The Philosophy of PCM*

B. M. OLIVER†, MEMBER, IRE, J. R. PIERCE†, FELLOW, IRE, AND C. E. SHANNON†

***Summary*—Recent papers[1,6] describe experiments in transmitting speech by PCM (pulse code modulation). This paper shows in a general way some of the advantages of PCM, and distinguishes between what can be achieved with PCM and with other broadband systems, such as large-index FM. The intent is to explain the various points simply, rather than to elaborate them in detail. The paper is for those who want to find out about PCM rather than for those who want to design a system. Many important factors will arise in the design of a system which are not considered in this paper.**

I. PCM and Its Features

THERE ARE SEVERAL important elements of a PCM (pulse-code modulation) system. These will be introduced, and the part each plays in PCM will be explained in this section.

Sampling

In general, the object of a transmission system is to reproduce at the output any function of time which appears at the input. In any practical system only a certain class of functions, namely, those limited to a finite frequency band, are admissible inputs. A signal which contains no frequencies greater than W_0 cps cannot assume an infinite number of independent values per second. It can, in fact, assume exactly $2W_0$ *independent* values per second, and the amplitudes at any set of points in time spaced τ_0 seconds apart, where $\tau_0 = 1/2W_0$, specify the signal completely. A simple proof of this is given in Appendix I. Hence, to transmit a band-limited signal of duration T, we do not need to send the entire continuous function of time. It suffices to send the finite set of $2W_0T$ independent values obtained by sampling the instantaneous amplitude of the signal at a regular rate of $2W_0$ samples per second.

If it surprises the reader to find that $2W_0T$ pieces of data will describe a continuous function completely over the interval T, it should be remembered that the $2W_0T$ coefficients of the sine and cosine terms of a Fourier series do just this, if, as we have assumed, the function contains no frequencies higher than W_0.

Reconstruction

Let us now proceed to the receiving end of the system, and assume that, by some means, the sample values representing the signal are there and available in proper time sequence, and can be used at the regular rate $2W_0$. To reconstruct the signal it is merely necessary to generate from each sample a proportional impulse, and to pass this regularly spaced series of impulses through an ideal low-pass filter of cutoff frequency W_0. The output of this filter will then be (except for an over-all time delay and possibly a constant of proportionality) identical to the input signal. Since the response of an ideal low-pass filter to an impulse is a $\sin x/x$ pulse, and since the total output is the linear sum of the responses to all inputs, this method of reconstruction is simply the physical embodiment of the method indicated in Appendix I.

Ideally, then, we could achieve perfect reproduction of a signal if we could transmit information giving us exactly the instantaneous amplitude of the signal at intervals spaced $1/2W_0$ apart in time.

Quantization

It is, of course, impossible to transmit the *exact* amplitude of a sample. The amplitude of a sample is often transmitted as the amplitude of a pulse, or as the time position of a pulse. Noise, distortion, and crosstalk between pulses will disturb the amplitude and position, and hence cause errors in the recovered information concerning the size of the sample. Ordinarily the error becomes greater as the signal is amplified by successive repeaters, and hence the accumulation of noise sets a limit to the distance a signal can be transmitted even with enough amplification.

It is possible, however, to allow only certain discrete levels of amplitude or position of the transmitted pulse. Then, when the signal is sampled, the level nearest the true signal level is sent. When this is received and amplified, it will have a level a little different from any of the specified levels. If the noise and distortion are not too great, we can surely tell which level the signal was supposed to have. Then the signal can be reformed, or a new signal created, which again has the level originally sent.

Representing the signal by certain discrete allowed levels only is called *quantizing*. It inherently introduces an initial error in the amplitude of the samples, giving rise to *quantization noise*. But once the signal is in a quantized state, it can be relayed for any distance without further loss in quality, provided only that the added noise in the signal received at each repeater is not too great to prevent correct recognition of the particular level each given signal is intended to represent. By quantizing we limit our "alphabet." If the received signal lies between a and b, and is closer (say) to b, we guess that b was sent. If the noise is small enough, we shall always be right.

* Decimal classification: R148.6. Original manuscript received by the Institute, May 24, 1948.

† Bell Telephone Laboratories, Inc., New York, N. Y.

[1] W. M. Goodall, "Telephony by pulse code modulation," *Bell Sys. Tech. Jour.*, vol. 26, pp. 395–409; July, 1947.

[2] D. D. Grieg, "Pulse count modulation system," *Tele-Tech.*, vol. 6, pp. 48–52; September, 1947.

[3] D. D. Grieg, "Pulse count modulation," *Elec. Commun.*, vol. 24, pp. 287–296; September, 1947.

[4] H. S. Black and J. O. Edson, "PCM equipment," *Elec. Eng.*, vol. 66, pp. 1123–25; November, 1947.

[5] A. C. Clavier, D. D. Grieg, and P. F. Panter, "PCM distortion analysis," *Elec. Eng.*, vol. 66, pp. 1110–1122; November, 1947.

[6] L. A. Meacham and E. Peterson, "An experimental multichannel pulse code modulation system of toll quality," *Bell Sys. Tech. Jour.*, vol. 27, pp. 1–43; January, 1948.

Coding

A quantized sample could be sent as a single pulse which would have certain possible discrete amplitudes, or certain discrete positions with respect to a reference position. However, if many allowed sample amplitudes are required, one hundred, for example, it would be difficult to make circuits to distinguish these one from another. On the other hand, it is very easy to make a circuit which will tell whether or not a pulse is present. Suppose, then, that several pulses are used as a *code group* to describe the amplitude of a single sample. Each pulse can be on (1) or off (0). If we have three pulses, for instance, we can have the combinations representing the amplitudes shown in Table I.

TABLE I

Amplitude Represented	Code
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

The codes are, in fact, just the numbers (amplitudes) at the left written in binary notation. In this notation, the place-values are 1, 2, 4, 8,—; i.e., a unit in the right-hand column represents 1, a unit in the middle (second) column represents 2, a unit in the left (third) column represents 4, etc. We see that with a code group of n on-off pulses we can represent 2^n amplitudes. For example, 7 pulses yield 128 sample levels.

It is possible, of course, to code the amplitude in terms of a number of pulses which have allowed amplitudes of 0, 1, 2 (base 3 or ternary code), or 0, 1, 2, 3 (base 4 or quaternary code), etc., instead of the pulses with allowed amplitudes 0, 1 (base 2 or binary code). If ten levels were allowed for each pulse, then each pulse in a code group would be simply a digit of an ordinary decimal number expressing the amplitude of the sample. If n is the number of pulses and b is the base, the number of quantizing levels the code can express is b^n.

Decoding

To decode a code group of the type just described, one must generate a pulse which is the linear sum of all the pulses in the group, each multiplied by its place value ($1, b, b^2, b^3, \ldots$) in the code. This can be done in a number of ways. Perhaps the simplest way which has been used involves sending the code group in "reverse" order, i.e., the "units" pulse first, and the pulse with the highest place value last. The pulses are then stored as charge on a capacitor-resistor combination with a time constant such that the charge decreases by the factor $1/b$ between pulses. After the last pulse, the charge (voltage) is sampled.

A Complete PCM System

A PCM system embodies all the processes just described. The input signal is band-limited to exclude any frequencies greater than W_0. This signal is then sampled at the rate $2W_0$. The samples are then quantized and encoded. Since only certain discrete code groups are possible, the selection of the nearest code group automatically quantizes the sample, and with certain types of devices it is therefore not necessary to quantize as a separate, prior operation. The code groups are then transmitted, either as a time sequence of pulses (time division) over the same channel, or by frequency division, or over separate channels. The code groups are regenerated (i.e., reshaped) at intervals as required. At the receiver the (regenerated) code groups are decoded to form a series of impulses proportional to the original samples (except quantized), and these impulses are sent through a low-pass filter of bandwidth W_0 to recover the signal wave.

II. Transmission Requirements for PCM

Suppose we consider what requirements exist, ideally, on the channel which is to carry the encoded PCM signal; that is, ruling out physically impossible devices, but allowing ideal components such as ideal filters, ideal gates, etc.

Bandwidth

If a channel has a bandwidth W cps, it is possible to send up to $2W$ independent pulses per second over it. We can show this very simply. Let the pulses occur (or not occur) at the time $t=0, \tau, 2\tau, \cdots, m\tau$ where $\tau = 1/2W$, and let each pulse as received be of the form

$$V = V_0 \frac{\sin \frac{\pi}{\tau}(t - m\tau)}{\frac{\pi}{\tau}(t - m\tau)} \cdot \qquad (1)$$

The shape of this pulse is shown in Fig. 1. It will be seen that the pulse centered at time $m\tau$ will be zero at $t=k\tau$

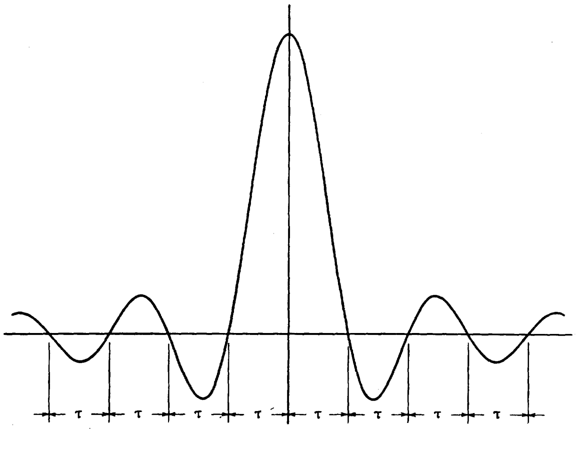

Fig. 1—Pulse of the form $V_0 \frac{\sin(\pi t/\tau)}{\pi t/\tau}$.

where $k \neq m$. Thus, if we sample the pulse train at the time $t = m\tau$, we will see only the pulse belonging to that time and none of the others.

Further, the pulse given by (1) contains no frequencies higher than W. It is the pulse one would get out of an ideal low-pass filter of cutoff W, on applying a very short impulse to the input.

Now, to send a signal of bandwidth W_0 by PCM, we must send $2W_0$ code groups per second and each code group contains (say) n pulse places. We must be prepared, therefore, to send $2nW_0$ pulses per second, and this requires a bandwidth $W = nW_0$. The pulses may be sent in time sequence over one channel or by frequency division. In either case the total bandwidth will be the same. Of course, if double-sideband transmission is used in the frequency-division case, or if the time-division signal is sent as double-sideband rf pulses, the total bandwidth will be $2nW_0$.

In short, the bandwidth required for PCM is, in the ideal case, n times as great as that required for direct transmission of the signal, where n is the number of pulses per code group.

Threshold Power

To detect the presence or absence of a pulse reliably requires a certain signal-to-noise ratio. If the pulse power is too low compared to the noise, even the best possible detector will make mistakes and indicate an occasional pulse when there is none, or vice versa. Let us assume that we have an ideal detector, i.e., one which makes the fewest possible mistakes. If the received pulses are of the form (1), and if the noise is "white" noise (i.e., noise with a uniform power spectrum and gaussian amplitude distribution as, for example, thermal noise), ideal detection could be achieved by passing the signal through an ideal low-pass filter of bandwidth W ($= nW_0$ in the ideal case) and sampling the output at the pulse times $k\tau$. If the signal when sampled exceeds $V_0/2$, we say a pulse is present; if less than $V_0/2$, we say there is no pulse. The result will be in error if the noise at that instant exceeds $V_0/2$ in the right direction. With gaussian noise, the probability of this happening is proportional to the complementary error function[7] of

$$\frac{V_0}{2\sigma} = \sqrt{\frac{P_s}{4N}}$$

where

σ = rms noise amplitude
P_s = signal (pulse) "power" $= V_0^2$
N = noise power in bandwidth $W = \sigma^2$.

As the signal power P_s is increased, this function decreases very rapidly, so that if P_s/N is large enough to make the signal intelligible at all, only a small increase will make the transmission nearly perfect. An idea of how rapidly this improvement occurs may be had from Table II. The last column in the table assumes a pulse rate of 10^5 per second.

TABLE II

Signal to Noise $\frac{P_s}{N}$	Probability of Error	This Is About One Error Every
13.3 db	10^{-2}	10^{-3} sec
17.4 db	10^{-4}	10^{-1} sec
19.6 db	10^{-6}	10 sec
21.0 db	10^{-8}	20 min
22.0 db	10^{-10}	1 day
23.0 db	10^{-12}	3 months

Clearly, there is a fairly definite *threshold* (at about 20 db, say) below which the interference is serious, and above which the interference is negligible. Comparing this figure of 20 db with the 60- to 70-odd db required for high-quality straight AM transmission of speech, it will be seen that PCM requires much less signal power, even though the noise power is increased by the n-fold increase in bandwidth.

The above discussion has assumed an on-off (base 2) system. In this system pulses will be present half the time, on the average, and the *average* signal power[8] will be $P_s/2$. If a balanced base 2 system were used, i.e., one in which 1 is sent as a + pulse (say) and 0 as a − pulse, the peak-to-peak signal swing would have to be the same as in the on-off system for the same noise margin, and this swing would be provided by pulses of only half the former amplitude. Since either a + or − pulse would always be present, the signal power would be $P_s/4$.

If pulses are used which have b different amplitude levels (i.e., a base b system), then a certain amplitude separation must exist between the adjacent levels to provide adequate noise margin. Call this separation $K\sigma$, where K = a constant. (From the preceding discussion we see that K is about 10.) The total amplitude range is therefore $K\sigma(b-1)$. The signal power will be least if this amplitude range is symmetrical about zero, i.e., from $-K\sigma(b-1)/2$ to $+K\sigma(b-1)/2$. The average signal power S, assuming all levels to be equally likely, is then[8]

$$S = K^2\sigma^2 \frac{b^2-1}{12}$$

$$= K^2 N \frac{b^2-1}{12}. \qquad (2)$$

It will be noticed that the required signal power increases rapidly with the base b.

Regeneration: The Pay-Off

In most transmission systems, the noise and distortion from the individual links cumulate. For a given

[7] Complementary error function of $x = 1/\sqrt{2\pi}\int_x^\infty e^{-\lambda^2/2}d\lambda$.

[8] See Appendix II.

quality of over-all transmission, the longer the system, the more severe are the requirements on each link. For example, if 100 links are to be used in tandem, the noise power added per link can only be one-hundredth as great as would be permissible in a single link.

Because the signal in a PCM system can be regenerated as often as necessary, the effects of amplitude and phase and nonlinear distortions in one link, if not too great, produce no effect whatever on the regenerated input signal to the next link. If noise in a single link causes a certain fraction p of the pulses to be regenerated incorrectly, then after m links, if $p \ll 1$, the fraction incorrect will be approximately mp. However, to reduce p to a value $p' = p/m$ requires only a slight increase in the power in each link, as we have seen in the section on threshold power. Practically, then, the transmission requirements for a PCM link are almost independent of the total length of the system. The importance of this fact can hardly be overstated.

III. Performance of a PCM System

We have seen that PCM requires more bandwidth and less power than is required with direct transmission of the signal itself, or with straight AM. We have, in a sense, exchanged bandwidth for power. Has the exchange been an efficient one? Are good signal-to-noise ratios in the recovered signal feasible in PCM? And how sensitive to interference is PCM? We shall now try to answer these questions.

Channel Capacity

A good measure of the bandwidth efficiency is the information capacity of the system as compared with the theoretical limit for a channel of the same bandwidth and power. The information capacity of a system may be thought of as the number of independent symbols or characters which can be transmitted without error in unit time. The simplest, most elementary character is a binary digit, and it is convenient to express the information capacity as the equivalent number of binary digits per second, C, which the channel can handle. Shannon and others have shown that an ideal system has the capacity[9]

$$C = W \log_2 \left(1 + \frac{P}{N}\right) \tag{3}$$

where

W = bandwidth
P = average signal power
N = white noise power.

Two channels having the same C have the same capacity for transmitting information, even though the quantities W, P, and N may be different.

In a PCM system, operating over the threshold so that the frequency of errors is negligible,

$$C = sm$$

where

s = sampling rate = $2W_0$
m = equivalent number of binary digits per code group.

If there are l quantizing levels, the number of binary digits required per code group is given by $l = 2^m$, while the actual number of (base b) digits n will be given by

$$l = b^n.$$

Thus,

$$2^m = b^n$$
$$m = n \log_2 b$$

and

$$C = sn \log_2 b.$$

Now sn is the actual pulse frequency, and is ideally twice the system bandwidth W.
Therefore,

$$C = 2W \log_2 b$$
$$= W \log_2 b^2.$$

Substituting for b the power required for this base (from (2)), we have

$$C = W \log_2 \left(1 + \frac{12S}{K^2 N}\right). \tag{4}$$

Comparing (4) with (3), we see they are identical if $S = (K^2/12)P$. In other words, PCM requires $K^2/12$ (or about 8) times the power theoretically required to realize a given channel capacity for a given bandwidth.

Perhaps the most important thing to notice about (4) is that the form is right. Power and bandwidth are exchanged on a logarithmic basis, and the channel capacity is proportional[10] to W. In most broadband systems, which improve signal-to-noise ratio at the expense of bandwidth, C is proportional only to log W.

Signal-to-Noise Ratio

There are two types of noise introduced by a PCM system. One of these is the quantizing noise mentioned in the section on quantization. This is a noise introduced at the transmitting end of the system and nowhere else. The other is *false pulse* noise caused by the incorrect interpretation of the intended amplitude of a pulse by the receiver or by any repeater. This noise may arise anywhere along the system, and is cumulative. However, as we have seen earlier, this noise decreases so rapidly as the signal power is increased above threshold that in any practical system it would be made negligible by design. As a result, the signal-to-noise ratio in PCM systems is set by the quantizing noise alone.

[9] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Jour.*, vol. 27, July, October, 1948.

[10] Provided S is increased in proportion to W to compensate for the similar increase in N.

If the signal is large compared with a single quantizing step, the errors introduced in successive samples by quantizing will be substantially uncorrelated. The maximum error which can be introduced is one-half of one quantizing step in either direction. All values of error up to this maximum value are equally likely. The rms error introduced is, therefore, $1/2\sqrt{3}$ times the height of a single quantizing step.[8] When the signal is reconstructed from the decoded samples (containing this quantizing error), what is obtained is the original signal plus a noise having a uniform frequency spectrum out to W_0 and an rms amplitude of $1/2\sqrt{3}$ times a quantizing step height. The ratio of peak-to-peak signal to rms noise is, therefore,

$$R = 2\sqrt{3}\, b^n,$$

since b^n is the number of levels. Expressing this ratio in db, we have

$$\begin{aligned} 20 \log_{10} R &= 20 \log_{10} 2\sqrt{3} + n(20 \log_{10} b) \\ &= 10.8 + n(20 \log_{10} b). \end{aligned} \tag{5}$$

In a binary system, $b=2$, and

$$20 \log_{10} R \cong 10.8 + 6n.$$

In examining (5) let us remember that n, the number of digits, is a factor relating the *total bandwidth used in transmission* to the *bandwidth of the signal to be transmitted*, i.e., $W=nW_0$. It is something like the index of modulation in FM. Now, for every increment of W_0 added to the bandwidth used for transmission, n may be increased by one, and this increases the signal-to-noise ratio by a constant number of db. In other words, in PCM, *the signal-to-noise ratio in db varies linearly with* the number of digits per code group, and hence with the *bandwidth.* Of course, as the bandwidth is increased the noise power increases, and a proportional increase in signal power is required to stay adequately above threshold.

A binary PCM system using ten times the bandwidth of the original signal will give a 70-db signal-to-noise ratio. Higher base systems will require less bandwidth.

Ruggedness

One important characteristic of a transmission system is its susceptibility to interference. We have seen that noise in a PCM circuit produces no effect unless the peak amplitude is greater than half the separation between pulse levels. In a binary (on-off) system, this is half the pulse height. Similarly, interference such as stray impulses, or pulse crosstalk from a near-by channel, will produce no effect unless the peak amplitude of this interference plus the peak noise is half the pulse height. The presence of interference thus increases the threshold required for satisfactory operation. But, if an adequate margin over threshold is provided, comparatively large amounts of interference can be present without affecting the performance of the circuit at all. A PCM system, particularly an on-off (binary) system, is therefore quite "rugged."

When a number of radio communication routes must converge on a single terminal, or follow similar routes between cities, the ruggedness of the channels is a particularly important consideration. If the susceptibility of the channels to mutual interference is high, many separate frequency bands will be required, and the total bandwidth required for the service will be large. Although PCM requires an initial increase of bandwidth for each channel, the resulting ruggedness permits many routes originating from, or converging toward, a single terminal to occupy the same frequency band. Different planes of polarization for two channels over the same path can often be used, and the directivities of practical antennas are such that only a small difference in direction of arrival will separate two routes on the same frequency. As a result, the frequency occupancy of PCM is exceptionally good, and its other transmission advantages are then obtained with little, if any, increase in *total* bandwidth.

IV. Comparison of PCM and FM

One feature of PCM is that the signal-to-noise ratio can be substantially improved by increasing the transmission bandwidth. This is an advantage shared with certain other pulse systems and with FM. As FM is the best known of these other systems, it is interesting to compare PCM and FM.

Broadband Gain

In going to high-deviation FM, the gain in signal-to-noise voltage ratio over AM (with the same power and the same noise per unit bandwidth) is proportional to the deviation ratio, or to the ratio of half the bandwidth actually used in transmission to the bandwidth of the signal to be transmitted. This ratio corresponds to n in our notation. If noise power is uniformly distributed with respect to frequency, and if one desires to provide the same margin over threshold in FM with various bandwidths, the transmitter power must be proportional to bandwidth (to n). If we so vary the power in varying the bandwidth of wide-deviation FM, the signal-to-noise voltage ratio will vary as $n(n^{1/2})$, where the factor $n^{1/2}$ comes about through the increased signal voltage. Thus the signal-to-noise ratio R will be given by

$$\begin{aligned} R &= (\text{const})n^{3/2} \\ 20 \log_{10} R &= 30 \log_{10} n + \text{const}. \end{aligned} \tag{6}$$

For binary (on-off) PCM we have, from (5), for the same simultaneous variation of bandwidth and power

$$20 \log_{10} R = 6n + 10.8.$$

Or, for ternary (base 3) PCM,

$$20 \log_{10} R = 9.54n + 10.8.$$

We see that, as the bandwidth (proportional to n) is increased in FM, the signal-to-noise ratio varies as log n, while in PCM it varies as n. Thus, as bandwidth is increased, PCM is bound to exhibit more improvement in the end. Further, a more elaborate analysis shows that, ideally at least, PCM can provide, for any bandwidth, nearly as high a signal-to-noise ratio as is possible with any system of modulation.

Why is PCM so good in utilizing bandwidth to increase the signal-to-noise ratio? A very clear picture of the reason can be had by considering a simple PCM system in which four binary digits are transmitted on four adjacent frequency bands with powers just sufficient to over-ride noise. In Fig. 2(a) the signals in these four channels B_1, B_2, B_3, B_4 are shown versus time. A black rectangle represents a pulse; a white rectangle, the absence of a pulse. The rectangles are $\tau_0=(1/2W_0)$ long. The particular sequence of code groups shown in the figure represents a quantized approximation to a linear change of amplitude with time, as shown in Fig. 2(b).

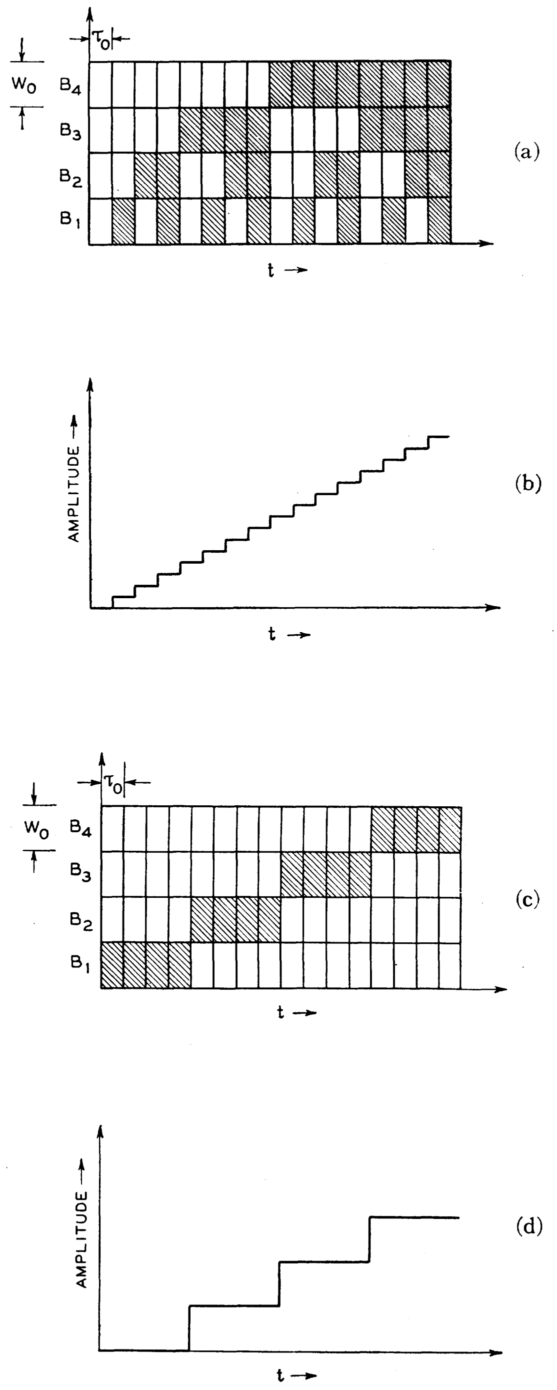

Fig. 2—The signals in channels B_1, B_2, B_3, and B_4. (a) Signal in a frequency-division PCM system. (b) Amplitudes corresponding to (a). (c) Signal in a quantized FM system. (d) Amplitudes corresponding to (c).

Now suppose, instead, that we confine ourselves to sending a pulse in only one channel at a time, as shown in Fig. 2(c). The best quantized representation of the signal we can get is shown in Fig. 2(d). Here the number of levels is four, while in Fig. 2(b) there are sixteen. In other words, Fig. 2(b) represents four times as good a signal-to-noise amplitude ratio as Fig. 2(d).

The total energy transmitted is in each case represented by the total black area; we see that on the average twice as much power is used in Fig. 2(a) as in Fig. 2(c). Thus we obtain a 12-db increase in signal-to-noise ratio with a power increase of only 3 db by sending the signal according to Fig. 2(a) rather than Fig. 2(c). If we had started out with six channels instead of four, we would have obtained a signal-to-noise improvement of 21 db for 4.77 db more average power. The greater the number of channels, and hence the wider the frequency band used, the better the method of transmission represented by Fig. 2(a) as compared to that represented by Fig. 2(c).

Now Fig. 2(a) represents PCM, while Fig. 2(c) represents what is essentially quantized FM with sampling. The signal in Fig. 2(c) varies with frequency according to the amplitude of the signal. Hence, we have compared PCM and a sort of FM, to the obvious advantage of PCM.

The trouble with the FM type of signal of Fig. 2(c) is that only a few of the possible signals which might be sent in the four bands B_1–B_4 are ever produced; all the others, those for which there is a signal in more than one band at a time, are wasted. Ideally, PCM takes advantage of every possible signal which can be transmitted over a given band of frequencies with pulses having discrete amplitudes.[11]

The relation between FM and PCM is closely analogous to the relation between the two types of computing machines: the so-called analogue machines and the digital machines. In analogue machines the numbers involved are represented as proportional to some physical quantity capable of continuous variation. Typical examples are the slide rule, network analyzers, and the differential analyzer. An increase in precision requires,

[11] It might be objected that one could have signals with a finer structure in the frequency direction than those shown in Fig. 2(a). This is possible only if τ is made larger, so that the pulses representing samples occur less frequently, are broader, and have narrower spectra. This means reducing W_0.

in general, a proportional increase in the range of physical variables used to represent the numbers. Furthermore, small errors tend to accumulate and cannot be eliminated. In digital machines the numbers are expressed in digital form, and the digits are represented by the states of certain physical parts of the machine which can assume one of a finite set of possible states. Typical digital machines are the abacus, ordinary desk computers, and the Eniac. In this type of machine the precision increases exponentially with the number of digits, and hence with the size of the machine. Small errors, which are not large enough to carry any part from one state to another state, have no effect and do not cumulate.

In FM (analogue), the amplitude of the audio signal is measured by the radio frequency. To improve the precision by 2 to 1 requires roughly a 2 to 1 increase in the frequency swing, and hence the bandwidth. In PCM doubling the bandwidth permits twice the number of digits, and therefore *squares* rather than doubles the number of distinguishable levels.

Other Factors

There are other considerations in a comparison between PCM and ordinary, unquantized FM, however. For instance, PCM allows the use of regenerative repeaters, and FM does not. PCM lends itself, like other pulse systems, to time-division multiplex. On the other hand, when the received signal rises considerably above threshold during good reception, the signal-to-noise ratio improves with FM but not with PCM. When we come to consider transmitters and receivers, we find that, for high signal-to-noise ratios at least, an FM transmitter and receiver will be somewhat less complicated than those for PCM are at present.

V. Conclusions

PCM offers a greater improvement in signal-to-noise than other systems, such as FM, which also depend upon the use of wide bands.

By using binary (on-off) PCM, a high-quality signal can be obtained under conditions of noise and interference so bad that it is just possible to recognize the presence of each pulse. Further, by using regenerative repeaters which detect the presence or absence of pulses and then emit reshaped, respaced pulses, the initial signal-to-noise ratio can be maintained through a long chain of repeaters.

PCM lends itself to time-division multiplex.

PCM offers no improvement in signal-to-noise ratio during periods of high signal or low noise.

PCM transmitters and receivers are somewhat more complex than are those used for some other forms of modulation.

In all, PCM seems ideally suited for multiplex message circuits, where a standard quality and high reliability are required.

Appendix I

We wish to show that a function of time $f(t)$ which contains no frequency components greater than W_0 cps is uniquely determined by the values of $f(t)$ at any set of sampling points spaced $1/2W_0$ seconds apart. Let $F(\omega)$ be the complex spectrum of the function, i.e.,

$$F(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} f(t)dt.$$

By assumption, $F(\omega)=0$ for $|\omega|>2\pi W_0$. $F(\omega)$ can be expanded in the interval $-2\pi W_0$ to $+2\pi W_0$ in a Fourier series having the coefficients

$$a_n = \frac{1}{4\pi W_0}\int_{-2\pi W_0}^{2\pi W_0} F(\omega)e^{-i(\omega n/2W_0)}d\omega. \qquad (1)$$

Now, since $F(\omega)$ is the Fourier transform of $f(t)$, $f(t)$ is the inverse transform of $F(\omega)$.

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}d\omega$$

$$= \frac{1}{2\pi}\int_{-2\pi W_0}^{2\pi W_0} F(\omega)e^{i\omega t}d\omega,$$

since $F(\omega)$ is zero outside these limits.

If we let $t=n/2W_0$, we have

$$f\left(\frac{n}{2W_0}\right) = \frac{1}{2\pi}\int_{-2\pi W_0}^{2\pi W_0} F(\omega)e^{i(\omega n/2W_0)}d\omega. \qquad (2)$$

Comparing (1) and (2), we see that

$$a_n = \frac{1}{2W_0} f\left(\frac{-n}{2W_0}\right).$$

Thus, if the function $f(t)$ is known at the sampling points, $\cdots$ $-(2/2W_0)$, $1/2W_0$, 0, $1/2W_0$, $2/2W_0 \cdots$, then the coefficients a_n are determined. These coefficients determine the spectrum $F(\omega)$ and $F(\omega)$ determines $f(t)$ for all values of t. This shows that there is exactly one function containing no frequencies over W_0 and passing through a given set of amplitudes at sampling points $1/2W_0$ apart.

To reconstruct the function, given these amplitudes, we note that

$$F(\omega) = \sum_n a_n e^{i(\omega n/2W_0)} \text{ for } |\omega| < 2\pi W_0$$

$$F(\omega) = 0 \qquad \text{ for } |\omega| > 2\pi W_0.$$

Taking the inverse transform, we have

$$f(t) = 2W_0 \sum_n a_n \frac{\sin \pi(2W_0 t + n)}{\pi(2W_0 t + n)}$$

$$= \sum_n f\left(-\frac{n}{2W_0}\right)\frac{\sin \pi(2W_0 t + n)}{\pi(2W_0 t + n)}$$

$$= \sum_n f\left(\frac{n}{2W_0}\right)\frac{\sin \pi(2W_0 t - n)}{\pi(2W_0 t - n)}.$$

In other words, the function $f(t)$ may be thought of as the sum of a series of elementary functions of the form $\sin x/x$ centered at the sampling points, and each having a peak value equal to $f(t)$ at the corresponding sampling point. To reconstruct the function $f(t)$, then, we merely need to generate a series of $\sin x/x$ pulses proportional to the samples and add the ensemble.

Appendix II

We wish to find the average power in a series of pulses of the form

$$f(t) = \frac{\sin \pi \dfrac{t}{\tau}}{\pi \dfrac{t}{\tau}}$$

occurring at the regular rate $1/\tau$.

The signal wave may then be written

$$v(t) = \sum_{k=1}^{n} V_k f(t - k\tau)$$

where V_k = peak amplitude of pulse occurring at the time $t = k\tau$. The average "power" (i.e., mean-square amplitude) S of the signal will then be

$$S = \overline{v^2} = \lim_{n\to\infty} \frac{1}{n\tau} \int_{-\infty}^{\infty} v^2(t)dt$$

$$= \lim_{n\to\infty} \frac{1}{n\tau} \Bigg[\sum_{k=1}^{n} V_k{}^2 \int_{-\infty}^{\infty} f^2(t - k\tau)dt + \sum_{j=1}^{n} \sum_{k=1}^{n} V_j V_k \int_{-\infty}^{\infty} f(t - j\tau) f(t - k\tau)dt \Bigg]_{j\neq k} .$$

For the assumed pulse shape, the first integral is equal to τ, while the second integral is equal to zero. Thus

$$S = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} V_k{}^2.$$

S is simply the mean-square value of the individual pulse peak amplitudes, and may also be written

$$S = \int_{-\infty}^{\infty} V^2 p(V)dV$$

where

$p(V)dV$ = probability that pulse amplitude lies between V and $V + dV$.

Suppose the pulses have b discrete amplitude levels $K\sigma$ apart, ranging from 0 to $(b-1)K\sigma$. Each pulse then has an amplitude $aK\sigma$ where a is an integer. The average power will be

$$S = K^2\sigma^2 \sum_{a=0}^{a=b-1} p(a)a^2$$

where $p(a)$ = probability of level a. If all levels are equally likely, $p(a) = 1/b$, and

$$S = K^2\sigma^2 \frac{1}{b} \sum_{0}^{b-1} a^2$$

$$S = K^2\sigma^2 \frac{(b-1)(2b-1)}{6} .$$

The quantity

$$\frac{1}{b} \sum_{0}^{b-1} a^2$$

is the square of the radius of gyration (i.e., the mean-square radius) about one end of a linear array of b points separated by unit distance. The average power of any amplitude distribution is the average of the squares of the amplitudes and is therefore proportional to the square of the radius of gyration of the distribution. The radius of gyration about any point is

$$r^2 = r_0{}^2 + d^2$$

where

r = radius of gyration about chosen point
r_0 = radius of gyration about center of gravity
d = distance to center of gravity from chosen point.

Obviously, $r_0 < r$, so that the average power will be least if the average amplitude is zero. S will be least if the pulse amplitude range is from $-K\sigma(b-1)/2$ to $+K\sigma(b-1)/2$, and will then be given by

$$S = K^2\sigma^2 \left[\frac{(b-1)(2b-1)}{6} - \left(\frac{b-1}{2} \right)^2 \right]$$

$$S = K^2\sigma^2 \frac{b^2-1}{12} .$$

This may also be written

$$S = \frac{A^2}{12} \frac{(b+1)}{(b-1)}$$

where A = total amplitude range $= (b-1)K\sigma$. As $b \to \infty$,

$$S \to \frac{A^2}{12} .$$

Thus, if all amplitude levels in a range A are possible and equally likely, the rms amplitude of the distribution will be $\sqrt{S} = (A/2\sqrt{3})$.