Digital Recording In Low Cost Transports

by **Clark E. Johnson, Jr.**

Clark E. Johnson, Jr. *is President of Micro Communications Corporation in Waltham, Massachusetts.*

Historically, most digital engineers have looked upon magnetic tape peripherals as black boxes which were fairly easily connected to the computer without much concern for the magnetic recording process and its requirements. The advent of microprocessors and microprocessor-based systems has greatly increased the need for low cost tape storage, requiring the software to do more and more of the data manipulation, encoding and decoding. This article reviews the system constraints imposed by the magnetic recording and playback process and shows you how to overcome them. This article is restricted to direct digital recording; it does not include such analog and quasi-analog techniques as frequency selective keying (FSK) or other tone-burst approaches.

introduction

A magnetic recording system may be thought of as a bandwidth limited communications channel. We must, of course, ignore the time delay between the writing and reading of a given unit of information. Fig 1 shows the usual model of a communications system. It consists of five components: an encoder, a modulator, the communications channel itself, a detector and a decoder. Some of these elements may be combined in whole or in part and each of them will be briefly explored as to its relevance in a magnetic recording system.

Since we are concerned only with recording digitally, we first must review how a pattern of ones and zeros is laid down on the tape. Clearly, only two unique states of magnetization are necessary for the storage of binary data. These states are normally positive and negative saturation of the magnetic tape surface to give maximum differentiation between the two states. Saturation recording takes maximum advantage of the non-linear saturation characteristic of the magnetic recording medium. It is to be clearly distinguished from a normal analog recording in which only about 25% of the magnetic moment of the media is put to use. Thus, the output from a saturated recording is approximately 12 db higher than that of an audio recording. The magnetic material in saturation recording provides inherent limiting action which greatly reduces the effects of variations in write current, tape-to-head spacing and media parameters. Because of the use of saturation techniques, you cannot use a direct correspondence between the input and output of a magnetic tape channel and a conventional signal transmission system.

the encoder

The function of the encoder is to take raw data at the input to the system (a bit stream of zeros and ones) and transform it into another data stream that has been specifically conditioned for transmission through the channel. This conditioning serves a number of purposes: for example, the addition of parity bits, error propagation limitation, spectral shaping and speed tolerance. Note that with a microprocessor-controlled system, software easily accomplishes this encoding and subsequent decoding. A later section will go into detail on various encoding techniques.

the modulator

The purpose of the modulator is to take the digital sequence produced by the encoder and convert it into an analog waveform that can then be transmitted through the channel. In a communications system with linear but bandwidth-limited channels, the modulator normally serves the purpose of producing a waveform whose spectrum is matched to the bandpass of the channel. In magnetic recording systems where

magnetic and mechanical characteristics of the media itself such as surface finish, coating thickness, coercive force and saturation magnetization. The most accessible parameter to the user is tape speed which is directly proportional to the frequency response of the channel. Since the fundamental limiting feature of the magnetic recording channel is the tape/read-head interface, the read head is considered a part of the channel.
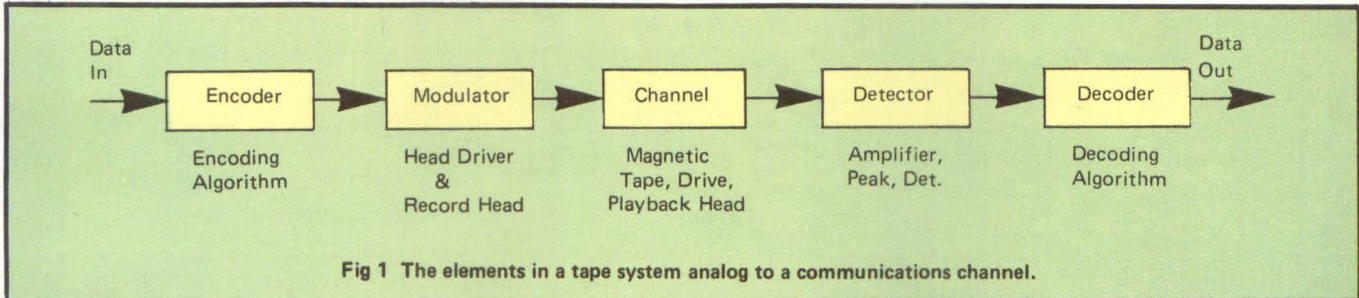

Data In → Encoder → Modulator → Channel → Detector → Decoder → Data Out

Encoder: Encoding Algorithm
Modulator: Head Driver & Record Head
Channel: Magnetic Tape, Drive, Playback Head
Detector: Amplifier, Peak, Det.
Decoder: Decoding Algorithm

Fig 1 The elements in a tape system analog to a communications channel.

the channel exhibits both saturation and hysteresis effects, as well as a frequency limitation, the modulator must serve the additional function of setting the proper write current amplitude and switching characteristic. The modulator in a tape system therefore consists of the magnetic recording head and the head driver. Recording takes place at the trailing edge of the record gap by the fringing flux, since this is the area of the highest flux gradient. Excessive current in the record head simply moves the magnetizing zone downstream from the gap.

the channel

In the usual communications channel, there is normally very little control over the channel characteristics, and the rest of the system is designed around it. Things are not so fixed in the magnetic recording channel. The transmission (storage) medium can be controlled in several ways. These include the

Fig 2 Relative transfer characteristic of the magnetic tape "channel" showing playback output voltage as a function of recorded wavelength.

Fig 2 shows the magnetic tape channel frequency response. Notice that a low frequencies the output rises at 6 db per octave in accordance with Faraday's Law. The output goes exactly to zero at the point where the recorded wavelength is equal to one-half of the playback head gap length. The peak of the response curve depends upon the magnetic parameters of the tape and can be pushed to higher frequencies by increasing the coercive force and decreasing the coating thickness. A main contributor to the roll-off of response with decreasing recorded wavelength on the tape is self-demagnetization. That is, as the recorded wavelength decreases, a decreasing amount of flux finds its way through the read head, and more of it is short-circuited through the tape itself. Clearly, significant advances in performance can be expected as tape materials and manufacturing techniques improve.
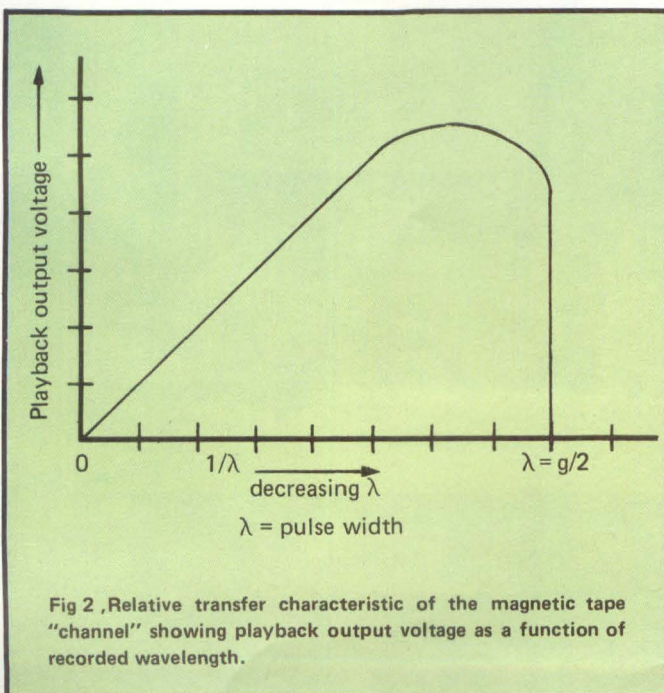
detector

The detector is the device that takes the incoming analog signal from the channel and converts it into a digital signal identical to that of the input to the modulator. Since, in the magnetic tape system, the read head is part of the channel, the detector is merely the amplifier and bit-by-bit detector, normally a peak detector. Note that the information content of the signal, located at the zero crossings of the record current, becomes translated to peaks of the detected voltage. The detector also recovers the timing information necessary to reconstruct the digital data stream.

decoder

The function of the decoder is exactly opposite that of the encoder: it takes the output of the detector and converts it back into the input data bit stream. Depending upon the type of encoding used, the decoder may correct errors in addition to performing the inverse algorithm of the encoder.

recording density limits

Since the recording process itself takes place at the trailing edge of the write-head gap, bit density on the tape does not

depend, at least to the first order, on the write-head-gap length. The cutoff wavelength (Fig 2) is directly proportional to the playback-gap length. There are, however, several other factors involved. One of these is head-to-tape spacing which greatly affects the playback output voltage and waveform. Reduction in output voltage causes errors in the detected signal because of insufficient amplifier gain to drive the detector. This reduction is dependent upon the ratio of band-to-tape spacing divided by read head gap length. Wave-shape broadening increases the effect of pulse crowding. Any non-uniformity of the head-tape interface such as debris on the tape or non-uniform surface, can cause an increase in head-to-tape spacing. As an example of how serious this is, a system which may exhibit one error in 10^7 bits at 800 bits-per-inch could easily exhibit one error in 10^3 bits at 1200 bits-per-inch using the same encoding scheme.

Pulse crowding is another effect that needs to be considered. In essence, the playback system (i.e., the detector) is linear and subject to the superposition theorem. The record process is very non-linear and depends upon the state of magnetization of the previous pulse, assuming that the magnetization did not reach zero before the next pulse came along. As a result, there is a phase shift of higher density pulses with respect to lower density pulses. Consequently, the detection window and amplitude of each pulse is dependent not only upon its own characteristics but those of its predecessor. Most manufacturers give specifications as to the acceptable maximum flux changes per unit length of tape for their system for low error rates.

signal processing techniques

There are basically two types of digital magnetic recording schemes: those providing a separate clock track independent-

ly recorded, and those which are self-clocking — that is, in which the data stream and the clock are encoded together into a single bit stream. Since this discussion is related to single-track recording, we are restricting ourselves to the latter case. One might legitimately ask why it is necessary to provide timing information at all. Why can't the data be recorded on tape and then played back using an independent timing oscillator identical to that used during recording? The problem is, of course, that tape systems, being mechanical, cannot provide the precise tape speed control necessary to make this feasible. In fact, it is even difficult to take the serial output from a UART, record it on tape, and then play it back on a different machine into a UART and recover error-free data. The reason for this is that the data timing between input and output can only change by 4% over one byte to faithfully recover the data. As a consequence, it is essential to provide independent clocking on the channel itself.

There are a number of self-clocking codes that can be used with digital recording. One author reports over a hundred of them, but most of them can be reduced to half a dozen or so basically different schemes. If we now refer to Fig 2, we note that dc cannot be transmitted through the magnetic tape channel. In fact, to reduce equalization, it is desirable to reduce the bandwidth spread as much as possible. Fig 3 shows the characteristics of various types of self-clocking codes (with NRZ included for the sake of comparison).

The abscissa of Fig 3 is the fundamental frequency bandwidth component required in terms of the data rate. The columns at the right indicate various characteristics of these codes. "DC Present" is whether or not the encoded waveform is asymmetric; that is, does it include a dc component. The second column is the "Ratio of the Transition Density to the Bit Density," which is a measure of the efficiency of

the recording scheme. Note that NRZ recording has a ratio of one. The third column gives the bandwidth requirements of the channel in terms of the "Ratio of the Maximum Frequency to the Minimum Frequency." The fourth column is whether a "Preamble" is required to provide read clock synchrony. A preamble is a known bit length and pattern appended to the front of each record. Generally, a preamble is required for all systems, but it may be only one bit long, as in biphase or ratio recording. Methods which do not provide a clock pulse for every bit require long, formalized preambles. Only long preambles have a "yes" in this column. The next two columns give the relative "Speed Tolerance" on both a bit-to-bit basis and on a long-term basis. Finally, the last column gives some additional comments.

Now let's look at each of these encoding schemes by itself and try to draw some conclusions about them. Fig 4 shows the encoded data stream for some of the encoding schemes of Fig 3 for an arbitrary 16 bits of data.

RZ recording starts out with unmagnetized tape (not such an easy requirement if one wants to reuse an already recorded tape) and simply goes positive for a one and negative for a zero, always returning to the demagnetized state between bits. The densities are relatively low, due to need to get back to zero between bits. The bandwidth requirements run roughly from 0.25 times the data rate to the data rate. RZ recording is particularly attractive for low density recording because it is virtually asynchronous, making minimal constraints on the density of the recorded data. RZ recording does not require a preamble.

S-NRZ recording is simply NRZ recording with an extra bit at every eight bits to provide a lower bound to the bandwidth requirements. Rather tricky electronics and buffering are required to squeeze an extra bit in for every eight in encoding and then clip it out again in decoding.

Bi-phase recording is a class of double frequency self-clocking schemes, many of which require no preamble and have a bandwidth requirement of only two to one. Bi-phase goes under many names, such as Manchester Code, Phase Enc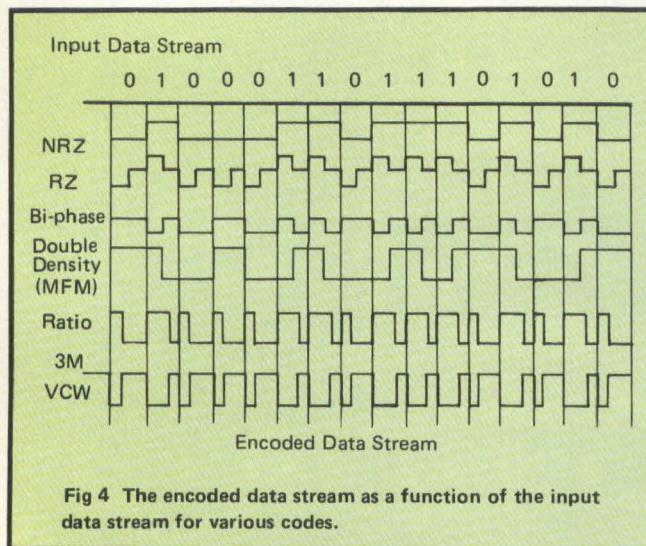oding and Frequency Modulation. These are all essentially the same scheme in which a one bit is represented by two flux changes and a zero bit by one flux change. All the other schemes in this category are simply variations on this with inversions and phase shifts brought into play. All have essentially the same mathematical characteristics as far as the channel is concerned. Bi-phase (which I have elected to use as the generic name) generally has no dc component present and is relatively insensitive to small speed changes both on a bit-to-bit or over a long-term basis. Fancy electronics can take into account long term speed changes by simply altering the sampling clock rate to agree with the average bit cell.

Double density recording, used in various disk files, do not provide a clock bit for every data bit and thus requires a preamble to provide read clock synchronization. With the exception of ZM recording, discussed below, all double density schemes have dc present and are very sensitive to speed changes. Their bandwidth requirements, however, are equal to one-half the data rate.
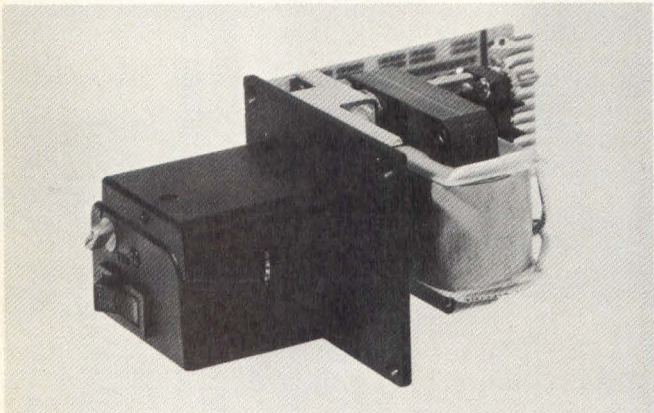
Ratio recording is a relatively inefficient scheme, but one which has significant virtues for low cost systems. In ratio recording, a positive-going flux change at the leading bit cell edge always corresponds to a clock pulse while the position of the negative-going change (whether it is in the first half or the second half of the bit cell) determines the data. Since each bit cell stands alone, no preamble is required; a lost bit has no effect on adjacent bits as happens with double density and some bi-phase schemes. In addition, the speed tolerance is theoretically ±50% on a bit-to-bit basis and very high on a long term basis, limited only by the ability of the amplifier to provide sufficient gain for slow-moving tape or to handle the bandwidth requirements of rapidly-moving tape. You pay a price for this capability, however, in that there are essentially three flux changes per bit cell, thereby limiting the maximum recorded density. Detection is done by charging a capacitor (or turning on an up-counter) during the first portion of the cycle and then discharging the capacitor (or causing the counter to count downward) during the second part of the cycle Since the capacitor is discharged at the end of each bit cell to start over again (or the counter reset to zero), the presence of dc in the detector is of no consequence.

Zero modulation or ZM recording was developed by IBM for their 3850 Honeycomb Mass Storage System. This is simply an algorithm of other double density schemes, but one which has zero dc component, thereby eliminating the accumulated unbalance between positive and negative pulse durations which cause baseline shift. This scheme is very sophisticated in terms of generating the encoded bit stream, since it requires an algorithm which looks both forward and backward in the data pattern, thereby requiring external memory. It is too sophisticated for simple cassette and cartridge tape systems.

Variable Cell Width, developed by 3M for use with their DC-100 transport, uses a combination of bi-phase and ratio recording. In this scheme, the bit cell length for a one is 50% longer than that for a zero. A zero is two pulses of opposite polarity for half the time, whereas for a one, the leading or

Input Data Stream

0 1 0 0 0 1 1 0 1 1 1 0 1 0 1 0

NRZ
RZ
Bi-phase
Double Density (MFM)
Ratio
3M
VCW

Encoded Data Stream

Fig 4 The encoded data stream as a function of the input data stream for various codes.

negative pulse is equal to the leading or negative pulse for the zero, but the positive pulse is twice as long.

what encoding scheme to use?

This is not an easy question to answer since it depends upon the following factors:
- Mechanical stability
- The nature of the data being recorded, and
- Whether one can use error-correcting codes.

The mechanical stability of the system includes such things as the need for interchangeability of tapes recorded on one machine to play on another, the accuracy and uniformity of tape speed and the mechanical rigidity of the tape-handling system to minimize structurally caused azimuth problems, etc. Apropos of the latter, the alignment between the gap in the playback head and the recorded data on the tape is very critical at high flux-change densities and requires the tape to pass over the head at a constant angle no matter what machine the tape is played on and no matter when and under what conditions. At high data densities, a slight azimuth misalignment of the tape with the head will cause a serious decrease in output and increased errors. Nonuniform tape speed requires the use of a speed-tolerant recording system such as ratio recording.

A second factor is the nature of the recorded data. Some of the encoding techniques require a preamble which must be appended to each record — expensive in terms of tape utilization if the records are short. Clearly, it is not any problem for a long, unformatted record since the preamble will represent only a minuscule fraction of the total record length. However, for short records, the requirement to append a preamble will cause an appreciable increase in the record length, thereby negating the gains made by going to a higher density recording scheme. In addition, the use of double density schemes on tape is particularly tricky, since they are virtually intolerant of any long term speed changes because their detection window is extremely narrow and difficult to change in concert with a change in data rate.

Finally, the use of error-correcting codes, either cyclic codes added on to fixed-length records or other more exotic codes, can provide automatic error correction in the decoding process itself. Again, the addition of error-correcting codes may or may not be required, depending upon what is to be done with the data. Most users find it sufficient to add a checksum at the end of a record in order to tell if the record has been received correctly by the decoder.

using a microprocessor to replace the encoder and decoder

Clearly, a microprocessor can be used to generate the encoded waveforms from the data stream, based on which encoding algorithm is required. The microprocessor can be run in interrupt mode during encoding, since the timing for writing the data is not critical. In playback, however, since the timing comes off the tape, the microprocessor must be dedicated to the reading of tape. If the system already has a microprocessor in it, then it may be easy to dedicate a small amount of the program to writing and reading tape, thereby saving the considerable expense of hardware implementation of encoding and decoding.

CIRCLE 29▶