HISTORY OF COMMUNICATIONS

EDITED BY MISCHA SCHWARTZ

INTRODUCTION

The article for this month's History of Communications Column, a summary by Hisashi Kobayashi of the history of partial-response signaling, is one in a continuing series in which we have pioneers in a significant communications discipline describe their work and that of associates in the field. You will note that Dr. Kobayashi, while focusing on the applications of partial-response maximum likelihood technology to digital magnetic recording (he was working for IBM at the time), does not neglect to signal out the early pioneering work of the late Adam Lender on duobinary transmission, the analogous technology applied to communication transmission. Other early contributors to this work are noted as well. Such is continuously the case with many of our significant systems and technologies: they have relevance and application in multiple fields. This is what makes the study of the history of communications so fascinating and so up-to-date. As one of my younger colleagues mentioned just the other day, we are continually in danger of re-inventing the wheel. This is why it is important to scan the history of our field and related areas, not just for the excitement of revisiting the early stages of an important invention or system concept, but to note that the original ideas and concepts of the pioneering workers in the area still have relevance and significance today. As Dr. Kobayashi cogently notes, these original ideas and developments of the 1960s and early 1970s in partial-response signaling have evolved into the vital and huge magnetic recording industry of today.

PARTIAL-RESPONSE CODING, MAXIMUM-LIKELIHOOD DECODING: CAPITALIZING ON THE ANALOGY BETWEEN COMMUNICATION AND RECORDING

HISASHI KOBAYASHI, PRINCETON UNIVERSITY

ABSTRACT

Signal processing and coding technology for digital magnetic recording is the core technology of the channel electronics module in a hard disk drive (HDD) that processes signals read from magnetic media. In this historical review I focus on what is now widely known as partial-response, maximum-likelihood (PRML) technology, which takes advantage of the inherent redundancy that exists in signals read out of magnetic media; its theoretical foundation goes back to 1970, and it capitalizes on the analogy between high-speed data transmission and high-density digital recording, and that between a convolutional code and a partial-response signal.

The first PRML-based product was introduced by IBM in 1990, and PRML technology soon became the industry standard for all digital magnetic recording products, ranging from computers' HDDs and tape drives to micro hard discs used in PCs, mobile phones, and MP3 players; use of the PRML principle has recently been extended to optical recording products such as CDs and DVDs. Its improved version, called NPML (noise-predictive, maximum-likelihood), and variants have been adopted by the HDD industry since 2000.

Today, a large number of communication and information theory researchers are investigating use of advanced techniques such as turbo coding/decoding to further improve the density and reliability of both magnetic and optical recording systems.

INTRODUCTION

The IBM RAMAC, the first HDD introduced in 1956, had storage capacity of a mere 4.4 Mbytes, and the price per megabyte was as high as $10,000, whereas 50 years later, in 2005, a microdrive contained 9 Gbytes, and the price per megabyte is less than $0.03. In this 50-year period the areal density has grown from 2 × 10–3 Mb/in^2 to 3.4 × 104 Mbs/in^2, ia phenomenal gain of 17 million times! Such dramatic growth in storage capacity and shrinking cost per bit is a result of the compounding effects of significant progress made in key components: track position control, head sensitivity, high-speed writing, media signal-to-noise ratio (SNR), head disk spacing, and signal processing. The signal processing and coding technology for HDDs is the essence of the channel electronics module in an HDD that processes signals read from the magnetic media [1].

PRE-1970 SIGNAL PROCESSING AND CODING FOR MAGNETIC RECORDING

The conventional method of magnetic recording used either the non-return-to-zero (NRZ) or NRZ-inverse (NRZI) method. In NRZ recording, one direction of magnetization corresponds to a 1, while the opposite direction corresponds to a 0 in data; in NRZI, 1 is recorded as a transition of magnetization and 0 as no transition. If the read-head uses an inductive coil head, the rate of change in the magnetic flux as the read-head passes over the medium will be proportional to the induced voltage at the read-head output. Thus, the relationship between the readback voltage $r(t)$ and the magnetization $m(t)$ should be written as [2]

$$r(t) = h(t) \otimes \frac{dm(t)}{dt}, \tag{1}$$

where ⊗ means the convolution operation, and $h(t)$ represents the magnetic-head field distribution characterized by the response due to a unit step function in $m(t)$. The conventional detection method of NRZI recording interpreted the presence of a pulse in the readback signal as 1 and the absence of a pulse as 0. This was often realized by passing the output voltage signal through a rectifier and then through a threshold detector. Furthermore, the conventional signal processing method for the readback signal used the so-called peak detection (PD) method (see, e.g., [3]), in which peak levels in the output voltage signal were searched, and the sampled values were compared to the threshold for binary decision. But as one attempted to store the information bits more densely on the medium, the PD method failed because:

- The height of a peak became not much larger than background noise.
- Neighboring peaks came closer and collapsed into one peak.
- The position of peaks significantly shifted, sometimes beyond the neighboring bit boundaries.

These "pulse crowding" effects set the limit on recording density in the conventional technique. The run-length limited (RLL) codes pioneered by Donald Tang [3, 4, references therein] were the main techniques available to mitigate adverse effects of pulse crowding.

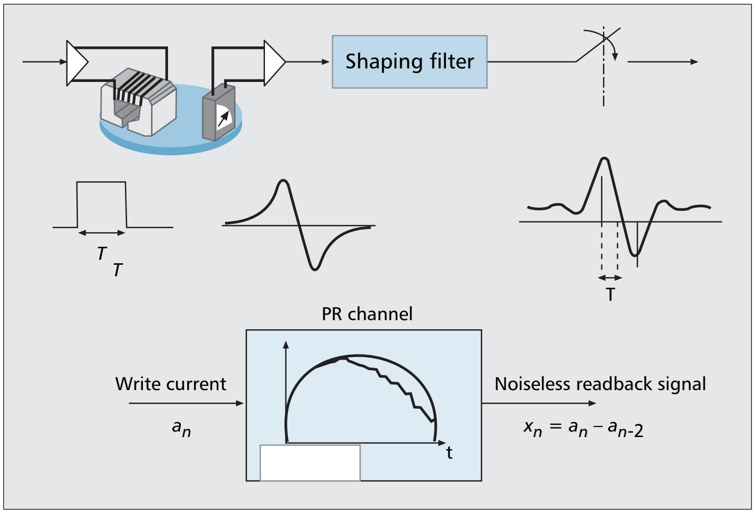

Figure 1. *Partial-response class-4 (PR4) channel:* $G(D) = 1 - D^2$. *The sampling rate is 1/T.*

ANALOGY BETWEEN MAGNETIC RECORDING CHANNEL AND PARTIAL-RESPONSE CHANNEL

The "pulse crowding" effect alluded to by the digital recording community prior to 1970 was equivalent to intersymbol interference (ISI) in digital data transmission. Unlike analog signal (e.g., audio) recording, digital recording uses *saturation recording* in that the driving current in the recording head coil is switched from one saturated level to the opposite saturated level so that the readout signal should have large SNR. This magnetization process is inherently nonlinear.

I joined the Communication Theory group at the IBM Research Center at Yorktown Heights, New York in 1967, and my primary assignment was to investigate methods to mitigate the ISI problem in data transmission over voice-grade lines. I got attracted, as a side line, to magnetic recording research on which my colleague Don Tang was working. Although I immediately noticed the similarity between pulse crowding and ISI, my attempt to treat the digital recording system as a linear channel was not readily accepted by magnetic recording experts in IBM. Use of saturation recording, its hysteresis characteristics and signal-dependent noise, all compounded to discourage them from treating a magnetic recording system as a linear system.

But once binary information is stored as a saturated bipolar signal $m(t)$, the readout process is a linear operation as given in Eq. 1. Thus, my argument was that if nonlinear distortion introduced in the writing process was negligible or could be precompensated by proper shaping of the writing current, the magnetic recording system could be approximated by a linear model as far as the readback process is concerned. If the ISI was introduced by an increase in recording density, it should be eliminated by an *equalizer*; so went my argument.

My 1970 paper with Don Tang [2] proposed that the recording channel should be treated just like a data transmission channel, and that the readout signal $x(t)$ should be sampled at regular intervals, $t = nT$ $(n = 0, 1, 2, \ldots)$, instead of sampling $x(t)$ at instants of peak values as practiced in the conventional peak detection method. If the ISI is removed by an equalizer, the sampled output $x_n = x(nT)$ is a three-level signal, represented by + 1, 0, –1 after proper scaling. In NRZ recording the sampled sequence $\{x_n\}$ is related to the binary data sequence $\{a_n\}$ by

$$x_n = a_n - a_{n-1},\ n = 0, 1, 2, \ldots, \tag{2}$$

which can be compactly written in a polynomial form,

$$X(D) = (1 - D)A(D) = G(D)\,A(D), \tag{3}$$

where D is the delay operator. The transfer function $G(D) = 1 - D$ is the "difference operator," which is a discrete-time counterpart of the "differential operator" involved in the readback process represented by Eq. 1.

In data transmission the subject of my primary assignment, I learned that Adam Lender (1921–2003) of GTE Lenkurt discovered in 1963 that as he increased the transmission rate of binary signals close to the Nyquist rate of a bandlimited channel, the ISI became so pronounced that the output signal suddenly turned into three level signals: if two adjacent pulses are both positive and move close to each other, they merge into a large positive pulse; if two negative pulses push closer together, they end up as a large negative pulse; if the adjacent pulses are opposite in their polarities, they result in zero by canceling each other as they are pushed closer together. So the sampled channel output forms a three-level sequence. If we label these three levels 0, 1, and 2, the corresponding channel is represented by $G(D) = 1 + D$; Lender called this high-speed signaling scheme the *duobinary technique* [5]. Similarly, he termed a data transmission channel with $G(D) = 1 - D^2$ *modified duobinary*. A general class of signaling scheme that can be characterized by a finite polynomial $G(D)$ with integer coefficients is referred to as *correlative-level* coding (see Adam Lender, *IEEE Spectrum*, February 1966). Ernest R. Kretzmer of Bell Telephone Laboratories coined the term *partial-response channel* for this class of binary data transmission channels, and referred to duobinary and modified duobinary as Class-1 and Class-4, respectively [6].

Note that $G(D) = 1 + D$ in Lender's duobinary signaling is a result of *intentionally* pushing the transmission speed well beyond the conventionally tolerable rate, whereas the term $G(D) = 1 - D$ we defined for the magnetic recording channel is due to the *inherent* differential operation in the readout process. But mathematically they are quite similar.

Don Tang and I showed in [2] that a magnetic recording channel can be shaped into a partial-response channel with the transfer function $G(D) = (1 - D)\,P(D)$, where $P(D)$ is any polynomial of D. The simplest choice is $P(D) = 1 + D$, which gives $G(D) = (1 - D)(1 + D) = 1 - D^2$, which we termed *Interleaved NRZI* [7]. The overall transfer function of Interleaved NRZI is equivalent to Lender's modified-duobinary and Kretzmer's partial-response Class-4 for data transmission. Thus, in the magnetic recording community, our interleaved scheme is often referred to as the "PR4" signal [3, 8] (Fig. 1). The next simple choice is $P(D) = (1 + D)^2 = 1 + 2D + D^2$, also proposed in our paper [2], which results in $G(D) = (1 - D)(1 + D)^2 = 1 + D - D^2 - D^3$. This partial-response channel is referred to as *extended PR4* or EPR4 in the magnetic recording community [3].

MAXIMUM-LIKELIHOOD DECODING ALGORITHM AND EQUALIZATION OF THE PR SIGNAL

From September 1969 to April 1970 I took a sabbatical leave from IBM Research to teach signal detection theory and information theory in the System Science Department of the

University of California at Los Angeles, where I had an opportunity to learn directly from Andrew Viterbi about his new nonsequential decoding algorithm for convolutional codes [9], that is, the Viterbi algorithm he published in 1967. Jim Omura, who joined the department as an assistant professor in 1969, had just shown the equivalence of the Viterbi algorithm to Bellman's dynamic programming (*IEEE Transactions on Information Theory*, January 1969).

I soon recognized an analogy between a convolutional encoder and a partial-response channel: they can both be represented as a linear finite state machine, the former being defined over a binary Galois field and the latter over the real number field. Then it became quite apparent that the Viterbi algorithm should be equally applicable to a partial-response (PR) channel. The analysis and simulation I performed soon after I returned to IBM Yorktown Heights confirmed that the maximum likelihood (ML) decoding algorithm could gain as much as 3 dB in SNR compared with bit-by-bit detection. Its advantage over the "ambiguity zone" detection method [10] — an algebraic decoding algorithm with an "erasure" option that I had been working on with Don Tang — was also demonstrated. I published these results in the *IBM Journal of Research & Development* [11] for the magnetic recording audience, and in the *Information Theory Transactions* [12]. These papers [2, 11, 12] laid the theoretical foundations of what was later called PRML in the digital recording community [3, 8].

Around the same time Dave Forney was developing the idea of applying the Viterbi algorithm to a general class of ISI channels, as discussed in his seminal paper [13]. Digital communication products based on Forney's maximum likelihood sequence estimation (MLSE) scheme, referred to as the Viterbi equalizer in GSM-related literature, were introduced to the mass market finally around 1995.

DEVELOPMENT OF PRML-BASED HDD PRODUCTS

Although the potential significance of the proposed scheme of combining the partial-response (PR) channel coding and maximum-likelihood (ML) decoding was recognized by some of IBM's magnetic recording experts, the scheme was considered too expensive to implement circa 1970, when microprocessor-based signal processing technology was in its infancy. Even analog-to-digital conversion was an expensive proposition. In 1971 the mission of communications research within the IBM Research moved to the Zurich Laboratory, and I was appointed manager of a newly created System Measurement and Modeling group in the Computer Science Department; thus, I was no longer able to further work on PRML or push its technology transfer. Several industrial laboratories in the United States and Japan reportedly conducted experiments and built prototypes (e.g., Robert Price of Sperry Research Center and the late Dr. K. Yokoyama of NHK Laboratory in Tokyo) by 1980. In the 1980s a team of several researchers led by François Dolivo in Gottfried Ungerboeck's group at IBM Zurich Research Laboratory conducted extensive simulations and built a working prototype that incorporated novel timing recovery and equalization algorithms during the 1980s, and they succeeded in transferring PRML technology to the IBM Storage System Division in Rochester, Minnesota. Their series of technological developments are reported in [8, references therein].

In 1990 IBM Corporation introduced a new generation of 5.25-inch HDD by incorporating a PRML channel. Magnetoresistive (MR) read heads, another major breakthrough technology, were incorporated in the following year, 1991. Since then, practically all HDDs have adopted the MR read heads and PRML channel, and the rate of increase in HDD areal density has jumped from the traditional 25 percent compound growth rate (CGR) to 60 percent CGR or higher, as an external analog filter, digital finite impulse response (FIR) filter, and equalization technology associated with the PRML channel were further improved, together with great advances in the MR read head and film disk technologies.

The PRML technology is now adopted not only in HDDs, but also tape drives and micro hard discs installed in laptop PCs, cell phones, and MP3 players; the PRML principle has recently been extended to optical recording products such as CDs and DVDs.

NOISE PREDICTIVE MAXIMUM LIKELIHOOD

Evangelos Eleftheriou and his coworkers at IBM Zurich Laboratory [14] more recently proposed to enhance the performance of the traditional PR equalizer by using noise prediction techniques. The resulting noise-predictive PR equalizer consists of a forward linear PR equalizer followed by a linear predictor to whiten noise. Their scheme, which combines the noise-predictive PR equalizer and ML sequence estimation, is termed noise-predictive maximum likelihood (NPML) detection. Introduction of NPML into HDD products since 2000 has led to a 50–60 percent increase in recording density and has resulted, together with the introduction of the giant magneto-resistive (GMR) read sensor, in 100 percent CGR in areal recording density.

Sophisticated signal processing techniques such as PR channel coding, maximum likelihood sequence estimation, and noise predictive equalization, contribute to the significant increase in density. With use of a proper Reed Solomon code and run-length limited (RLL) code, a BER as low as 10^{-15} can be achieved. Today a read channel architecture based on NPML detection and noise-predictive parity-based post-processing techniques has become the new de facto industry standard for HDDs.

RECENT PROGRESS IN PRML SYSTEMS

Signal processing and coding for PRML-based digital recording, both magnetic and optical, is now a well established area of research and development, actively pursued by researchers with communication and information theory backgrounds. Turbo decoding or iterative decoding of partial-response channel output sequence has been discussed by Kobayashi and Bajcsy [15], Souvignier *et al.* (*IEEE Transactions on Communications*, August 2000) and Bajcsy *et al.* (*IEEE Journal on Selected Areas in Communications*, May 2001). Kavcic *et al.* (*IEEE Transactions on Information Theory*, May 2005) discuss low density parity check (LDPC) codes for partial response channels. Recent studies of hidden Markov models (HMMs) show that the Viterbi algorithm and maximum a posteriori (MAP) algorithm used in turbo decoding are special cases of forward-backward algorithms (FBAs) for hidden Markov chains, and the FBA in turn is a special case of the expectation-maximization (EM) algorithm. Therefore, we anticipate a further advance in algorithmic developments for signal processing of digital recording data.

ACKNOWLEDGMENTS

I would like to thank my former IBM colleague, Dr. Donald D. T. Tang, who introduced me to magnetic recording research; and Drs. François Dolivo, Evangelos Eleftheriou, and their team members at IBM Zurich Laboratories and IBM Storage System Division in Rochester, Minnesota for their efforts in turning the theoretical concept into a working prototype and finally to real products. I am also indebted to the late Dr. Adam Lender, Drs.

Andrew Viterbi, Jim Omura, and David Forney for sharing their knowledge and insights with me during my research on PRML. This article draws on my joint paper with François Dolivo and Evangelos Eleftheriou [1]. I thank Prof. Mischa Schwartz for inviting me to prepare this article, and Dr. Dolivo and anonymous reviewers for their suggestions to improve this manuscript. Because the editorial policy requires that the number of references be limited to 15, I fear that I am doing injustice to many authors by not including their worthy papers.

REFERENCES

[1] H. Kobayashi, F. Dolivo, and E. Eleftheriou, "35 Years of Progress in Digital Magnetic Recording," *Proc. 11th Int'l. Symp. Problems of Redundancy in Info. and Control Sys.*, Saint-Petersburg, Russia, July 2–6, 2007, pp. 1–10.

[2] H. Kobayashi and D. T. Tang, "'Application of Partial-Response Channel Coding to Magnetic Recording Systems," *IBM J. R&D.*, vol. 14, no. 4, July 1970, pp. 368–75.

[3] P. H. Siegel and J. K. Wolf, "Modulation and Coding for Information Storage," *IEEE Commun. Mag.*, vol. 29, no. 12, Dec. 1991, pp. 68–86.

[4] H. Kobayashi, "A Survey of Coding Schemes for Transmission or Recording of Digital Data," *IEEE Trans. Commun. Tech.*, COM-19, no. 6, Dec. 1971, pp. 1087–1100.

[5] A. Lender, "The Duobinary Technique for High-Speed Data Transmission," *IEEE Trans. Commun. Elec.*, vol. 82, May 1963, pp. 214–18.

[6] E. R. Kretzmer, "Generalization of a Technique for Binary Data Transmission," *IEEE Trans. Commun. Tech.*, COM-14, Feb. 1966, pp. 67–68.

[7] H. Kobayashi and D. T. Tang, "Magnetic Data Storage System with Interleaved NRZI CODING" U.S. Patent no. 3,648,265, Mar. 7, 1972.

[8] R. D. Cideciyan *et al.*, "A PRML System for Digital Magnetic Recording," *IEEE JSAC*, vol. 10, no. 1, Jan. 1992, pp. 38–56.

[9] A. J. Viterbi, "'Error Bounds for Convolutional Codes and Asymptotically Optimum Decoding Algorithm," *IEEE Trans. Info. Theory*, IT-13, no. 2, Apr. 1967, pp. 260–69.

[10] H. Kobayashi and D. T. Tang, "On Decoding of Correlative Level Coding with Ambiguity Zone Detection," *IEEE Trans. Commun. Tech.*, COM-19, no. 8, Aug. 1971, pp. 467–77.

[11] H. Kobayashi, "Application of Probabilistic Decoding to Digital Magnetic Recording Systems," *IBM J. R&D*, vol. 15, no. 1, Jan. 1971, pp. 69–74.

[12] H. Kobayashi, "Correlative Level Coding and Maximum Likelihood Decoding," *IEEE Trans. Info. Theory*, IT-17, no. 5, Sept. 1971, pp. 586–94.

[13] G. D. Forney, Jr., "Maximum Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," *IEEE Trans. Info. Theory*, IT-18, no. 3, May 1972, pp. 363–78.

[14] J. D. Coker *et al.*, "Noise-Predictive Maximum Likelihood Detection," *IEEE Trans. Magnetics*, vol. 34, no. 1, Jan. 1998, pp. 110–17.

[15] H. Kobayashi and J. Bajcsy, "System and Method for Error Correcting A Received Data Stream in A Concatenated system," U.S. Patent no. 6,029,264, Feb. 22, 2000.

BIOGRAPHY

HSASHI KOBAYASHI [LF] is the Sherman Fairchild University Professor Emeritus of Princeton University, where he served as dean of the School of Engineering and Applied Science (1986–1991). Prior to joining the Princeton faculty he was with the IBM Research Division (1967–1986), where he held many managerial positions, including founding director of the IBM Tokyo Research Laboratory (1982–1986). Among his technical contributions is his 1970 invention of the high-density digital recording scheme called partial-response coding and maximum-likelihood decoding (PRML) discussed in this article. For this contribution he was awarded, together with Drs. François Dolivo and Evangelos Eleftheriou of IBM Zurich Research Laboratory, the 2005 Eduard Rhein Technology Award. He has also contributed to data transmission theory and system performance evaluation methodology, especially diffusion process approximation, queuing and loss network models, and their computational algorithms. He authored *Modeling and Analysis* (Addison Wesley, 1978) and coauthored with Brian L. Mark *System Modeling and Analysis* (Pearson/Prentice Hall, 2008). He received the Humboldt Prize (Senior U.S. Scientist Award) from the Alexander von Humboldt Foundation (1979) and IFIP's Silver Core Award (1980). He was elected to the Engineering Academy of Japan (Japan's national academy of engineering) in 1992. He has served as a scientific advisor for numerous organizations in the United States, Japan, Canada, and Singapore. Currently he resides in Manhattan and is authoring textbooks on probability, statistics, and random processes; network protocols, performance, and security; and digital communications and networks. He also serves as a technical advisor for the National Institute of Information and Communications Technology of Japan on their new-generation network architecture project called AKARI.