Apr 1st, 8:00 AM

The Utility of Coded Multilevel Communications Systems

R. F. Filipowsky
International Business Machines Corporation

Follow this and additional works at: http://commons.erau.edu/space-congress-proceedings

Scholarly Commons Citation
http://commons.erau.edu/space-congress-proceedings/proceedings-1968-5th/session-5/3
THE UTILITY OF CODED MULTILEVEL COMMUNICATIONS SYSTEMS

R. F. Filipowski
International Business Machines Corporation
Huntsville, Alabama

Summary

A design tool, called the Utility Chart, was introduced at this Congress in 1967. It is based on the utility, a definition of the efficiency of a communications link. The utility relates the normalized transmission rate (bit density) to the signal-to-noise ratio. This kind of definition of the communications efficiency is based on a paper by R. W. Sanders published in 1960. The definition has been modified to meet more closely the requirements of practical communications designers.

The present paper compares the various methods of coded multilevel communications systems with the help of the utility chart. "Coded communications" is a term introduced by Viterbi in 1959. This term is generally used to designate all kinds of digital communications systems that translate a set of input messages into a set of transmission messages. The translation rule is called a "code" or, more specifically, a "code book."

Multilevel systems, also called quantized pulse amplitude modulation systems, are systems with transmission signals appearing in more than two levels (or states). This characteristic makes the multilevel systems members of the much larger class of nonbinary systems. Binary systems will appear as a special case using only two levels. If the two levels are the positive and negative peak value, bi-level signals of rectangular shape become identical with bi-phase signals.

In a more general sense the paper deals primarily with one-dimensional nonbinary systems; i.e., with systems using more than two members in the set of transmission messages and modulating only one parameter of the corresponding transmission signals. The detection of such one-dimensional signals takes place one at a time.

The conclusions of the paper are in line with many other publications on nonbinary systems. A review of several hundred original papers on this general subject will be published soon. The theoretical analyses indicate that in general the utility of any nonbinary (or higher order systems, as they are also called) approaches the utility of Shannon’s most ideal system as the size of the set of transmission messages increases. This increase in the size of messages can be achieved by permitting more than two states of the same dimension (amplitude, frequency and phase) or by increasing the number of dimension; i.e., by encoding messages into several independent parameters and detecting the message as a whole, rather than detecting the parameters individually. The first class is the one-dimensional coded communications systems, the second class is the multidimensional coded communications systems. The present paper is restricted to a subclass of the first class. The second class will be discussed elsewhere.

Viterbi’s coded phase-coherent communications system, despite the fact that it uses binary elements, must be classified as a multidimensional system.

The practical measurements carried out so far do not substantiate the theoretical improvement of coded communications systems beyond a certain size of the set of transmission messages. This discrepancy indicates that the present theory is still based on rather ideal assumptions. It will be important to refine the theoretical models. The utility chart may be a welcome tool in this effort to make theory and practical measurements coincide.

Introduction

Many variations are known for the process called modulation and many related variations are known for the process called coding. The processes of modulation and of coding are intimately linked with each other and it is one purpose of this paper to explain this interconnection. Several classes of modulation methods have been reviewed in the previous paper on the utility chart. The present paper deals only with the class of digital modulation. Multilevel (or MASK, which means multi-amplitude shift keying) systems are systems using a kind of digital modulation that is usually designated as higher order modulation or nonbinary modulation. It is the second purpose of this paper to give an example of the application of the utility chart to such nonbinary digital systems.

It is a third purpose of this paper to compare the efficiency of a number of recently developed coded multilevel systems and to relate these results to the efficiency of uncoded multilevel systems and to the theoretical efficiency.

Difference Between Modulation and Coding

There is no well-defined discrimination between the terms modulation and coding. They are loosely used as alternate words for the same or for similar processes. Yet for the purpose of the present paper we may reserve each term for one of two rather distinct processes. This can best be shown in Fig. 1 by using as an example a coded communications system that makes use of Viterbi’s 1959 mode of orthogonal coding.
We assume that a digital source, normally a binary source, delivers nonredundant data in words of a given length. In the example chosen for the explanation of Fig. 1, there are words of 4 bits each. The first of the two processes is a digital conversion process, called the encoding process. It is usually combined with a serial-to-parallel translation process. A good example is the conversion of the 4-bit input word to a 7-bit transmission word by the Hamming encoding process explained in Fig. 2. Returning to Fig. 1 we may assume that an eighth bit will be added for service purposes.

Such 8-bit words now proceed at the same rate as the 4-bit input words to the second of the two processes, which may be called the modulation process. This process is frequently combined with a multiplexing process to permit the simultaneous transmission of more than one transmission signal (or waveform). This modulation process is a hybrid process combining digital and analog operations. The purpose of the modulation process is the generation of waveforms with a frequency spectrum that does not contain significant energy outside the limits of the assigned frequency band of the transmission channel. Waveforms meeting this requirement are called bandlimited waveforms. The modulation process is usually but not necessarily executed with the help of two separate entities: a carrier generator and a modulator. The first one produces a sinusoidal signal; the second one modifies one or more of the parameters of this signal in accordance with the content of the 8-bit word delivered from the encoder.

In the example of Fig. 1 we assume that a biorthogonal coded phase-coherent modulation process will be applied. According to this process binary words are modulated to a carrier so that each binary zero changes the phase of the carrier for 180° but each binary one leaves the phase unchanged. To keep the amount of hardware small, two 8-element transmission words are generated on two different subcarriers (or carriers) and multiplexed into the common channel. This is done in preference to the generation of one 128-element word on a single subcarrier. Fig. 3 shows the code book for the 16 bi-orthogonal transmission words. Notice also that the selected mode of operation using two transmission words requires a much smaller WT product (product of transmission bandwidth [BT] with the duration [T] of the transmission word) than the second mode of operation. On the other hand, it is evident that less work is required in the second case than in the first case. It is the task of the designer of a special system to make the tradeoff calculations for finding the most efficient solution for his case. The utility chart will help him in that effort.

Fig. 1 further shows the transmission channel where noise and disturbances are added and where the transmission signals suffer distortions. At the receiving side one must perform the complementary processes of the transmission side. The extractor-demultiplexer-demodulator block of the receiver corresponds to the modulation block of the transmitter. One function of this block is the separation of the transmission signals from any nonrelated signals that do not occupy the same transmission band or the same time slots (extractor function). Another function of this first block in the receiver is the separation of several subchannels, if any multiplexing had been applied in the transmitter (demultiplexer function). The most important function, however, is the demodulator function aiming at the separation of the modulation from the carrier or, in other words, performing a transformation of the spectrum of the transmission signals from the carrier-band range to the low-pass range. The output of this first block will frequently be in the form of analog samples forwarded over parallel lines to the second block in the receiver.

There are many modes of instrumenting the receiver circuits. In the special case used as an example in Fig. 1, one preferred solution will be to correlate the output of each of the two demodulators separately with each of the eight orthogonal binary sequences of the phase-coherent code of Fig. 3. This means that two groups of eight lines will convey the output of all 16 correlators to the second block of the receiver in the form of analog samples.

The second block in the receiver of Fig. 1 is the complementary block to the first block in the transmitter in the same figure. This second receiver block (on the lower left side of Fig. 1) contains a decoder, one or more deciders, and a parallel-to-serial converter. The decoder and decider may be combined into one intermingled unit with hybrid circuits, partly performing analog processes such as finding the largest of many analog samples, and partly performing logical digital operations such as parity checks and Hamming decoding processes.

The result of all the four processes (encoding—modulating—demodulating—decoding) should be a communications system operating in a mode best suited for the case at hand. This paper is restricted to the consideration of systems receiving digital input information, usually in binary form, though the transmission itself takes place with the help of nonbinary sets of bandlimited analog transmission signals. Thus we prefer to call the systems under consideration, "systems with nonbinary modulation methods."

Nonbinary modulation methods are characterized by the fact that one single signal limited in time and frequency (width of its power spectrum) carries more than one bit of information. There are virtually hundreds of variations and tens of classes of nonbinary modulation methods. This makes it difficult for a systems engineer to decide which approach would efficiently solve his particular design problem. He also needs to know how much he gains, either in transmission rate or in power savings or in lower error ratio, when going from a binary to a nonbinary design. He then should determine his cost increase due to the more complex circuitry and he should determine any changes in his system's reliability between a binary and a nonbinary design. At the present state-of-the-art a cross-over point between achievement and cost will occur at rather low orders of such "higher order
systems." Quaternary systems (four phases in place of a bi-phase system, for example) or octo-

tary systems (using eight frequencies in place of the two frequencies of the standard FSK [fre-

quency shift keying] system) have been thoroughly

explored. These are examples of a multiphase shift keying system (MPSK) and a multifrequency shift keying system (MFSK). This paper deals

with MASK (multi-amplitude shift keying) systems and it will be shown that this class of nonbinary systems recently has been successfully explored

for much higher orders (up to \( n = 32 \)).

The optimum combination of coding and modu-

lation has not been investigated in detail. The slow progress towards nonbinary systems apparently is due to the lack of a convenient engineering
tool for evaluating and comparing the various approaches, including the binary approach always as a special case. The slow progress was also

caused by the high cost of hardware. The cost

will ultimately be of negligible importance, due
to the evolution of large scale integrated (LSI)
circuitry. The first cause may be removed by a

further development of the utility chart, as will be shown in the next section.

The Utility Chart

Descriptions of the theory and the applica-
tions of the utility chart have been published
in 1967. The results will be summarized in con-

nection with Fig. 4. Both axes are in a logarithmic scale of decibels.

The horizontal axis represents the bit den-
sity, a measure for the achievement of a communi-
cations link.

\[
D = \frac{R}{B_T} \quad \text{or in decibels:}
\]

\[
D_{db} = 10 \log_{10} \frac{R}{B_T}
\]

\[
R \quad \text{transmission rate in bits per}
\text{second} \quad \text{sec}^{-1}
\]

\[
B_T \quad \text{Transmision bandwidth in}
\text{Hz} \quad \text{sec}^{-1}
\]

The bit density is defined as the ratio between two magnitudes of the same physical dimension (sec\(^{-1}\)); consequently, it is a number without dimension and can therefore be expressed in
decibels. The transmission rate is counted at the input to the system (top left corner in Fig.

1). The transmission bandwidth is the 3-db

bandwidth of the transmission channel, including

any receiver-filters as part of the channel.

(For more detailed definitions the reader is

referred to chapter 4 of Space Communications Systems\(^7\) or "Performance Criteria for the Com-

parison of Modulation Methods\(^8\).)

The vertical axis of Fig. 4 represents the

utility of the link, a measure of the efficiency

in making use of the available transmitter energy.
The utility is defined as:

\[
u = \frac{D}{\sigma} \quad u_{db} = D_{db} - \sigma_{db}
\]

\[
= 10 \log_{10} N_o - 10 \log_{10} E_B
\]

\[
D \quad \text{bit density as defined in}
\text{equation 1} \quad \text{units}
\]

\[
\sigma \quad \text{SNR (signal-to-noise}
\text{power ratio} \quad \text{units}
\]

\[
N_o \quad \text{noise power density} \quad \text{VAS}
\]

\[
P_e \quad \text{energy per bit of input}
\text{information} \quad \text{VAS}
\]

The expressions of bit density and of SNR are

numbers without physical dimensions. They can be readily expressed in decibels. The expressions

\( N_o \) and \( P_e \) carry the physical dimension of energy. They have to be used in equation 2 in identical units (for example, either ergs or volt-ampere-

seconds). Under this condition it is possible to express their ratio in decibels. Actually it has been suggested that \( \sigma \) be called the power

contrast and that the expression \( P_e/N_o \) be called the energy contrast\(^8\), in an effort to gain shorter words for these frequently used terms. Caution

has to be exercised when applying either \( \sigma \) or \( N_o \). Only for normally distributed (Gaussian) noise is it sufficient to specify the noise characteristics

by a single parameter; more complex definitions

are required for more irregular noise statistics.

However, even in such cases it is possible to
define an "equivalent noise power" (\( N_e^* \)) or an

"equivalent noise power density" (\( N_e^{**} \)). (For

more details consult reference 2 or appendix B in
reference 7.)

A utility chart is a plot of \( u_{db} \) as a func-
tion of \( D_{db} \). It is a characteristic of nonbinary

systems that \( u \) is invariably a nonlinear function

of \( D \) while this is not the case of ideal binary

systems. Fig. 4 shows two typical utility curves. The curve for the ideal binary system is a hori-
zontal straight line. This indicates that an

ideal binary system has always the same utility

no matter at what bit density it operates. How-
ever, the absolute value of this utility depends

on the quality of the transmission. This forces

the user of a utility chart to consider one more

magnitude: the error ratio, a quality measure.

For the purpose of the utility chart the error

ratio is defined as:

\[
\bar{e}_B = \frac{\text{number of bits received in error}}{\text{number of bits taken from the source}}
\]

The bar over \( e_B \) indicates that \( e_B \) is an average value. The counts of the numbers defining \( e_B \)
have to be taken over the same interval and the

interval has to be long enough to achieve a con-

stant value for \( e_B \). Notice that the number of

bits received in error has to be counted at the

output of the system where the information is

being delivered to the final sink. This means

that any errors caused by distortions or by lack

of synchronism will be counted as well as all
the errors caused by noise. Such additional systems errors or processing errors are usually not considered in analytical models of the system. On the other hand it is well known that many coded systems use error correcting codes. In such cases there will be a rather high number of errors at the output of the demodulator (or decoder), but many of those will be corrected in the decoder. Because the quality criterion for the utility chart is defined as the error ratio at the output of the total system, one can see that only the uncorrected errors will be counted.

The error ratio for an ideal binary system was derived in 1958. In terms of the magnitude used in the utility chart, the error ratio may be expressed as:

\[ \frac{E_B}{N_0} = Q \left( \frac{2^D}{u} \right) \quad \text{with } Q(x) \]

\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} \, dt \]

Assuming \( Q(3.1) = 10^{-3} \) in equation 4 gives \( x = 3.1 \) and yields the constant value \( u = 0.208 \) or \( u = -6.8 \) db for the utility of the ideal binary system. This line is plotted in Fig. 4. For smaller error ratios it would be lower and for larger error ratios it would be higher.

The second typical utility curve shown in Fig. 4 is the curve for C. E. Shannon's ideal communications system published in 1948. This mathematical model does not specify any particular coding and modulation system but rather attempts to arrive at an upper limit of the utility that can not be exceeded no matter how cleverly any system may be designed in the future. It is indeed gratifying to notice that no system known so far has ever exceeded the utility of this upper limit of Shannon's ideal system. When expressed in the terminology of the utility chart one arrives at the following equation for Shannon's upper bound:

\[ u_{db} = p_{db} - 10 \log_{10} \left( 2^D - 1 \right) \quad (5) \]

This equation has two approximate expressions that may be applied for very small or very large values of \( D \) respectively:

\[ u_{db} = -10 \log_{10} \left( 0.6931 + 0.2404D \right) \]

\[ + 0.035D^2 + \ldots \]  \quad (5a)

\[ u_{db} = p_{db} - 3.01D \]  \quad (5b)

From equation 5a one may learn that the largest value for \( u_{db} \) is 1.592 db. It is reached for \( D = 1 \). From equation 5b one may learn that for very large values of \( D \) the utility in decibels becomes about three times the numerical value of \( D \), but with a negative sign. This means that a system transmitting at a rate about ten times the bandwidth of the channel (\( B = 10 \)) can never have a utility larger than \( -10 \) db; it must have an efficiency of 1 thousandth or less. The curve representing the utility of Shannon's upper bound of any average power-limited communications system is shown in Fig. 4. Notice that, according to Shannon’s theory, this ideal system works completely error-free. Yet to achieve this feat would require a coding delay (code word lengths) of infinity. Naturally, certain assumptions about the characteristics of the noise and the channel have to be fulfilled to meet the requirements of the theorem.

It is one of the advantages of the utility chart that it makes possible the comparison in the same plot of theoretical models with practical measurements of equipment in operation. The interesting result of such comparisons is that all practical systems tested so far performed far below Shannon's ideal system. Thus, one must conclude that the curve shown in Fig. 4 is an unattainable goal for any designer of a practical system. Yet a designer may measure how many decibels of utility he is below this ideal goal and may compare this figure with the same achievement of other designers.

A designer of a practical communications link will attempt to improve his design by adding better hardware (more circuits, better filters, larger sets of waveforms, etc.) until the increase of the cost for such improvements will outweigh the value of the improvements. Once this point is reached the designer can read off for his design the number of decibels below the ideal system.

On the other hand, the theoretical analysts are at work to derive more realistic models of an ideal system. Slepian, for example, derived in 1963 bounds for the utility of an ideal communication system that needs only a finite coding delay and not an infinite delay like Shannon's model. He showed that a restriction to a delay of 10^4 independent samples (dimensions) causes a reduction in utility of only less than 2 decibels.

When following the course of a practical designer using the utility chart, we may see him starting at the lower left corner of Fig. 4. This place will correspond to a rather inefficient design. He must now decide which way to go to improve his design. He has three choices. If the output power of his device is costly, as it is in spacecraft transmitters, he will give first priority to an increase in utility. His path will point vertically upward. If the cost of his design shows a strong increase he may attempt to improve his design by going to the right; i.e., by increasing the bit density until the additional cost increase will outweigh the improvement in transmission rate.

A different approach will be followed by a designer of a modem (modulator-demodulator equipment) for transmitting data over telephone lines. This designer's cost is primarily determined by the lease for the lines; i.e., he is paying for the bandwidth offered to him and not for the transmitter power. This designer will follow a path through the chart that goes first horizontally to the right. His main concern is an increase in bit density and not so much an increase in utility. Only after he finds that any further increase in bit density will become too
costly will the designer try to follow a vertical path, to increase the utility as much as possible without further significant cost increase.

It is apparent that any attempt to approach Shannon's upper bound very closely, either in utility or in bit density, will become increasingly costly. This points the way to the low cost design. This design proceeds diagonally through the chart along a line of constant signal-to-noise ratio. Interestingly we notice that this design leads automatically into the area of the conventional binary systems.

Thus we recognize that the utility chart allows the designer to follow one of the three approaches: the power limited design (vertically up), the bandwidth limited design (horizontally to the right), or the cost limited design (diagonally up to the right). Each of these approaches will run into increased cost, the closer it approximates Shannon's ideal system. Somewhere on this path the designer will find his optimum trade-off point.

We now recognize that the step from binary to nonbinary systems offers the designer more freedom of action. Indeed it has been proved\(^4\) that only nonbinary systems offer the possibility of approaching the upper bound of the channel capacity represented by Shannon's ideal system.

Fig. 5 should further stress the importance of the utility chart by demonstrating how one can convert the conventional error curves (plotted at the right side) into utility curves (plotted on the left side). The example is the ideal coded phase-coherent communications system of Viterbi\(^4,12\). The right side are curves expressing the utility \(U = \log_2 P_e\) as a function of the error ratio. The bit density is a constant parameter for each curve at the right side (\(n\) = constant). To plot a utility curve (with constant error ratio as curve parameter) one must read off all the values along a vertical line in the right chart and calculate the D value in decibel for each \("n"\) value marked at the right-hand curves. The \("n"\) values indicate the number of message bits in each orthogonal transmission word of Viterbi's system\(^12\). The number of binary elements in each transmission word is \(2^n\). All curves at the left side of Fig. 5 start from a small vertical part at \(n=2\) and go from there to the left. The lower end of this vertical part is the point for a binary system (\(n=1\)) operating with the error ratio indicated as curve parameter. It is easy to see that with decreasing error ratio the curves start at lower utility values, but also farther to the left. The reason for the shift to the left is that distortions have to be reduced when operating at a lower total error ratio. This can be achieved only when allotting a larger bandwidth for the transmission signal. Increasing the transmission bandwidth \(B_p\) when keeping \(R\) constant means a decreasing \(D\); thus we see that the curves must shift slightly to the left when being plotted for a smaller error ratio. It is evident that such details would not be visible in the conventional error probability curves at the right.

Having explained how the utility charts can be used for the analysis and the design of nonbinary systems, we turn now to special applications. Before concentrating on one particular subclass of nonbinary modulation systems, the subclass of multilevel systems, it may be desirable to give a classification of all nonbinary modulation methods.

### Nonbinary Modulation Methods

The nonbinary modulation methods may be subdivided into one-dimensional methods and multi-dimensional methods. The one-dimensional methods are also known as \(n\)-ary methods or multifrequency methods or multiphase methods, designating respectively the amplitude, the frequency, or the phase of a standard waveform as the parameter to be coded into one of \(n\) possible states. No other less frequently applied one-dimensional methods are the quantized pulse position method and the quantized pulse duration method.

These five multistate methods may best be discriminated by looking at Fig. 6. This figure depicts the sinusoidal wave of a carrier or subcarrier, plainly showing that there are 6 degrees of freedom (independent parameters) that specify completely such a piece of a sine wave (also called a rectangular AC pulse). In a one-dimensional modulation system only one of these parameters will be modulated and all the others will be kept at a constant value. The parameter \(b\), the dc component, rarely will be used; but must be zero in all radio systems with \(f_c\) as the radio carrier frequency. The parameters \(a\), \(f_c\), and \(\phi_c\) respectively correspond to amplitude, frequency, and phase. They are the most frequently used parameters. The modulation of the amplitude \(a\) will be discussed in details in other sections of this paper. The parameter \(t_1\) is the absolute time of the beginning of a pulse. It may be used in QPPM (quantized pulse position modulation systems)\(^6,14\). The parameter \(t_2\) is the end of the waveform; it determines, together with \(t_1\), the duration (or width) of the pulse as \(t_2 - t_1\). It is occasionally used as QFM (quantized pulse duration modulation)\(^6,15\). Both these systems depend on the accurate rectangular waveform of the pulse or some approximation to it. It can be easily shown that such methods are rather inefficient substitutes for the three basic \(n\)-ary modulation methods (MASK, MFSK, MPSK) standing for multiple amplitude (frequency or phase) shift keying. The term keying is borrowed from telegraphy to symbolize the discontinuous character of the modulation methods as compared with the continuous character of analog amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). Notice that the latter should not be confused with pulse modulation such as that used in the term PCM for pulse code modulation. FM is a continuous phase modulation (i.e., an analog process); whereas PCM is a code modulation (i.e., a digital process). One of the three basic one-dimensional multistate modulation methods will now be discussed in more detail.
Mathematical Models of Multilevel Systems

This may be the oldest n-ary modulation method that has been subjected to analytical investigation. The utility curves for such multilevel systems are all on the right side of the utility chart. This is understandable, as the provision of more than two amplitude levels does not change the width of the frequency spectrum of the waveform beyond that of a binary waveform of two levels. Neither does the multilevel character alter the length of the waveform. This means that the WT product (also called the signal base) of the waveform is the same for binary and for n-ary signals. This product will preferably be as small as possible, usually close to unity. The information transmission rate at the other side will increase to log₂n if n levels will be used. Accordingly, it is evident that the bit density will increase with n. On the other hand, it is necessary to increase the energy of the signals with increasing n if the error ratio is to stay constant. This accounts for the decreasing utility with increasing n. The results are curves as shown in Fig. 7. Curve 1 is again Shannon's ideal system following equation 5. Curve 2 is the ideal binary system. When applying it to a lowpass channel with six x over x waveforms, one may extend this ideal model to a bit density of +3 db corresponding to the so-called Nyquist rate. The curve 3 for the ideal multilevel system postulated in 1948 by Oliver, Pierce, and Shannon starts correctly at this point and extends generally parallel to curve 1, approaching that curve more closely for increasing n. This mathematical model of a completely idealized system for a lowpass channel is not very useful. Another mathematical model published in 1952 by Arthurs and Dym is shown by curve 4, in Fig. 7. This model tries to avoid the unrealistic assumptions of the model of curve 3. The Arthurs and Dym model assumes a bandpass channel and rectangular waveforms as shown in Fig. 6. This brings the system to a point for n=2 to a much lower bit density. The model is derived for a noncoherent amplitude demodulator and this fact is responsible for the much lower utility at corresponding n values when compared with the lowpass model of curve 3. Thus one may say that the Arthurs and Dym model is a kind of worst-case model. Yet it assumes distortion-free operation, an assumption that is difficult to realize with ideal rectangular pulse waveforms.

The Utility of Practical Designs of Multilevel Systems

Many efforts have been reported for the design of practical multilevel transmission systems. Some of the most successful designs will be mentioned here, but a more detailed review is given in part 2 of reference 5. Curve 5 and curve 6 in Fig. 7 represent measurements performed on a four-level data system that is based on an original design by McAuliffe.15 This particular kind of equipment has several practical improvements over the theoretical model of curve 4. One improvement is the use of an analog modulation method called vestigial sideband (VSB) system. This is one of the most efficient amplitude modulation methods because the transmission of the carrier and the low frequency sidebands permit the correct transmission of the bit rate and the carrier frequency without any decrease in the transmission rate. The result is that the bit density for VSB systems will be higher than for any other amplitude modulation method.

Another improvement applied in the VSB system is the special circuitry for carrier recovery that frees the designer from the constraint that the bit rate and the carrier frequency have to be synchronized. Such would be the case in single sideband systems (SSB) and in double sideband systems with suppressed carrier (DSB-SC).

A third improvement that makes the terminal rather independent of the specific channel characteristics is the use of a complementary VSB filter in the receiver. This filter has, in the vicinity of the carrier frequency, the opposite frequency response of the original VSB filter used in the transmitter modified for any variations of the frequency response of the channel.

The team that designed this modem performed functional tests with a four-level mode of this VSB data system. These tests were made in the laboratory over a channel of approximately 3.3 kż for the carrier channel, the characteristics of which are not specified in detail. Assuming an effective bandwidth of 33 kż for the carrier channel, one can plot the results in the utility chart in Fig. 7, curves 5 and 6. In particular, curve 5 corresponds to the utility of the system if the data carrier is 'patched through' between transmitter and receiver. Curve 6 corresponds to the utility of a design where the carrier is recovered from the noisy signal in the receiver. The interesting result is that the three improvements over older methods bring the practical measurements well in line with the theoretical model. Yet one may notice a number of differences. The theoretical curve 4 is plotted for eight levels at the bit density and utility where the practical system (curve 6) operates with four levels. The reason is that the theoretical model assumes rectangular wave-shape in DSB with envelope detector. The design of Critchlow et al. (curves 5 and 6) applies SSB and synchronous demodulation. It uses a wave-shape that is determined by special filters matched to the line characteristics. Actually, the Critchlow design, when assuming ideal hardware, should be represented by a single point higher up and more to the right. Indeed, when eliminating the jitter from the carrier recovered in the receiver by 'patching through' the carrier between transmitter and receiver (in a loop test), one arrives at curve 5. That curve, indeed, is higher up than curve 6. The fact that the Critchlow tests are represented by lines and not by points, despite the constant transmission rate, can be explained by the use of different filters and lines. Some of them have smaller bandwidth resulting in a higher bit density and smaller utility. The reduction in utility is caused by the increase in the energy per bit that is needed to maintain the constant error ratio for which all curves are plotted (one bit error.
in one thousand). This energy increase is required despite the assumption of constant noise power density (N0) because the smaller bandwidth causes higher intersymbol distortions and therefore would cause more errors without an increase in signal energy.

Coded Multilevel Systems

Intersymbol distortions are the principal cause of errors in systems with a large number of levels. Such distortions are the strongest for the largest difference in levels between successive symbols. Much effort therefore has been devoted to a coding operation that reduces the probability for large transitions. Some systems essentially establish "forbidden transitions." Thus we arrive at a scheme that follows Fig. 1. Multilevel systems of this kind may be called "coded multilevel systems."

The coder and the modulator of Fig. 1 will now have different circuits and different outputs from these in the coded phase-coherent system. This is shown in Fig. 8, where a special example of binary to quaternary encoding scheme is assumed, following the translation rules and the code shown in the figure. This leads to a code book of 16 code words, also shown in the figure. Assuming the same input message as in Fig. 1 (0110), one can see that this message is translated in the coder into the quartenary message -3, +1, +3, -1. The modulator may be just a low-pass filter in this case, shaping the rectangular signals to a more bandlimited waveform for transmission over a voice channel. It is evident that this particular code leads to transmission signals with a spectrum without dc component and without very low frequency components. Other codes may be applied to shape the spectrum at the high frequencies, still avoiding, if required, any dc component. To achieve this goal an encoder is used with a translation rule that excludes the maximum change directly from level 1 to level 3 or vice versa. Correspondingly an encoder for the polybinary system excludes any change of more than one level between adjacent m-ary digits.

Similarly, one can design a binary-to-multilevel encoder with equal input and output rate if one follows the rule that a binary zero at the input causes the level to stay constant, while a binary one at the input causes the output level to change for one step. This change continues in the same direction until a peak value is reached and then the direction of change is reversed. An equipment performing such a "variable-level coding" operation, thus generating stepped sawtooth-like waveforms, has been described by Shagena and Kvarda.

It is not easy to derive a utility chart for the two varieties of coded multilevel systems that we discussed above. The difficulty is due to the fact that the waveform distortions rather than the noise are causing the errors. Gerrish overcame this problem by comparing the coded multilevel system with an uncoded binary system of the same transmission rate, the same bandwidth, and the same error ratio. Curve 8 shows the utility curve of this binary reference system as we derived it from the eye pattern measurements by Gerrish. It is interesting to note that the bit density may be increased up to +2 db without any decrease in utility. In the first part the curve is parallel to curve 3, the curve of the ideal binary system. The constant loss of about 3 db in utility is due to the code and distortions in the non-ideal circuits. Beyond +2 db bit density (3 db corresponds to the Nyquist rate) the reference system loses utility fast. This is due to the severe distortions when operating at the Nyquist rate or faster. Notice that Gerrish operates with a lowpass channel that simulates the attenuation and phase characteristics of a telephone line.

This reference curve 8 can now be compared with the utility curve for a duobinary system (curve 9), a standard quaternary system without coding (curve 10), and the variable multicode system (curve 11). The results indicate that all the multilevel systems show a maximum utility at about the Nyquist rate and they are better than the binary system above the Nyquist rate. When significantly below the Nyquist rate, the binary system has the higher utility. Among the multilevel systems, the variable multicode system offers the least advantage over the binary system. The duobinary system shows a significant
improvement at bit densities above +2db. The uncoded four-level system shows a well-marked maximum utility at the Nyquist rate, but falls off fast at higher rates.

When comparing Gerrish's results for an uncoded four-level system (curve 10) with the results of Critchlow (curve 5), one should remember that Gerrish made his measurements under simulated practical conditions while the Critchlow tests were loop tests over actual telephone lines. The principal difference is that Gerrish investigated a lowpass system with "coherent amplitude modulation" (ideal resampling instants) while Critchlow investigated vestigial sideband modulation with resampling instants derived from the noisy signals, even in the case where the carrier was 'patched through'. From this comparison one must conclude that the coded multilevel systems of curves 9 and 11 would be 4 to 6 db below the present values when operated over actual links. This number would account for demodulator and synchronization losses.

High Order Multilevel Systems

All systems discussed so far operated with a small number of levels. Publications are available about two developments of multilevel systems with 16 and 32 levels respectively. F. K. Becker24 describes a multilevel VSB data terminal for use over high grade voice facilities (specially selected telephone lines) that is the result of many earlier research and development projects. This terminal applies an automatic line equalization system described independently by Lucky25. The coding equipment in this system uses an error correcting technique described by Burton and Weldon26. The combined effect of the application of VSB, equalization, and coding results in an integrated data system with a utility curve as shown in curve 12, figure 7. The end points of this line show the operation of the equipment with 8 or 16 levels at transmissions rates of 7200 and 9600 bits per second respectively. Only about 70 percent of the rates are used for information digits; the remainder is used for error control digits and for synchronization and formatting. In the calculation of the bit density, an effective bandwidth of 2000 Hz was assumed24 and only the rate of information digits was applied as transmission rate. It is interesting to note that curve 12 of the practical system is closely parallel to curve 4, a mathematical model. Yet the practical system would require a quite different mathematical model that has not yet been derived but that might have a curve more nearly like curve 2 because of the VSB modulation and the coherent detection. The VSB system of curve 12 applies two pilot carriers, at 600 Hz and at 3000 Hz. Their power requirement has been included when calculating the utility values.

The second coded multilevel system to be mentioned in this connection has 32 levels and is an experimental system described by staff members of the Lincoln Laboratory of the Massachusetts Institute of Technology27. The utility curve is shown as curve 13 in Fig. 7. This system, like the system described by Becker, is an integrated system with coder and modulator operating in highly sophisticated modes. The encoder is a self-regulating error correcting coder-encoder (SECO)28 which operated during the tests with one out of three different redundancies: 3/5, 2/5, or 1/5, meaning that 3, 2, or 1 out of 5 transmitted bits are information bits and the rest are check bits. This coding scheme operates on a sequential basis with a coding delay of 60 bits. The output of the encoder is converted to transmission signals having one out of 32 levels. A special "signals synthesis network" matches the lowpass signals of the multilevel coder to the line phase characteristics. A VSB modulator then matches the signal spectrum to the bandpass characteristic of the line, using a carrier of 2500 Hz. A pilot of 3000 Hz, together with the 2500 Hz carrier, enables synchronous tracking in the receiver. Additional adaptive procedures permit the system to match its mode of operation to the changing condition in the transmission plant.

For complete utilization of all these capabilities, the SECO system needs a return channel. In the tests that formed the basis of curve 13 the SECO system operated over an 800-mile loop of a type K carrier system and achieved a throughput of 6000 to 9000 bits per second. Notice, however, that in this case the return channel was 'patched through' between the two terminals. This has to be taken into account when comparing the utility of the SECO system with the smaller utility of the system described by Becker. Curve 13 is an extended line, although it is plotted for a constant number of levels. This is due to the adaptive character of the system. When applying less energy per bit (i.e., when increasing the utility), the system will automatically select a higher redundancy; i.e., it will decrease the bit density. The same effect occurs when the noise power density increases and the energy per bit remains constant.

It is interesting to note that the SECO operating line (curve 13) comes very close to the operating point of an ideal lowpass system, suggested as early as 1952 in a paper by Gilbert29, known as a quantized pulse amplitude system. Point 14 represents Gilbert's results for an ideal bipolar system, operating with 5 levels and with sin x over x waveforms.

Conclusions

Summarizing the interpretation of the utility curves of multilevel systems, one can recognize that all curves are on the right side of the utility chart. Multilevel systems definitely have the purpose of operating in the range of high bit densities. Their applications will be in cases where the cost for bandwidth is high and the cost for power is low. Data transmission over telephone lines is a typical example. Fig. 8 indicates that a number of sophisticated systems are able to come close to the utility of the theoretical models. Yet one should notice that the models have been derived for extremely simple cases. More complex models are required that could better represent the complex practical systems. One would then recognize how much more margin for improvement is still available. This becomes particularly evident when comparing the only
system at the chart that had been tested under completely normal operating conditions, the system described by Becker and shown in curve 12, with the best possible system, curve 1, representing Shannon’s upper bound of the channel capacity. The practical system is more than 20 db below the ideal system. Yet this practical system is already highly sophisticated, can operate only over carefully selected private lines, and still makes one error in one thousand bits. The ideal system of Shannon is error-free. Future designers have the field wide open to produce data transmission systems with more than 1 percent efficiency and with less error ratio.

References


Hamming Encoding

Bi-orthogonal Words

Input Word

0110

Service Digit

0110

One out of 16

One out of 16

Figure 1. Coded Communications Systems
Message | Encoding |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>First Bit</td>
<td>0 1 0</td>
</tr>
<tr>
<td>Second Bit</td>
<td>1 0 1</td>
</tr>
<tr>
<td>Fourth Bit</td>
<td>0 1 1</td>
</tr>
</tbody>
</table>

Transmitted Error Received

First check | 1 0 0 0 1 1 0 |
Second check | 0 0 1 1 0 |
Third check | 0 1 1 0 |

Position of error (binary number) | X |
Corrected Message | 1 1 0 0 1 1 0 |
Output Word | 0 1 1 0 |

Figure 2. Hamming Error Correction Code

Bi-orthogonal Transmission Words

<table>
<thead>
<tr>
<th>Message</th>
<th>0000</th>
<th>0001</th>
<th>0010</th>
<th>0011</th>
<th>0100</th>
<th>0101</th>
<th>0110</th>
<th>0111</th>
<th>1000</th>
<th>1001</th>
<th>1010</th>
<th>1011</th>
<th>1100</th>
<th>1101</th>
<th>1110</th>
<th>1111</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Figure 3. Bi-orthogonal Codes
Utility = \frac{\text{Bit density}}{\text{SNR}} = \frac{\text{Noise power density}}{\text{Energy per bit}} = \frac{N_0}{E_B}; \text{Bit density} = \frac{\text{Transmission rate}}{\text{Bandwidth}}

Figure 4. Fundamentals of the Utility Chart
Figure 5. Generating Utility Curves (Left) from the Error Probability Curves (Right)
Figure 6. The Six Parameters of a Typical Rectangular Carrier Signal
Figure 7. The Utility of Multilevel Systems (All curves except curve 1 are plotted for an error ration of $10^{-3}$.)
Digital Source  \( \rightarrow \) Coder  \( \rightarrow \) Filter

**Translation Rules**

\[
\begin{align*}
\beta_1 &= \alpha_1 (\alpha_3 + 2) \\
\beta_2 &= \alpha_2 (\alpha_4 + 2) \\
\beta_3 &= -\beta_1 \\
\beta_4 &= -\beta_2
\end{align*}
\]

**Code**

\[
\begin{align*}
11 &\rightarrow +1+1 \rightarrow +3 \\
10 &\rightarrow +1-1 \rightarrow +1 \\
00 &\rightarrow -1+1 \rightarrow -1 \\
01 &\rightarrow -1-1 \rightarrow -3
\end{align*}
\]

**Code Book**

<table>
<thead>
<tr>
<th>Code</th>
<th>Quaternary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 1 1</td>
<td>#1 $+3 +3 -3 -3$</td>
</tr>
<tr>
<td>1 1 1 0</td>
<td>#2 $+3 +1 -3 -1$</td>
</tr>
<tr>
<td>1 1 0 1</td>
<td>#3 $+3 -3 -3 +3$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0 1 1 0</td>
<td>#9 $-3 +1 +3 -1$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0 0 0 1</td>
<td>#15 $-1 -3 +1 +3$</td>
</tr>
<tr>
<td>0 0 0 0</td>
<td>#16 $-1 -1 +1 +1$</td>
</tr>
</tbody>
</table>

**Figure 8. Four-Level Code**