

CODING METHODS FOR HIGH-DENSITY OPTICAL RECORDING

by K. A. SCHOUHAMER IMMINK

Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

Abstract

Maximum-likelihood sequence estimation of binary coded and uncoded information, stored on an optical disc, corrupted with additive Gaussian noise is considered. We assume the presence of intersymbol interference and channel/receiver mismatch. The performance of the maximum-likelihood detection of runlength-limited sequences is compared against both uncoded information and information encoded by Hamming-distance-increasing convolutional codes.

ECO: 5.6.

1. Introduction

Optical disc recording has become an established method of storing large amounts of digital or analogue information. The optical disc contains a spiral-shaped track of successive shallow depressions, usually called pits, in a reflective layer. The information is stored in the lengths of the pits and the intervals of land between them¹). The rotating disc is scanned by a focused laser beam. The reflected light, modulated by the information layer, is detected with a photodiode and subsequently electronically processed. The read-out is contactless: electromechanical servosystems focus the laser spot on the disc and follow the track within the specified accuracy. An important specification of a disc-oriented storage device is the amount of information that can be stored per unit of surface. This quantity is determined by two parameters: the track pitch and the linear information density. The difficulties encountered when improving the linear information density by using codes will be considered in this paper.

A major impairment found in high-density digital optical storage systems is intersymbol interference (ISI). It is well known that a single pulse read out by the system is smeared in time due to the convolution with the light intensity profile of the read-out spot. A sample at the centre of a symbol interval is a

weighted sum of amplitudes of pulses in several adjacent intervals. To make matters worse, the impulse response of the read-out mechanism is strongly time-variant. The time-variance is a consequence of the constantly changing physical characteristics of the read-out mechanism. Possible sources of disturbance are: defocusing, mistracking and reflection variations of the disc. For example, a 1–2 μm defocusing leads to a dramatic reduction of the resolving power (or bandwidth) of the read-out mechanism. Fingerprints on the disc may reduce the amount of reflected light to 20% of the nominal amount. To the designer of the recording device the impulse response of the channel, disc plus read-out, appears to be of a random nature. All the effects mentioned, combined with random additive noise, may lead to errors in the retrieved data.

Many so-called recording codes (or modulation codes) were designed to enhance the information density and the reliability of the recorded information²). Recording codes should not be confused with error-correcting codes. Error-correcting codes, mostly based on Reed-Solomon codes, are normally used in recording practice as an outer code of a recording code. This partitioning into two codes is not strict, but at this moment there seems to be no better practical solution. A complication of a code design for information storage systems of these kind is that the designer does not have a free hand in selecting code patterns. The physics of the optical disc only allows the recording of 'full- T ' pulses having positive or negative polarity, called pits and lands, respectively.

Runlength-limited (RLL) sequences are the state-of-the-art cornerstone of all current data storage systems, whether their nature is magnetic or optical. RLL sequences possess the property that the minimum and maximum runlengths, i.e. the number of consecutive like symbols, is constrained between $(d + 1)$ and $(k + 1)$, where $d \geq 0$ and $k > 0$, $d < k$, are predefined parameters. In the Compact Disc system, which is a 1 Gbyte optical read-only memory, the EFM code based on RLL sequences ($d = 2$, $k = 10$) is used³). ISS used a ($d = 2$, $k = 11$) code, called 3PM⁴), in its 8434 disk drives and a ($d = 1$, $k = 7$) code in its 8470 drive⁵). In the IBM 3370–3380 family of high-end magnetic disk drives, considered by many as the workhorses of main-frame computer memories, the Eggenberger ($d = 2$, $k = 7$) code is used as the recording code^{6,7}). The detection circuitry normally employed is threshold detection in the Compact Disc case and a peak detector is often used in magnetic recording⁸).

The data transmission literature shows that much progress has been made in the improvement of channel utilization by using the so-called maximum-likelihood sequence estimator (MLSE), sometimes called Viterbi detector^{9–11}). Here the complete received sequence is used to detect any symbol or group of

symbols. Only quite recently MLSE detection of RLL sequences has found interest in (magnetic) recording practice^{12,13}). MLSE in the presence of inter-symbol interference rests on the basic assumption that the receiver has a prior knowledge of the channel characteristics. In many practical situations the receiver does not have exact information about some of the parameters, such as time-varying impulse response, sampling instants, colour of the noise, etc. These unknown parameters should be measured (estimated) at the receiver but in practice there are always measurement errors. As an additional complication the impulse response might change so rapidly that the adaptation circuitry lags behind the actual situation. The MLSE detection method will not function properly if insufficient or even incorrect knowledge of the actual channel characteristics is available. The ability of the detection method in combination with the applied recording code to cope with tolerances of the channel characteristics, the robustness of the method, is one of the major topics to be considered here. We shall not statistically characterize the time-variant channel, but rather investigate the performance of MLSE plus recording code when the detection is based on incorrect assumptions regarding the channel characteristics. In particular we will deal with the robustness of recording codes in combination with the detection method in the face of gain and bandwidth variations of the channel.

It is customary when dealing with ISI-free channels to define the so-called coding gain. The coding gain is the improvement of noise margin over the uncoded case when coding is applied. The (asymptotic) coding gain equals the product of the code rate and the code's minimum (free) Hamming distance. The minimum Hamming distance of a set of distinct RLL sequences is unity so that there is an obvious coding loss when these sequences are used on ISI-free channels. RLL-based codes have not found widespread interest in the coding literature and their performance has occasionally been misunderstood. Most of the misunderstanding of the performance of RLL sequences stems from the fact that in general the performance of a code is calculated assuming the receiver has an ideal prior knowledge of the channel. When, however, the actual channel differs from the channel model a completely different result can be expected. The next sections show the value of RLL sequences on recording channels with severe ISI and unknown gain or bandwidth variations when MLSE detection is employed. This is quite surprising since this class of codes was conceived for detection with simple detectors such as for example a peak detector.

Sec. 2 deals with some properties of RLL sequences followed in sec. 3 by the basics of MLSE detection in the presence of ISI. The error event probability of the MLSE detector, when channel and receiver are mismatched, is cal-

culated in sec. 4. In sec. 5 the theory is applied to the optical recording channel. Specifically we compare the noise margin of recording codes based on RLL sequences with that of uncoded sequences. Furthermore we study the noise margin of sequences generated by (free) Hamming distance increasing convolutional codes under optimal and mismatched conditions. The results of computer simulation tests are presented in sec. 6, comparing the noise margin of the detection processes when an RLL ($d = 1$) coded sequence or uncoded data are stored on the optical disc.

Though results are derived for the optical recording channel, any of the results can with small modifications be applied to other transmission systems with a dispersive character.

2. Codes based on runlength-limited sequences

We assume that binary user information with a bit rate of $1/T$ sec⁻¹ is translated into a coded channel sequence having a channel bit rate $1/T_c$. The ratio $R = T_c/T$ is the rate of the code. Binary codes such as the Miller code¹⁴), EFM³) and 3PM⁴) are examples of recording codes with applications in magnetic and optical recording. These recording codes are all based on so-called binary runlength-limited sequences (RLL sequences). A string of bits is defined to be runlength-limited if the number of consecutive like symbols, the so-called runlength, is bounded between a certain minimum and a maximum value. The theory of binary sequences with restrictions on minimum and maximum runlength goes back to Kautz¹⁵) and Tang and Bahl¹⁶). For an exhaustive treatment of this subject the reader is referred to ref. 16. Their most important results are summarized here. A dk -limited sequence simultaneously satisfies the following two conditions:

- a) d constraint — any two logical ones are separated by a run of at least d consecutive logical zeros,
- b) k constraint — the length of any run of consecutive logical zeros is at most k .

Obviously $k > d$. A sequence only satisfying the d constraint is called a d -limited sequence. A d - or dk -limited sequence is not the sequence to be recorded, we actually record a runlength-limited sequence with at least $(d + 1)$ and at most $(k + 1)$ consecutive like symbols which is obtained by integrating modulo 2 a dk -limited sequence. The maximum runlength constraint guarantees a clock pulse within some specified time interval, which is needed for the clock regeneration at the receiver. The minimum runlength constraint is imposed to control intersymbol interference and consequently has a bearing on the distortion of the transmitted signal when the channel is bandwidth-limited^{17,18}). An important parameter when dealing with runlength constraints

is the information capacity of the constrained sequence. Tang et al.¹⁶⁾ calculated the maximum (information) rate or noiseless capacity $C(d, k)$ of the dk -limited sequence. The maximum rate $C(d, k)$ of dk sequences is given by ref. 16

$$C(d, k) = {}^2\log \lambda,$$

where λ is given by the largest real root of

$$z^{k+2} - z^{k+1} - z^{k-d+1} + 1 = 0.$$

Table I shows the capacity $C(d) = C(d, \infty)$ of d -constrained sequences versus d .

TABLE I

Capacity $C(d)$ of RLL sequences and T_{\min}/T versus d .

d	$C(d) = T_c/T$	$T_{\min}/T = (d + 1)C(d)$
0	1.00	1.00
1	0.69	1.38
2	0.55	1.65
3	0.47	1.88

For an RLL sequence with a constraint on the minimum and maximum runlength, the runlengths T_i are denoted by

$$T_i = iT_c, \quad d + 1 \leq i \leq k + 1.$$

The minimum and maximum physical distances between transitions, T_{\min} and T_{\max} , respectively, are given by $T_{\min} = (d + 1)C(d, k)T$ and $T_{\max} = (k + 1)C(d, k)T$.

An RLL sequence achieving the channel capacity, a so-called maxentropic RLL sequence, has the following useful property. Assuming that the lengths of the time intervals between the transitions of the RLL sequence are statistically independent, the information rate is maximized if the probability $P(T_i)$ of the occurrence of runlength T_i is given by¹⁹⁾

$$P(T_i) = \lambda^{-i}, \quad d + 1 \leq i \leq k + 1.$$

Many efficient procedures are available for the design of runlength-limited codes with a certain d and/or k constraint^{7,15,16)}. The term 'efficient' relates here to the ratio of the actual code rate and the capacity of a dk -limited channel with the given d and k constraints. In particular the methods presented by Tang and Bahl¹⁶⁾ and Beenker and Immink²⁰⁾ are attractive in recording practice because they are based on codewords of fixed length. Franaszek⁷⁾

showed that with little hardware, practical codes can easily achieve rates of 90–95% of the capacity. We therefore use in the next sections the capacity $C(d)$ as the rate R of the runlength-limited code. This unties the analysis to follow from a specific embodiment of an RLL encoder and decoder.

3. Channel mode, MLSE detection

By virtue of the physics of the recording channel the recorded sequence I consists of binary digits $I_i \in \{-1, 1\}$ that are generated each T_c second. Assuming a linear read-out mechanism the retrieved (or received) signal $r(t)$ is of the form

$$r(t) = \sqrt{E_c} \sum_{i=-N}^N I_i g(t - i T_c) + n_w(t), \quad (1)$$

where E_c is the received energy per channel symbol, $g(t)$ is the nominal channel waveform and $n_w(t)$ is additive white Gaussian noise with two-sided spectral density $N_0/2$.

The recorded sequence is taken to be of arbitrary odd length $2N + 1$. The coded channel sequence (or codeword) $I = (I_{-N}, \dots, I_N)$ is a member of a predefined set of codewords. We have normalized $g(t)$ in such a way that

$$\int_{-\infty}^{\infty} g^2(t) dt = 1.$$

Assuming that the signal is matched filtered with a filter having a response $\sqrt{E_c} g(-t)$ and subsequently sampled at $t = k T_c$, the equivalent vector channel is

$$\begin{aligned} r_k &= E_c \left\{ I_k + \sum_{\substack{i=-N \\ i \neq k}}^N I_i g_{k-i} \right\} + n_k \\ &= E_c \{ I_k + q_k \} + n_k, \quad -N \leq k \leq N, \end{aligned} \quad (2)$$

where n_k are coloured noise samples and q_k , given by

$$q_k = \sum_{i=-N, i \neq k}^N I_i g_{k-i},$$

is the intersymbol interference (ISI) at the sampling moments $t = k T_c$. The q_i are given by

$$g_i = \int_{-\infty}^{\infty} g(t) g(t + i T_c) dt.$$

Note that $g_i = g_{-i}$ and $g_0 = 1$. The statistics of q_k are directly related to the channel sequence $\{I_i\}$ and the waveform $g(t)$ of the channel.

The ISI can therefore be affected in two ways:

- 1) pulse shaping at the transmitter and/or receiver using filters, and
- 2) manipulation of the code structure and hence of the correlation in the channel sequence.

The usual approach to combat ISI has focused on the shaping of $g(t)$ for zero interference. Hard-limiting channels, such as encountered in optical or magnetic recording systems, only accept two pulse shapes, a positive or a negative 'full- T ' pulse, so that ISI can only be affected by the code structure or the receiver filter.

From eq. (2) we find that the received $(2N + 1)$ vector $r = (r_{-N}, \dots, r_N)$ can be written as

$$r = E_c M I + n,$$

where M is the sampled auto-correlation matrix of $g(t)$ with entries given by

$$M_{i,j} = g_{i-j}$$

and n is the noise vector $n = (n_{-N}, \dots, n_N)$, where n_i are zero mean, Gaussian distributed samples with auto-correlation function $E(n_j n_{j+k}) = E_c N_0 g_k/2$.

The minimum sequence error probability is obtained by choosing that message I' from the set of allowed sequences which maximizes $\Pr(I' | r)$.

Assuming that the codewords are equally likely, we find that $\Pr(I' | r)$ is proportional to

$$\exp \left\{ - \frac{1}{2E_c} (E_c M I' - r)^T M^{-1} (E_c M I' - r) \right\},$$

where ' T ' stands for transpose¹⁰). Taking logs and keeping only those terms which depend on I' , we need to choose that sequence I' which maximizes $\mu(I')$ given by

$$\mu(I') = 2 r^T I' - E_c I'^T M I'. \quad (3)$$

We notice that in a straightforward MLSE implementation the number of computations required grows exponentially with the length $2N + 1$ of the recorded sequence I . To limit the dimensionality we assume that the response of the channel can be ignored outside a certain time span, i.e. $g_i = 0$, $|i| > L$, where L is the memory of the channel.

The significance of the dynamic programming approach (Viterbi algorithm) is that the number of computations required for MLSE detection grows only linearly with the length of the sequence and exponentially with the memory L of the channel. In this paper we will not dwell on the details of the dynamic programming method; the reader is referred to refs 11 and 21.

4. Probability of bit error of MLSE detector

An expression for the probability of occurrence of an error event when complete knowledge of the channel is available was found by Forney⁹). Here we derive a generalized expression of the error event probability when the receiver's a priori assumptions regarding the channel characteristics are wrong. In the analysis of the error performance of MLSE a key role is played by the error sequence between two allowed sequences.

Let the sequences I and $I' = (I'_{-N}, \dots, I'_N)$ be two elements from the code-word set. The error sequence $e = (e_{-N}, \dots, e_N)$ is defined by

$$e_i = \frac{1}{2}(I'_i - I_i).$$

Obviously $e_i \in \{-1, 0, 1\}$.

We assume that the received channel waveform is $g'(t)$ instead of $g(t)$. The matched-filter and the detection algorithm are based on the (wrong) assumption that the channel waveform is $g(t)$. The received vector r , assuming the message I was recorded, is given by

$$r = E_c M' I + n, \quad (4)$$

where the cross-correlation matrix M' has entries given by

$$M'_{i,j} = \int_{-\infty}^{\infty} g(t + iT_c) g'(t + jT_c) dt. \quad (5)$$

In ideal circumstances $g(t) = g'(t)$. Note that in general

$$\int_{-\infty}^{\infty} g(t) g'(t) dt \neq 1.$$

In order to determine the performance of the detector, observe that if I is the actual recorded message, a message $I' \neq I$ is chosen if

$$\mu(I) - \mu(I') < 0.$$

From eqs (3) and (4) it follows that

$$\begin{aligned} \mu(I) &= 2r^T I - E_c I^T M I \\ &= 2E_c I^T M' I + 2n^T I - E_c I^T M I. \end{aligned}$$

In the same way we find

$$\mu(I') = 2E_c I'^T M' I' + 2n^T I' - E_c I'^T M I'.$$

After some algebra we find that $\mu(I) - \mu(I')$ is given by (using the fact that M is symmetric)

$$\mu(I) - \mu(I') = 4E_c \left\{ e^T M e + e^T (M - M') I + n^T \frac{e}{E_c} \right\}.$$

Obviously $\frac{1}{4E_c} \{\mu(I) - \mu(I')\}$ is a Gaussian random variable with a mean a given by

$$a = e^T M e + e^T (M - M') I \tag{6}$$

and variance N'

$$N' = \frac{b N_0}{2E_c}, \tag{7}$$

where b is given by

$$\begin{aligned} b &= \sum_i \sum_j e_i e_j \int_{-\infty}^{\infty} g(t - iT_c) g(t - jT_c) dt \\ &= e^T M e. \end{aligned}$$

Suppose now that the number of codewords is restricted to two. In this case the problem reduces to a simple two-hypothesis detection problem. The probability $\Pr(I, I')$ of choosing the incorrect sequence I' instead of the correct sequence I is

$$\begin{aligned} \Pr(I, I') &= \Pr\{\mu(I) - \mu(I') < 0\} \\ &= Q\left(\frac{a}{\sqrt{N'}}\right) \\ &= Q\left(d_{\min} \sqrt{\frac{2E_c}{N_0}}\right). \end{aligned}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{t^2}{2}\right) dt$$

and d_{\min} is defined by

$$d_{\min} = \frac{a}{\sqrt{b}}.$$

When the number of codewords is larger than two, Forney⁹) has shown that for moderate and high signal-to-noise ratio the error probability $\Pr(e)$ can be lower bounded in the following way:

$$\Pr(e) > K_1 Q\left(d_{\min} \sqrt{\frac{2E_c}{N_0}}\right),$$

where K_1 is a small constant and

$$d_{\min} = \min_{I, I' \neq I} \frac{a}{\sqrt{b}} \tag{8}$$

is the minimum dissimilarity between any pair of sequences in the codeword set. When $d_{\min} < 0$, the Viterbi detector has completely broken down and it systematically makes errors. When the receiver has a complete knowledge of the channel, i.e. $M = M'$, it can readily be verified that eq. (8) reduces to

$$d_{\min}^2 = \min_{e, e \neq 0} a = \min_{e, e \neq 0} e^T M e, \quad (9)$$

where d_{\min} is now called the minimum Euclidean distance. It is not difficult to show that the minimum Euclidean distance is non-negative. A natural function to optimize is the error probability at a given signal-to-noise ratio. From a computational point of view this is quite unrealistic. Therefore we choose the minimum dissimilarity as a parameter to be used as the figure of merit of the code performance. We call $\text{SNR} = 2d_{\min}^2 E_c / N_0$ the effective signal-to-noise ratio. Then

$$\Pr(e) > K_1 Q(\sqrt{\text{SNR}}).$$

It is a well-known result¹⁰⁾ that if two recorded sequences are to be easily distinguished at the output of a linear channel it is necessary that the difference waveform $e(t)$ propagates well on the channel. In other words the spectral properties of $e(t)$ rather than the spectral properties of the transmitted sequences determine the reliability of the retrieved data. If, however, the receiver is ignorant of the actual channel impulse response, the error probability is governed by eq. (8). This shows that the characteristics both of the transmitted sequence and of the difference waveform are of importance. This last assumption, that the receiver is mismatched to the channel, is in actual recording practice more the rule than the exceptionally simple 'full-knowledge' assumption. As we shall see later we have to be rather careful in the interpretation of the code performance predicted by eq. (9) if we are dealing with channel/receiver mismatch.

In order to evaluate the code performance it is worth defining the coding gain G :

$$G = 10 \log \frac{d_{\min}^2 E_c \text{ (coded)}}{E_c \text{ (uncoded)}}.$$

The coding gain G expresses the quotient of effective signal-to-noise ratio in the coded case and the signal-to-noise in the uncoded ISI-free case. For comparison purposes, the effective signal-to-noise ratio of an MLSE receiver in the absence of intersymbol interference is

$$\text{SNR} = \frac{2R w_H E}{N_0}, \quad (10)$$

where the minimum Hamming distance

$$w_H = \min_e \sum_{i=-N}^N e_i^2$$

denotes the minimum number of symbol errors between any two allowed data sequences and E is the received energy per user symbol, $E = E_c/R$. In the ISI-free case the coding gain is $10 \log R w_H$. Efficient and systematic procedures are available to design codes that achieve coding gain on the ISI-free channel¹¹⁾. The complex nature of the Euclidean distance measure when ISI is involved has hindered the design of good codes. The calculation of the error probability using eqs (6), (7) and (8) is tedious: the minimum distance d_{\min} has to be calculated between any source sequence I and any allowed sequence I' . The execution time can be enormous, even for simple codes. In the ideal case, $g(t) = g'(t)$, the minimum distance only depends on the difference of the sequences I and I' and not on the actual recorded sequence itself. For the ideal case Anderson et al.²²⁾ have computed the minimum distance and corresponding error sequence as a function of the channel memory L . When $g(t) \neq g'(t)$ the only method that seems to be available is exhaustive search using eq. (8) as a criterion.

It is obvious from table I that if RLL sequences are used on ISI-free channels that RLL sequences have a coding loss $10 \log C(d)$ (the minimum Hamming distance being unity). Perhaps this explains why RLL-based codes have not found widespread interest in the coding literature and why their performance has occasionally been misunderstood. The next sections show that on a recording channel with severe ISI and unknown gain or bandwidth variations, RLL sequences are extremely valuable.

4.1. Probability of error during gain mismatch

In order to get some insight into the performance of the detection system during gain mismatch we assume that the channel has a gain K while the receiver erroneously assumes that the gain is unity. In other words, $g'(t) = K g(t)$, $K > 0$. From eq. (6) we derive

$$\begin{aligned} a &= e^T M e + e^T (M - M') I \\ &= e^T M e + (1 - K) e^T M I. \end{aligned} \quad (11)$$

In this section we assume that the error performance is dominated by error sequences of weight one. This assumption is rather restrictive, but it gives us the opportunity to use an analytical approach. Further we assume, without

loss of generality, that this error occurs at $t = j T_c$ or $I'_j = -I_j$, so that $e_j = -I_j$ and $e_l = 0$, otherwise. Substitution in eq. (11) yields

$$\begin{aligned} a &= K + (K - 1) I_j \sum_{\substack{i=-N \\ i \neq j}}^N I_i g_{j-i} \\ &= K + (K - 1) I_j q_j, \end{aligned}$$

where q_j is the ISI at $t = j T_c$ (see eq. (2)). We further assume for reasons of normalization that the noise level is scaled with the same factor K , so that the reduction of d_{\min} and consequently the reduction in error performance is due only to the channel/receiver mismatch. We therefore define $b' = K^2 b$, so that for uncoded data we simply find

$$\begin{aligned} d_{\min} &= \min_{e, I} \frac{a}{\sqrt{b'}} \\ &= 1 - \left| 1 - \frac{1}{K} \right| \max_j (q_j) \\ &= 1 - 2 \left| 1 - \frac{1}{K} \right| \sum_{i=1}^L |g_i|. \end{aligned} \quad (12)$$

We observe from this simple relation that d_{\min} depends on two parameters: the gain mismatch and the intersymbol interference. Eq. (12) reveals that any tolerance of the actual channel gain imposes a penalty on the minimum distance which is proportional to the worst-case ISI. Eq. (12) is reminiscent of peak-eye-closure in the case of bit-by-bit detection using a linear receiver. Specifically, if $K = \frac{1}{2}$ we conclude that the error probability is similar to the error performance of the equivalent linear unequalized receiver. In retrospect it is quite surprising that the 'eye-closure' measure enters our analysis through the back-door.

Eq. (12) holds for the situation that uncoded data is recorded and the assumption that only one error will occur. For coded data or the general case of more than one error we could not find such a simple analytical result.

5. Applications to an optical read-out channel

This section illustrates the use of the theory developed in previous sections in the specific application of optical recording. The impulse response $h(t)$ of the optical read-out channel is often modelled according to a Gaussian shape¹), or

$$h(t) = \frac{1}{t_0 \sqrt{\pi}} \exp\left(-\frac{t^2}{t_0^2}\right), \quad (13)$$

where t_0 is the $1/e$ width of the read-out spot. The pits and lands in the optical disc are recorded as full- T pulses $p_{T_c}(t)$ with constant amplitude $\sqrt{\frac{E}{T}}$, or

$$p_{T_c}(t) = \sqrt{\frac{E}{T}}, \quad |t| < \frac{1}{2} T_c, \quad (14)$$

$$= 0, \text{ otherwise.}$$

E is the transmitted energy per user (or information) bit. The channel waveform $g(t)$ of the combined disc recording plus read-out is given by

$$\begin{aligned} \sqrt{E_c} g(t) &= (p_{T_c} * h)(t) \\ &= \sqrt{\frac{E}{4T}} \left[\operatorname{erf} \left(\frac{t + \frac{1}{2} T_c}{t_0} \right) - \operatorname{erf} \left(\frac{t - \frac{1}{2} T_c}{t_0} \right) \right], \end{aligned}$$

where '*' stands for convolution and the error function $\operatorname{erf}(\cdot)$ is defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt.$$

The auto-correlation coefficients g_i are found by means of a straightforward analysis:

$$\begin{aligned} E_c g_i &= \int_{-\infty}^{\infty} (h * p_{T_c})(t) (h * p_{T_c})(t + i T_c) dt = \\ &= \int_{-\infty}^{\infty} (p_{T_c} * p_{T_c})(t) (h * h)(t + i T_c) dt = \\ &= ER \left\{ 2Y \left(\frac{S}{R}, i \right) - Y \left(\frac{S}{R}, i - 1 \right) - Y \left(\frac{S}{R}, i + 1 \right) \right\}, \quad (15) \end{aligned}$$

where we used the fact that both $p_{T_c}(t)$ and $h(t)$ are symmetric. The normalized (information) density S is defined according to

$$S = \frac{t_0}{T} \quad (16)$$

and $Y(x, i)$ is found after an evaluation of eq. (15):

$$Y(x, i) = \frac{-i}{2} \operatorname{erf} \left(\frac{i}{\sqrt{2} x} \right) - \frac{x}{\sqrt{2\pi}} \exp \left(-\frac{i^2}{2x^2} \right). \quad (17)$$

The energy per channel symbol, E_c , is given by

$$E_c = \int_{-\infty}^{\infty} (p_{T_c} * h)^2(t) dt = ER \left\{ \operatorname{erf} \left(\frac{R}{\sqrt{2}S} \right) - \frac{2S}{R\sqrt{2\pi}} \left(-\exp \left[-\frac{R^2}{2S^2} \right] \right) \right\}. \quad (18)$$

The energy per channel symbol E_c can at high information densities be approximated by

$$E_c \cong \frac{1}{\sqrt{2\pi}} R^2 E \frac{1}{S}. \quad (19)$$

Eq. (19) reveals that at high information densities the penalty on a reduction of the code rate R is proportional to R^2 . It is quite clear that making profit out of coding is now much more difficult than on an ISI-free channel, where the penalty is proportional to R .

Fig. 1 shows the reduction of the received energy E_c/E as a function of the normalized density S (eq. (18)). The second curve in the figure shows the minimum Euclidean distance $d_{\min}^2 E_c/E$ between uncoded sequences. In the region

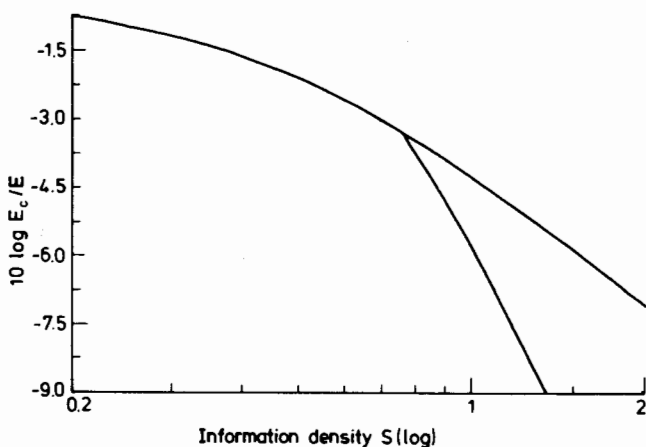


Fig. 1. Received energy E_c/E (upper curve) and minimum Euclidean distance $d_{\min}^2 E_c/E$ of uncoded sequences (lower curve) versus the normalized information density S .

$S < 0.75$ both curves are the same, but if $S > 0.75$ the ISI is so severe that d_{\min} becomes smaller than unity. In this region the error performance is dominated by twin errors of the form $e = (0, \dots, -1, 1, \dots, 0)$ and its inverse; the minimum distance is given by $2(1 - g_1)E_c < E_c$.

Fig. 2 shows the coding gain (or better stated, coding loss) of RLL sequences, $d = 1, 2, 3$ and uncoded ($d = 0$) sequences. This figure reveals that in the depicted region RLL-based codes show a coding loss with respect to uncoded

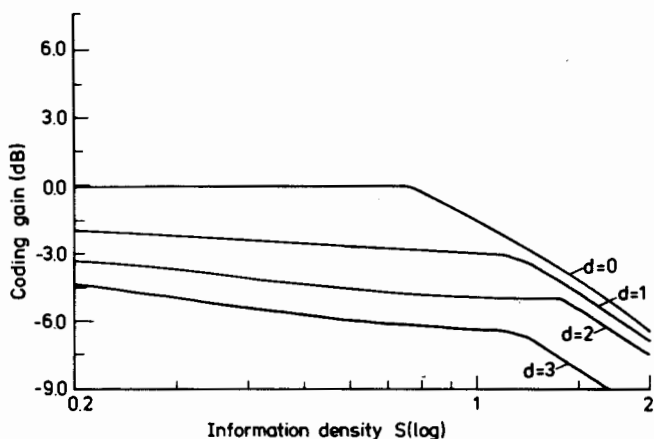


Fig. 2. Coding gain of d -constrained sequences as a function of the normalized information density S with d as a parameter.

information. As a reference, note that the Compact Disc has ample signal-to-noise ratio: $E/N_0 = 30$ dB. The normalized information density using simple bit-by-bit detection is approximately $S = 0.9$.

Fig. 3 shows the coding gain versus the density of two famous Hamming-distance-increasing codes:

a) a convolutional code, $R = \frac{1}{2}$, minimum Hamming (free) distance 5, with generator polynomials '111' and '101';

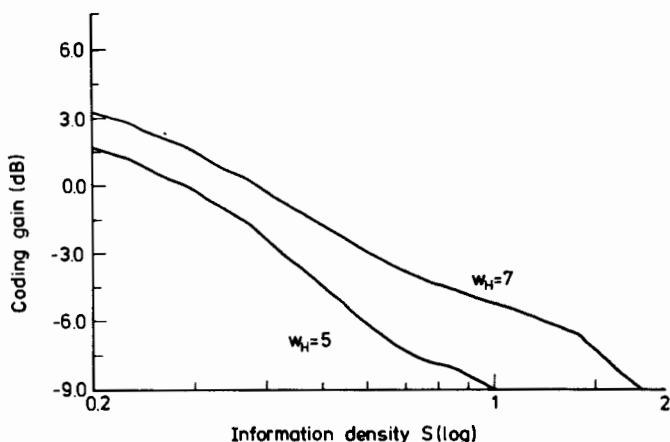


Fig. 3. Coding gain of rate- $\frac{1}{2}$ convolutional codes with minimum (free) Hamming distance 5 and 7, respectively.

b) a convolutional code, $R = \frac{1}{2}$, minimum Hamming (free) distance 7, with generator polynomials '11101' and '10011'.

The selection of these codes was based upon:

- (1) their optimality on ISI-free channels and
- (2) the fact that they achieve the minimum free Hamming distance with the minimum value of the encoder constraint length²⁴).

This last property has a beneficial influence on the complexity of the detector. Earlier results on the performance of Hamming-distance-increasing codes in the presence of ISI were given by Acampora²⁵) and Divsalar²⁶). All the results depicted in the figure were obtained with an exhaustive computer search using eqs (6) and (8). The Hamming-distance-increasing codes have a coding gain $R_{WH}(\frac{5}{2} = 3.9$ dB and $\frac{7}{2} = 5.4$ dB, respectively) in the range of (very) small densities.

From the figure we conclude that the codes are apparently not adapted to channels with severe ISI (after all this is no surprise because they were not designed to perform on ISI channels). There is a coding loss if the information density is increased above $S = 0.28$ and $S = 0.37$, respectively. Codes a) and b) show at a density $S = 1$ a coding loss of approximately 7 and 3 dB, respectively. The performance of the convolutional code b) can be improved at low densities by approximately 0.5 dB if the direction of switching at the encoder is reversed. This code with swapped generator polynomials exhibits, however, a 4 dB loss with respect to code b) at the higher densities. This phenomenon was already observed by Divsalar²³). He noted that a reversal of the direction of switching the encoder output is a simple demonstration of the fact that the well-known optimum codes for the linear AWGN channel are no longer necessarily optimum for a bandwidth-constrained channel. From figs 2 and 3 we conclude that the best choice for a code to be used at high densities ($S > 1$) is the ($d = 0$) code, i.e. no coding at all. The situation changes, however, if we assume that the receiver's knowledge is incorrect.

Fig. 4 shows the coding gain of selected codes versus the information density S , but now in the case that a 20% reduction of the channel gain is assumed. This value of the gain mismatch is chosen quite arbitrarily, but is certainly a value that can be met in average recording practice; peak values of the actual gain variations are much higher. We also plotted the loss in performance resulting from the use of uncoded sequences. Remind that the coding loss or gain as plotted in the figure relates to the performance of MLSE detection of uncoded sequences under optimum conditions, the so-called matched-filter or one-shot bound. For normalization purposes we reduced the power of the additive noise as discussed in sec. 4.1, so that the results shown are due only to the channel/receiver mismatch. We observe that the performance re-

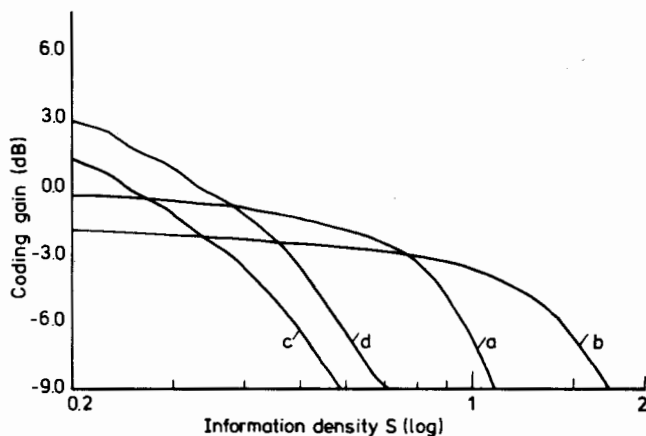


Fig. 4. Coding gain of selected codes versus the information density when a gain mismatch of 20% is assumed, *a*) uncoded sequences, *b*) RLL sequence $d = 1$, *c*) conv. code with minimum (free) Hamming distance 5 and *d*) idem with distance 7.

sulting from the use of uncoded sequences deteriorates when there is gain mismatch. At high relative information densities the loss with respect to the ideal case is dramatic. It is quite apparent that the Viterbi detector so heavily leans on the assumed correct knowledge of the channel impulse response that as soon as this knowledge is actually incorrect, it dramatically deteriorates in performance. Now the difference between uncoded and RLL encoded information is changed with respect to the ideal case. In the high density range the detection of an RLL sequence outperforms the detection of an uncoded sequence. We further note that in the low density region the performance of the two convolutional codes is quite immune to gain mismatch, but at higher densities it rapidly loses performance. When coding gain is our design criterion we find that an RLL code is preferable in the range $S > 0.6$.

Fig. 5 shows the effect of an unknown 20% bandwidth reduction, i.e. $t_o = 1.2$, while the detector assumes $t_o = 1$. This figure leads qualitatively to conclusions similar to those derived from fig. 4. More insight can be gained by rearranging the previous data.

Fig. 6 shows the coding gain of various codes as a function of the channel gain at an information density $S = 1$ (the receiver assumes the channel gain to be unity). The figure clearly reveals that the larger the d constraint the more loss is found in the optimum situation. However, a larger d constraint makes the system less vulnerable to unknown channel gain variations. Qualitatively we obtain the same results when the coding gain is calculated as a function of

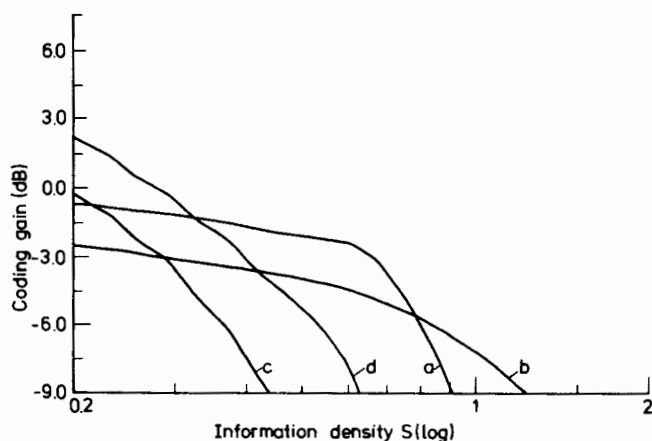


Fig. 5. Coding gain of selected codes versus the information density when a bandwidth mismatch of 20% is assumed, *a*) uncoded sequences, *b*) RLL sequence $d = 1$, *c*) conv. code with minimum (free) Hamming distance 5 and *d*) idem with distance 7.

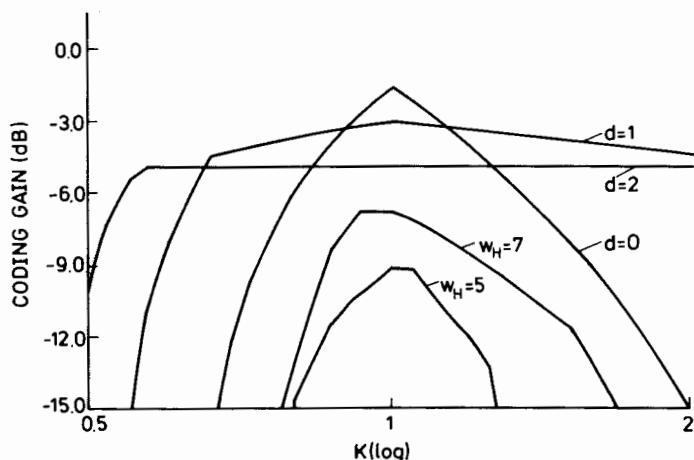


Fig. 6. Coding gain of selected codes as a function of the channel gain K at an information density $S = 1$. The receiver assumes the channel gain to be unity.

the bandwidth mismatch. We conclude that we have to sacrifice a certain noise margin, resulting in a larger error probability, to obtain a more robust system.

6. Simulation results

In order to verify the analytical results of the preceding section a runlength-limited ($d = 1$) code was tested by a computer program which measures the error probability using the Viterbi algorithm for detection. The results are compared with the Viterbi detection of uncoded data. The statistical distribu-

tion of the runlengths of the RLL sequence was chosen to be exponential according to the maxentropic property as described in sec. 2. In order to limit computer time and memory the ISI coefficients were ignored when smaller than 0.01. Tests showed that an increase of the accuracy did not lead to significantly different results. The simulations were stopped when more than 300 errors were counted. Viterbi detection of the ($d = 1$) runlength-limited sequence was established by a simple modification of the MLSE detector of uncoded data as described by Forney⁹). The states in the trellis diagram that do not satisfy the d constraint are simply discarded.

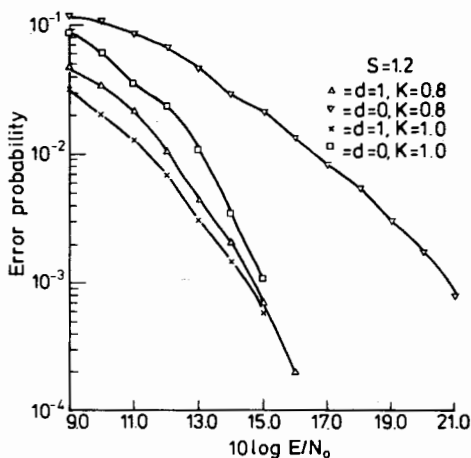


Fig. 7. Error probability versus signal-to-noise ratio E/N_0 at an information density $S = 1.2$.

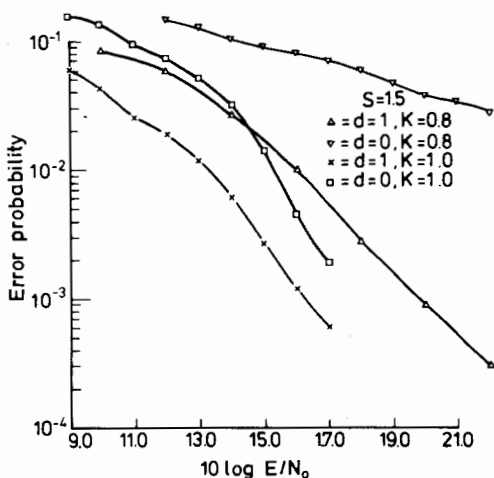


Fig. 8. Error probability versus signal-to-noise ratio E/N_0 at an information density $S = 1.5$.

Fig. 7 shows the error probability versus signal-to-noise ratio E/N_0 an information density $S = 1.2$ for the coded and the uncoded case. In addition the figure shows the effect of gain reduction of 20%, which is unknown to the receiver.

Fig. 8 shows the performance of the same codes but now at an increased information density $S = 1.5$. The detection of uncoded sequences leads to a systematic error probability of 0.013 in the absence of noise when a 20% gain reduction is assumed. The ($d = 1$) RLL code is still going strong at a 3 dB reduced noise margin with respect to the optimum situation.

7. Concluding remarks

In this paper we have discussed the error probability performance resulting from the use of candidate recording codes based on runlength-limited or Hamming-distance-increasing sequences. The detection method we employed was the celebrated maximum likelihood sequence estimation. Our results showed that under optimum conditions, in which the receiver has full knowledge of the channel characteristics, the RLL-based codes exhibit an actual coding loss. In other words, when RLL sequences are employed on a recording channel this recording channel definitely needs a better signal-to-noise ratio than when no coding at all is employed. The application of (Hamming) distance improving codes permits channels with an even worse signal-to-noise ratio. Unfortunately, small deviations from the ideal situation (as an example we studied the effect of gain and bandwidth mismatch) dramatically deteriorate the performance of the MSLE in conjunction with this class of codes. At this point it seems appropriate to quote Tukey²⁷), who addressed the IT community in 1966: 'the most robust method, the one least affected by small deviations from idealness of circumstances, is usually that one that performs worst in the utopian situation. For robustness one must pay, though the insurance premium is often surprisingly small'. Our study showed that the coding loss of RLL sequences is easily regained when ISI and channel/receiver mismatch are present.

8. Conclusions

At high normalized information densities of an optical disc, i.e. when the ISI becomes quite severe, maximum-likelihood detection of uncoded information is sensitive to unknown tolerances of the channel impulse response. Convolutional codes designed to increase the (free) Hamming distance were found not to perform satisfactorily under severe ISI conditions. Furthermore it was found that the performance of these codes rapidly deteriorates in the case of erroneous receiver knowledge about the actual channel characteristics. The

combination of an RLL-sequence-based recording code plus maximum-likelihood detection showed a relative immunity to ISI and unknown impulse response variations. Computer simulations based on the Viterbi detection algorithm verified our analytical results.

REFERENCES

- 1) G. Bouwhuis, J. Braat, A. Huijser, J. Pasman, G. van Rosmalen and K. Schouhamer Immink, Principles of optical disc systems, Adam Hilger Ltd, Bristol 1985.
- 2) H. Kobayashi, IEEE Trans. Commun. Techn. **COM-19**, pp. 1087-1100 (1971).
- 3) J. P. J. Heemskerck and K. A. Schouhamer Immink, Philips Tech. Rev. **40**, pp. 157-164 (1982).
- 4) G. V. Jacoby, IEEE Trans. Magn. **MAG-13**, pp. 1202-1204 (1977).
- 5) G. V. Jacoby and R. Kost, IEEE Trans. Magn. **MAG-20**, pp. 709-714 (1984).
- 6) J. S. Eggenberger and P. Hodges, 'Sequential encoding and decoding of variable length, fixed rate data codes', US Patent 4,115,768, 1978.
- 7) P. A. Franzaszek, IBM J. Res. Develop. **14**, pp. 376-383 (1970).
- 8) P. H. Siegel, IEEE Trans. Magn. **MAG-18**, pp. 1250-1252 (1982).
- 9) G. D. Forney, IEEE Trans. Inform. Theory **IT-21**, pp. 363-378 (1972).
- 10) G. Ungerboeck, IEEE Trans. Commun. **COM-22**, pp. 624-636 (1974).
- 11) A. J. Viterbi and J. K. Omura, Principles of digital Communication and Coding. McGraw-Hill, New York 1979.
- 12) R. Wood, Fourth Intern. Conf. on Video and Data Recording, Southampton, pp. 333-343, (1982).
- 13) H. Burkhardt, IEEE Trans. Magn. **MAG-17**, pp. 3337-3339 (1981).
- 14) J. C. Mallinson and J. W. Miller, Radio and Elec. Eng. **47**, pp. 172-176 (1970).
- 15) W. H. Kautz, IEEE Trans. Inform. Theory **IT-11**, pp. 284-292 (1965).
- 16) D. T. Tang and L. R. Bahl, Inform. Contr. **17**, pp. 436-461 (1970).
- 17) M. G. Pelchat and J. M. Geist, IEEE Trans. Commun. **COM-23**, pp. 878-883 (1975). Correction **COM-24**, p. 479 (1976).
- 18) P. D. Shaft, IEEE Trans. Commun. **COM-21**, pp. 687-695 (1973).
- 19) K. A. Schouhamer Immink, Philips J. Res. **38**, pp. 138-149 (1983).
- 20) G. F. M. Beenker and K. A. Schouhamer Immink, IEEE Trans. Inform. Theory **IT-29**, pp. 751-754 (1983).
- 21) J. G. Proakis, Digital Communications, McGraw-Hill, New York 1983.
- 22) R. R. Anderson and G. J. Foschini, IEEE Trans. Inform. Theory **IT-21**, pp. 544-551 (1975).
- 23) D. Divsalar and M. K. Simon, Proc. IEEE ICC 1983, pp. 908-914 (1983).
- 24) G. C. Clark and J. B. Cain, Error-correction coding for digital communications, New York; Plenum Press 1981.
- 25) A. S. Acampora, IEEE Trans. Commun. **COM-26**, pp. 766-776 (1978).
- 26) D. Divsalar, PhD thesis, University of California, Los Angeles (1978).
- 27) J. W. Tukey, IEEE Trans. Inform. Theory **IT-12**, pp. 87-92 (1966).