

# Marker codes for channels with insertions and deletions

Edward A. Rutzer

Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK  
Phone: +44 (0) 1223 337238, Fax: +44 (0) 1223 354599  
E-mail: ear23@mrao.cam.ac.uk

**Abstract:** *At low noise levels marker codes are shown to outperform watermark codes. Full iterative decoding enhances performance to close to the capacity bounds.*

**Keywords:** Insertion-deletion channel, marker codes, LDPC codes, turbo synchronization, watermark codes

## 1. INTRODUCTION

Examples of channels with synchronization errors include:

**Serial line** The clock speed of the transmitter may not be accurately known (for instance due to temperature variations in the clock) so the time of arrival of each transmitted bit is not known.

**Hard disc** Variations in the rotation speed (for instance due to mechanical vibrations or shock) mean the position of the head relative to the platter may be uncertain.

**DAT tape** Tape stretch leads to problems similar to those suffered by a hard disc.

In this paper the channel will be modelled by random uncorrelated insertion or deletion events at unknown positions. A flowchart of the channel model is shown in figure 1. The capacity for channels of this kind is not known exactly. A capacity lower bound from [11] and an upper bound from [9] have been used in this paper.

Marker codes [5] were originally designed to be able to deal with single insertion or deletion error events. The bit stream to be transmitted has a regular marker (or header) inserted in it. For example the marker '001' may be inserted between every 4 data bits:

01101100101010  $\mapsto$  01100011100001101000110

The decoder can then look out for the markers and use any shift in their position to deduce bit loss or gain. Errors in the matched sequence can then be corrected with a conventional code. With advances in computer power probabilistic sequence matching

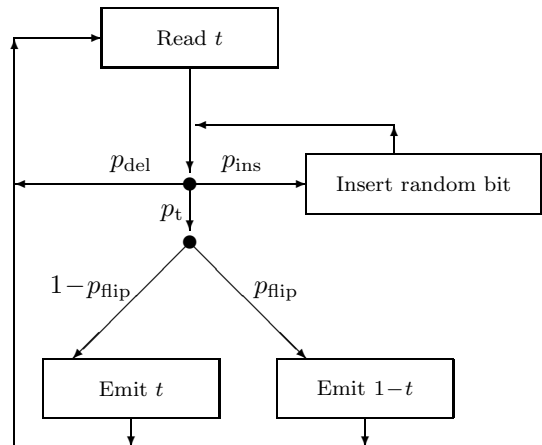


Figure 1: Flow chart describing the insertion-deletion channel with insertion probability  $p_{\text{ins}}$ , deletion probability  $p_{\text{del}}$ , transmission probability  $p_t = 1 - p_{\text{ins}} - p_{\text{del}}$ , and substitution probability  $p_{\text{flip}}$ . For simplicity  $p_{\text{flip}} = 0$  for the simulations in this paper.

[7] can be carried out. The coding system is shown in figure 2.

Watermark codes [3] are a similar scheme, but rather than having bursts of synchronization information and bursts of data, the information is distributed uniformly. To encode, the data bits are uniformly sparsified and then combined with a watermark sequence. To decode, probabilistic resynchronization can be carried out with the watermark sequence.

In this paper the results in [7] on the properties of the markers alone are extended by comparing complete marker code based systems with watermark codes. The benefit of iterative probabilistic resynchronization is also studied.

## 2. COMPARISON WITH WATERMARK CODES

In [7] it was shown that for a code system like figure 2 the capacity of the effective channel seen by an outer code is higher for marker codes than for watermark codes at low noise levels. The highest rate results for watermark code published [3] are at

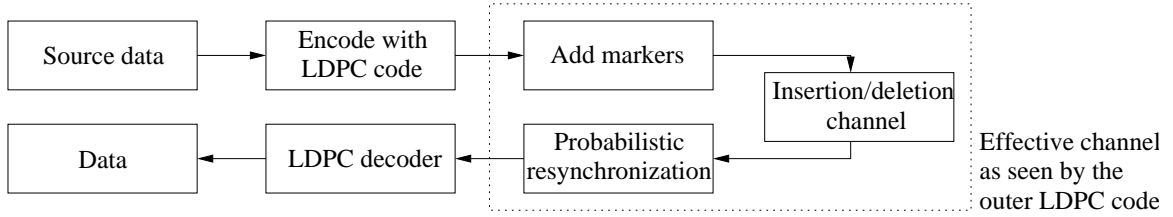


Figure 2: The code system used, with serial decoding illustrated

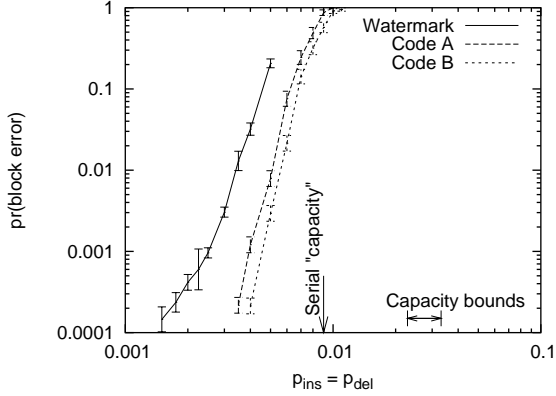
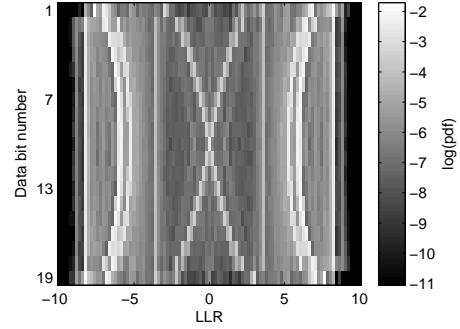


Figure 3:  $R = 0.71$ ,  $N = 4995$  codes over an insertion/deletion channel with serial decoding. The watermark code result is from [3]. Marker code A is with a LDPC code with with weight 2 and weight 3 columns. Marker code B has weight 10 columns in addition, table 1.

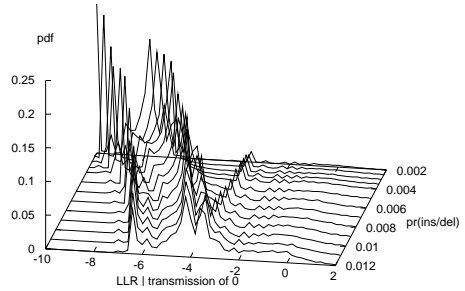
$R = 0.71$  with a block size of 4995. To match this a marker code of the same overall rate and block size was created, code A. The inner code was chosen to be good at a noise level  $p_{\text{ins}}=p_{\text{del}}=0.005$  using the effective capacity technique outlined in [7]. The outer code was a low-density parity-check (LDPC) code with weight 2 and 3 columns. The code parameters are shown in table 1. Decoding was as figure 2.

A comparison of the watermark and marker codes is shown in figure 3. The marker code outperforms the watermark code, despite the watermark code having been constructed with a LDPC code defined over a larger field (GF(16)) than the binary LDPC codes considered in this paper.

The marker code is not close to the channel capacity bounds. To try to approach the bounds, further optimization of the outer LDPC code was studied. The threshold of an infinite loop-free LDPC code as a function of column weight distribution was evaluated using a Monte Carlo approach [2]. In this procedure it is necessary to know the distribution of messages sent by the channel. Statistics of the messages received by the outer code were collected, figure 4. The distribution is almost symmetrical if given a data se-



(a) At a noise level of  $p_{\text{ins}} = p_{\text{del}} = 0.5\%$ . The vertical axis ranges over each data bit between two markers



(b) Marginalised over all data bits whose transmitted bit was '0'

Figure 4: Probability density functions of messages received by an outer code where the inner code has a marker sequence of '01' between every 19 data bits. The log-likelihood ratio (LLR) is  $\frac{\text{Pr}(\text{bit}=1)}{\text{Pr}(\text{bit}=0)}$ .

quence of i.i.d. bits. The simulation was done with an all zero-transmission but with noise statistics as if they were i.i.d. data bits. Despite the non-Gaussian form of the messages, degree sequences could not be found that significantly outperformed good degree sequences obtained from optimizations on the Gaussian channel [1]. This optimization was carried out with the message histograms marginalised over all bits, figure 4(b). Perhaps if a different degree sequence for each bit between markers is allowed a further gain could be achieved, as figure 4(a) shows bits received near markers are more reliable than bits

far away from markers.

Simulations using a good degree sequence from the Gaussian channel, code B, are also shown in figure 3. The waterfall region is not much closer to the insertion-deletion capacity, but it is close to the serial “capacity” (defined from the effective capacity of the channel seen by the outer code [7]).

### 3. A COMPLETE ITERATIVE SYSTEM

It has been shown [6], [10] that extending loopy belief propagation as used to decode LDPC codes to include estimating channel state can be beneficial. For the insertion-deletion resynchronization phase it is expected that if we give the resynchronization algorithm more information about the likely transmission the accuracy of the resynchronization will increase.

Watermark and regular marker codes were empirically discovered to have similar performance near  $R = 0.5$  [7]. Therefore  $R = 0.5$  codes with a block-size of 4000 (to match [3]) were studied.

For benchmarks, simulations were done of marker codes with the serial decoding algorithm from section 2. The first simulation (code C) was with identical markers of length 3 and a LDPC outer code with weight 2 and weight 3 columns. Figure 5 shows that the watermark code has better error floor behaviour. With identical markers catastrophic error propagation is possible as the resynchronization can be shifted by a multiple of a marker interval. Code D was similar to code C, however the markers were pseudo-randomly chosen from a set of two different markers. This improved the error floor and led to better performance than the watermark code. An irregular LDPC outer code (chosen to be good on the Gaussian channel) was also tested and a small further improvement found, code E.

Simulations were carried out with codes D and E using iterative resynchronization. The algorithm is similar to the serial resynchronization algorithm but with extrinsic information from the LDPC decoder fed back into the resynchronization stage (altering  $\Pr(\mathbf{r}|\text{path})$  in [7, eq 2]). The resynchronization was carried out every five iterations of the LDPC code to increase the decoding speed as a LDPC code iteration is faster than probabilistic resynchronization.

The simulation results are also shown on figure 5. The figure shows that the iterative approach significantly outperforms the serial approach and that the waterfall regions can be close to the channel capacity. It is worth noting that the ranking of the codes is reversed. The choice of a code to be used should be made when the decoding algorithm is known.

To look into this swap in performance, extrinsic information transfer (EXIT) charts [8] for the systems were studied at a noise level of  $p_{\text{ins}}=p_{\text{del}}=0.04$ .

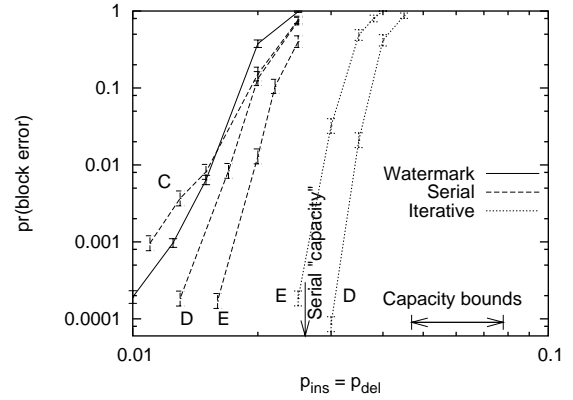


Figure 5: The benefit of full iterative decoding with  $R = 0.5$ ,  $N = 4000$  codes. Identical marker codes are shown simulated with iterative and non-iterative resynchronization. Also the reduction in error floor from identical markers (C) to non-identical markers (D) is shown. The outer LDPC code in codes C and D have only weight 2 and 3 columns, code E has weight 10 columns in addition. The watermark code result is from [3] and the details of the marker codes are in table 1.

The EXIT chart of the resynchronization was obtained by a Monte Carlo approach and a quadratic function fitted, figure 6(a). With no information passed from the code to the resynchronization stage the maximum rate of the LDPC outer code with serial decoding is shown. As the input information is increased the curve increases but does not reach 1. This is due to remaining uncertainty in the exact synchronization path and whether the bit in question may have been deleted (and possibly reinserted).

The behaviour of the decoding algorithm can be seen in terms of messages being passed between the variable nodes and check nodes. The variable node transfer function includes the EXIT chart of the resynchronization calculated above. The code that performs well on the Gaussian channel leads to an EXIT chart with an intersection, figure 6(b), ie decoding is not expected to converge. With the code with only weight 2 and weight 3 columns no intersection is observed, figure 6(c), and therefore decoding is expected to converge.

The width of the swath between the curves can be used as a metric to choose a better degree sequence. Better degree sequences were searched for using a global optimization package [4]. Better thresholds could only be found by increasing the number of weight 2 columns above the number of rows in the parity-check matrix. This is not expected to produce a good code as short cycles in weight 2 columns lead to low weight codewords.

## 4. CONCLUSION

We have shown that marker codes outperform watermark codes (as expected from effective capacity calculations). To avoid catastrophic decoding errors at higher noise levels the use of pseudo-random markers from a set of different markers is necessary.

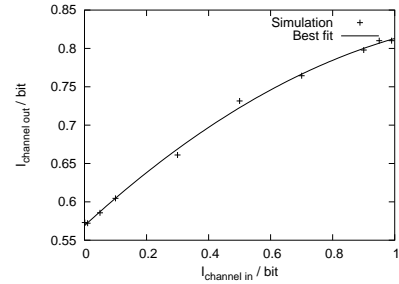
Iterative resynchronization provides better performance than serial resynchronization, bringing the waterfall region close to the bounds on the channel capacity.

## REFERENCES

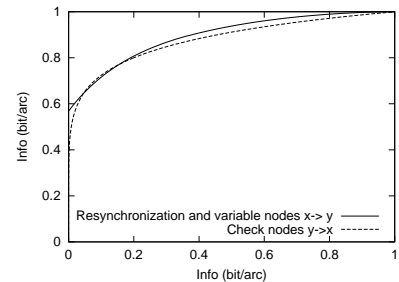
- [1] Sae-Young Chung. LDPC code design applet. <http://lids.mit.edu/~sychung/gaopt.html>.
- [2] Matthew C. Davey. *Error-Correction Using Low-Density Parity-Check Codes*. PhD thesis, University of Cambridge, 1999. Available at <http://www.inference.phy.cam.ac.uk/mcdavey>.
- [3] Matthew C. Davey and David J.C. MacKay. Reliable communication over channels with insertions, deletions, and substitutions. *IEEE Transactions on Information Theory*, 47(2):687–698, February 2001.
- [4] Joerg M. Gablonsky and C. Tim Kelley. A locally-biased form of the DIRECT algorithm. *Journal of Global Optimization*, 21:27–37, 2001.
- [5] Frederick F. Sellers Jr. Bit loss and gain correction code. *IEEE Transactions on Information Theory*, 8(1):35–38, January 1962.
- [6] Edward A. Ratzler. Low-density parity-check codes on Markov channels. In *Second IMA Conference on Mathematics in Communications*, December 2002.
- [7] Edward A. Ratzler and David J.C. MacKay. Codes for channels with insertions, deletions and substitutions. In *2nd International Symposium on Turbo Codes and Related Topics*, 2000.
- [8] Stephan ten Brink, Gerhard Kramer, and Alexei Ashikhmin. Design of low-density parity-check codes for multi-antenna modulation and detection. Submitted to *IEEE Transactions on Communications*, June 2002.
- [9] Jeffrey D. Ullman. On the capabilities of codes to correct synchronization errors. *IEEE Transactions on Information Theory*, 13(1):95–105, January 1967.
- [10] Andrew P. Worthen and Wayne E. Stark. Unified design of iterative receivers using factor graphs. *IEEE Transactions on Information Theory*, 47(2):843–849, February 2001.
- [11] Kamil Sh. Zigangirov. Sequential decoding for a binary channel with drop-outs and insertions. *Problemy Peredachi Informatsii*, 5(2):23–30, 1969.

	A	B	C	D	E
R	0.71	0.71	0.5	0.5	0.5
N	4995	4995	4000	4000	4000
d	19	19	9	9	9
m	01	01	001	001/110	001/110
$N_{LDPC}$	4521	4521	3001	3001	3001
$M_{LDPC}$	969	969	1001	1001	1001
$c_2$	968	963	1000	1000	927
$c_3$	3553	2785	2001	2001	1572
$c_{10}$	0	773	0	0	502

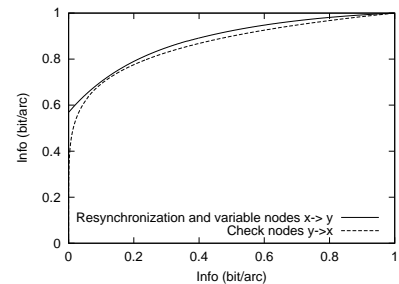
Table 1: The codes used in this paper. Markers (m) are inserted between every d data bits. If more than one marker is available this is chosen pseudo-randomly. The LDPC outer code was constructed with  $c_i$  columns of weight  $i$ .



(a) Insertion deletion channel



(b) Code E



(c) Code D

Figure 6: EXIT charts at  $p_{ins} = p_{del} = 0.04$