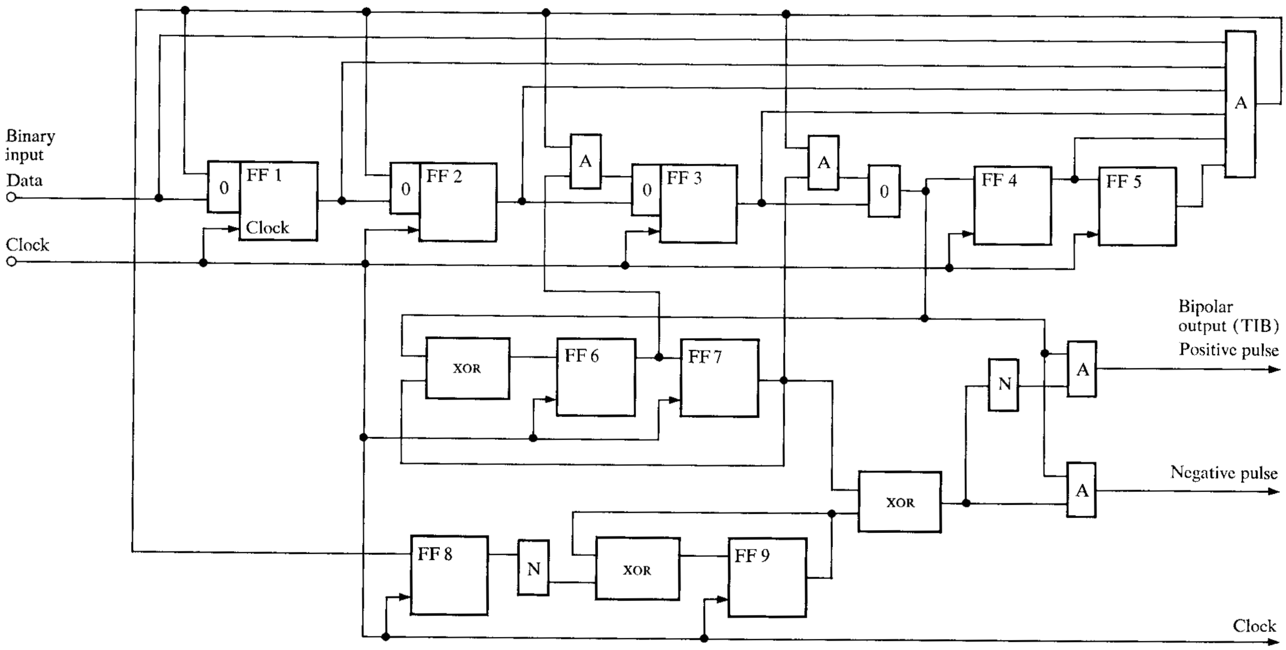


Figure 10 TIB 5 encoding example.

Figure 11 TIB 5 encoder. FF = flip-flop; other symbols as in Figure 6.



most 5 consecutive ZEROS). Figures 11, 12 and 13 represent encoding, decoding and error monitoring circuits for the TIB5 code. (In fact, only the encoder is particular to TIB5; the other two circuits work with any TIB code).

• *Summary of nonlinear PT codes*

Figure 14 groups the codes just described. As in Fig. 9, the positions of the spectral ZEROS and the DSV appear under each code name. The rate increase factor (relative to binary) has been added when applicable. The spectral densities of the PST, B6ZS, MS43 and the various CHDB codes are displayed in Fig. 15. They have been computed from the autocorrelation functions as described earlier in this paper. (The spectrum of TIB has not yet been computed; spectra of HDB codes will be found in Ref. 26).

Conclusions

There is obviously a great variety of pseudoternary codes. The choice of the proper design depends on such particular requirements as freedom from dc, ease of filtering at the Nyquist rate, simple encoding and decoding, and so on. Generally speaking, it seems to us that nonlinear codes offer more possibilities than the linear ones; they allow one to simultaneously control the position of spectral ZEROS, minimize the effects of low-frequency removal, guarantee a minimum pulse density and even increase the data rate in a given bandwidth (relative to binary transmission).

Alphabetic codes, like MS43 or VL43, allow data transmission at greater than the binary Nyquist rate while removing the dc component; they should be used whenever efficient use of the bandwidth can be paid in design complexity.

Modal filled bipolar codes, on the other hand, have statistical properties (digital sum variation, average power, minimum pulse density) which make them easy to detect by simple means; they are particularly suited for PCM transmission on lines with repeaters.

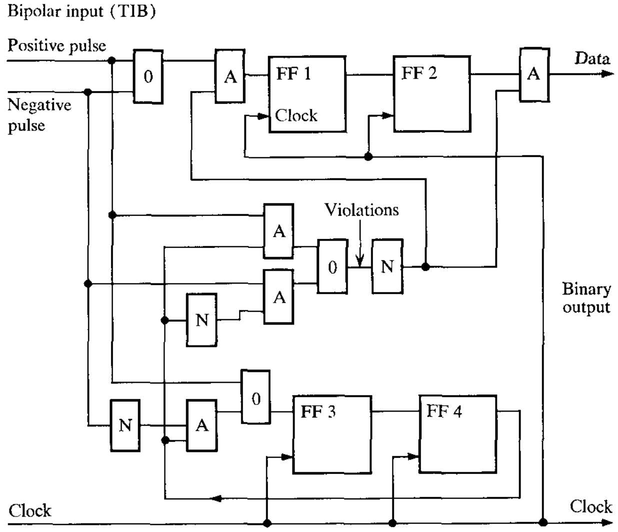


Figure 12 Error monitoring device for any TIB code.

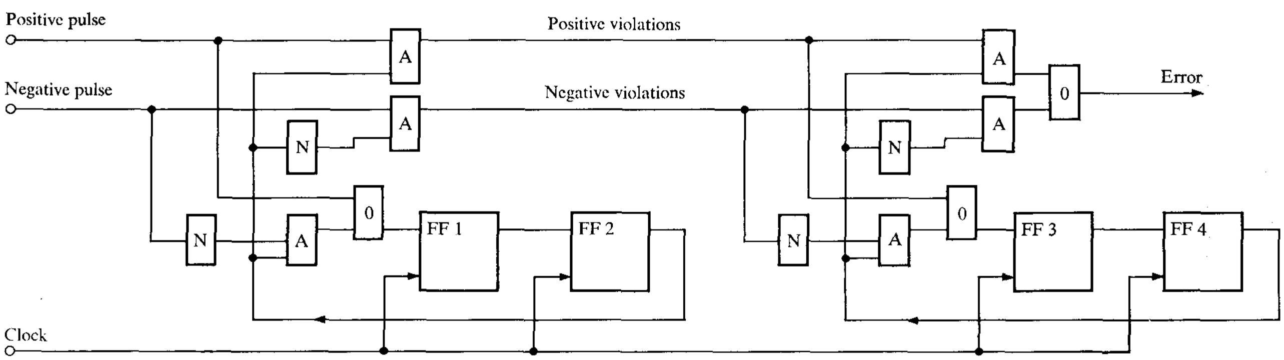
Acknowledgment

The author is indebted to R. Vermot-Gauchy and R. Mariotti for the development of the computer programs used in preparing the frequency spectra shown in Fig. 15.

References

1. M. R. Aaron, "PCM Transmission in the Exchange Plant," *Bell System Tech. J.* **41,** 99 (1962).
2. A. Lender, "Correlative Level Coding for Binary-Data Transmission," *IEEE Spectrum* **3,** 104 (Feb. 1966).
3. A. Lender, "The Duobinary Technique for High Speed Data Transmission," *IEEE Trans. Comm. and Electron.* **CS-5,** 214 (1963). Ref. 62-283).
4. F. K. Becker, E. R. Kretzmer and J. R. Sheehan, "A New Signal Format for Efficient Data Transmission," *Bell System Tech. J.* **45,** 755 (1966).
5. J. H. Davis, "T2: A 6.3 Mb/s Digital Repeatered Line," Int. Comm. Conf. (1969) Boulder, Colo., IEEE Ref. 69CP369 COC p. 34–9.

Figure 13 Universal TIB decoder.



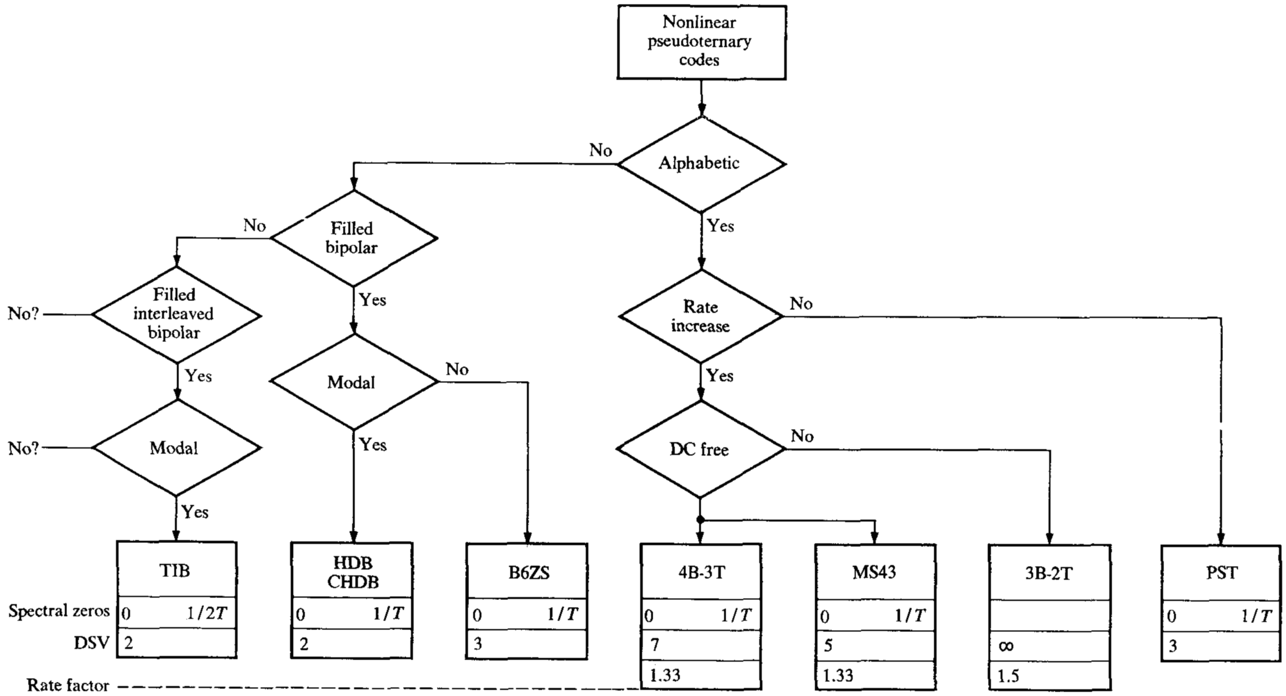
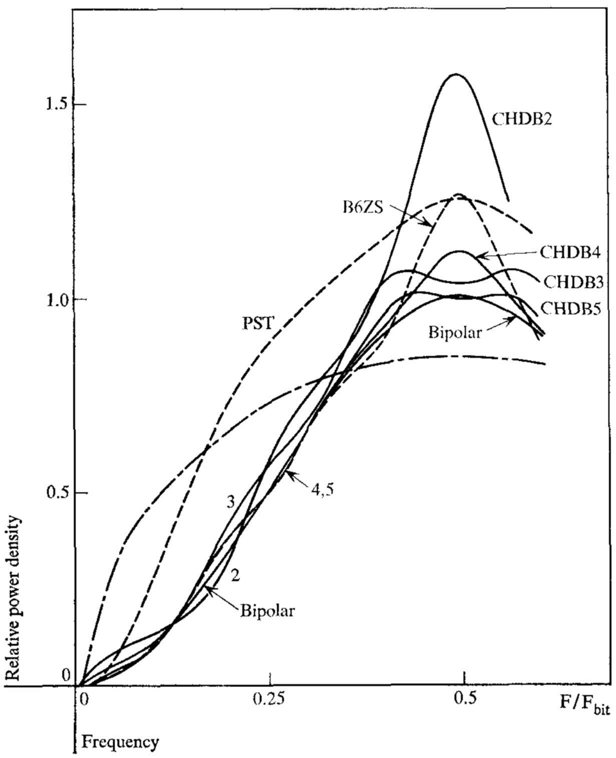


Figure 14 Relations among nonlinear pseudoternary codes.

Figure 15 Spectral densities of various PT codes.



6. H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *AIEE Transactions* **47,** 617 (1928).
7. R. W. Lucky, J. Salz and E. J. Weldon, Jr., *Principles of Data Communication,* McGraw-Hill, New York, 1968.
8. R. A. Gibby and J. W. Smith, "Some Extensions of Nyquist's Telegraph Transmission Theory," *Bell System Tech. J.* **44,** 1487 (1965).
9. L. S. Schwartz, *Principles of Coding, Filtering and Information Theory,* Cleaver-Hume Press, London, 1963, p. 113 ff.
10. W. W. Huggins, "Signal-Flow Graphs and Random Signals," *Proc. IRE* **45,** 74 (1957).
11. L. A. Zadeh, *Proc. IRE* **45,** 1413 (1957).
12. P. A. Franaszek, "Sequence-State Coding for Digital Transmission," *Bell System Tech. J.* **47,** 143 (1968).
13. M. Humbert, "4-Phase Digital Echo Modulation Technique," *CCITT COM SpA* Contribution Nr 150-E, (Oct. 30, 1967).
14. A. Croisier and J. M. Pierret, "High Efficiency Data Transmission Through Digital Echo Modulation," International Communication Conference (1969) Boulder, Colo., IEEE, Ref. 69CP364-COM.
15. E. Gorog and C. M. Melas, "A Digital Data Transmission System With Optimum Bandwidth Utilization," IEEE Tenth Nat. Comm. Symposium, Utica, New York, p. 340 (Oct. 1964).
16. P. J. Van Gerwen, "On the Generation and Application of Pseudo-Ternary Codes in Pulse Transmissions," *Philips Res. Repts.* **20,** 469 (1965).
17. E. R. Kretzmer, "Binary Data Communication by Partial Response Transmission," *First IEEE Comm. Conv. Conference Record,* Paper G-1E, 451 (June 1965).

18. E. R. Kretzmer, "Generalization of a Technique for Binary Data Communication," *IEEE Trans. Comm. Tech.* **COM-14,** 67 (1966).
19. J. M. Sipress, "A New Class of Selected Ternary Pulse Transmission Plans for Digital Transmission Lines," *IEEE Trans. Comm. Tech.* **COM-13,** 366 (1965).
20. I. Dorros, J. M. Sipress and F. D. Waldhauer, "An Experimental 224 Mb/s Digital Repeatered Line," *Bell System Tech. J.* **45,** 993 (1966).
21. A. Jessop, "High Capacity PCM Multiplexing and Code Translation," *IEE Electronics Division Colloquium on Pulse Code Modulation* (4th March 1968) IEE Coll. Digest No. 1968/7, p. 14/2 to 14/5.
22. E. Gorog, "Redundant Alphabets with Desirable Frequency Spectrum Properties," *IBM J. Res. Develop.* **12,** 234 (1968).
23. M. I. Pelekhatyy, "Some New Possibilities for Increasing the Transmission Rate of Discrete Information," *JPRS:* 36,028 TT:66-32462 U. S. Dept. Commerce Clearinghouse, Washington, D. C. (June 16, 1966). [English translation. *Telecommunications and Radio Eng.*, pt. 1, No. 3, 25–32 (March 1966)].
24. R. Kersten, "Signalarten und Signalformen bei der Uebertragung von PCM—Signalen auf Symetrischen Fernsprechkabeln," (Signal Types and Signal Forms in Transmitting PCM Signals on Balanced Telephone Cables, *Archiv der Elektrischen Uebertagung* **22,** 461 (1968).
25. V. I. Johannes, A. G. Kaim and T. Walzman, "Bipolar Pulse Transmission with Zero Extraction," *IEEE Trans. Comm. Tech.* **COM-17,** 303 (1969).
26. A. Falcoz and A. Croisier, "Le code bipolaire à haute densité, un procédé de transmission en bande de base," *Colloque International sur la Téléinformatique,* Paris 1969, Editions Chiron, Paris, Tome 1, p. 54.
27. A. Croisier, "Proposal for a Code Allowing Unrestricted Binary Transmission over Digital Transmission Systems," Contribution to the CCITT Special Study Group D (1969), Ref: Com Sp-D No. 33.

Received October 8, 1969