EXACT SYMBOL ERROR RATE FOR VARIABLE LENGTH CODES OVER BINARY SYMMETRIC CHANNEL

Jiantao Zhou, Xiaopeng Fan, Zhiqin Liang and Oscar C. Au

Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong, China
Email: {eejtzhou, eexp, zhiqin, eeau}@ust.hk

ABSTRACT

In this paper, we analyze the error recovery performance of variable length codes (VLCs) transmitted over binary symmetric channel (BSC). Simple expressions for the exact mean symbol error rate (MSER) and the exact variance of symbol error rate (VSER) for any crossover probability p_e are presented. We also prove that the mean error propagation length (MEPL) derived for single bit inversion error case is a scaled value of MSER when p_e tends to zero. Comparisons with simulations demonstrate the accuracy of the MSER and VSER expressions.

Index Terms— Variable length codes, Symbol error rate, Error recovery, BSC.

1. INTRODUCTION

Variable length codes (VLCs) have found widespread use for efficient encoding in many practical situations. However, a major drawback of VLC is that the decoder synchronization is required for correct decoding, and a loss of synchronization often leads to error propagation. To assess the error recovery (also called synchronization recovery) performance for VLCs, Maxted *et al.* developed a state model to calculate the statistical moments of the error span [1] (see also the corrections and additions of Monaco *et al.* [2]). In [3], Swaszek *et al.* simplified the model of [1] such that the mean and the variance of the delay until resynchronization can be obtained via the inversion and multiplication of dimension $n-1$ square matrices of constants. In [4], Zhou *et al.* derived a simple expression for the mean error propagation length (MEPL) and the variance of error propagation length (VEPL) by making use of the summation of semi-infinite series.

In the above methods a common assumption is that the transmission fault is a single inversion error [1, 3, 4]. Nevertheless, we favor a more general analysis that multiple errors, or comparably high error rates, are taken into consideration. In [5], Takishima *et al.* presented a formula to compute the average number of codewords received until synchronization is recovered for a BSC model with given crossover probability. It should be noted, however, that their results contain an unknown random variable and the expressions are equivalent to that of the single inversion error case, only with different definition of the transition matrix (details will be reported elsewhere). In this paper, we consider that the bit stream produced by VLC is transmitted over BSC. Simple expressions for the exact mean symbol error rate (MSER) and the exact variance of symbol error rate (VSER) for any crossover probability p_e are presented. We also prove that the MEPL derived for single inversion error case is a scaled value of MSER when p_e tends to zero. Comparisons with simulations demonstrate the accuracy of the MSER and VSER expressions.

The rest of this paper is organized as follows. Section II introduces the preliminary notions that we use in the sequel. Section III is the method for calculating the MSER and VSER. In section IV, we show the comparisons with simulations. Conclusions are briefly stated in the last section.

This work was supported by the Innovation and Technology Commission of the Hong Kong Special Administrative Region, China (project no. ITS/122/03 and project no. GHP/033/05

2. PRELIMINARY NOTATIONS

Let the source alphabet be $\mathcal{A} = \{a_1, a_2, \cdots, a_N\}$, and let the probability mass function (PMF) of this source be $p(a_1), p(a_2), \cdots, p(a_N)$. The source is encoded by a binary prefix code $\mathcal{C} = \{c_1, c_2, \cdots, c_N\}$, where c_i is the codeword of a_i. Let

$$\mathcal{C}^* \triangleq \bigcup_{n=0}^{\infty} \mathcal{C}^n$$

where $\mathcal{C}^n$ denotes the set of all sequences obtained by concatenating n codewords of $\mathcal{C}$. We call the elements in $\mathcal{C}^*$ sentences.

Given a code $\mathcal{C}$, we define $L_{max} = \max\{l(c)|c \in \mathcal{C}\}$, where $l(c)$ is the length of c, and length vector $\mathbf{L} = (L_1, L_2, \cdots, L_{L_{max}})$, where L_i is the number of codewords with length i. Let

Table 1. Five-character source

Symbol	Probability	Code
A	0.4	00
B	0.2	01
C	0.2	10
D	0.1	110
E	0.1	111

$$\Theta^* \triangleq \bigcup_{n:L_n>0} \{0,1\}^n$$

be the set of all possible received codewords when $\mathcal{C}$ is transmitted over BSC.

A square non-negative matrix $\mathbf{T}$ is said to be *primitive* if there exists a positive integer k such that $\mathbf{T}^k > \mathbf{0}$, where $\mathbf{0}$ denotes the matrix with all zero entries.

For a square matrix $\mathbf{T} = \{t_{ij}\}_{n\times n}$, suppose that there exists a real or complex number λ such that $\lambda\mathbf{u}' = \mathbf{u}'\mathbf{T}$, $\lambda\mathbf{v} = \mathbf{T}\mathbf{v}$, for some vector $\mathbf{u}$ and $\mathbf{v}$, where "$'$" stands for transpose. Then λ is called an eigenvalue of $\mathbf{T}$, and $\mathbf{u}$ ($\mathbf{v}$, respectively) is called a left (right) eigenvector of $\mathbf{T}$. The largest eigenvalue λ in magnitude of a primitive matrix is called the *Perron-Frobenius*(*PF*) eigenvalue, and according to *Perron-Frobenius* Theorem, λ is a real positive number [6]. The associated left and right eigenvectors $\mathbf{u}$ and $\mathbf{v}$, respectively, are called *PF* left and *PF* right eigenvectors if $\mathbf{u}$ and $\mathbf{v}$ are positive componentwise and $\mathbf{u}'\mathbf{1}_n = \mathbf{u}'\ \mathbf{v} = 1$, where $\mathbf{1}_n = (\underbrace{1, 1, \cdots, 1}_{n})'$ [6].

3. CALCULATION OF MSER AND VSER

3.1. Error Recovery and Extended Transition Matrix

To illustrate the error recovery process of VLC transmitted over BSC, let us first see the following example.

Example 1: Consider the encoding and decoding procedure shown in Fig. 1, where the symbols are encoded by the code given in Table 1. The bit inversion errors are underlined.

From this example, we find that there are three possible cases when the decoder parses a codeword received from BSC.

1) The codeword is correctly decoded [1]. In this case, the state after parsing is called *synchronization state without error*, denoted by SYN_1;

[1] As usual, the 'correctly decoded' means that the decoded symbol is exactly the same as the transmitted one, and all the received bits are consumed. Even though, appended with some remained bits, the received codeword may be decoded as a correct symbol concatenated with a sentence or a string of undecodable bits (see the first decoded symbol in desyn period 3 of Fig. 1), we still do not treat this case as correct decoding. Besides, in this paper, we only consider the synchronization in Levenshtein sense [7].

2) The codeword is incorrectly decoded, but all the received bits are consumed. In this case, the state after parsing is called *synchronization state with error*, denoted by SYN_2;

3) The codeword is incorrectly decoded, but not all received bits are consumed. The remained undecodable bits will match one of the internal nodes of the binary tree corresponding to the VLC. Since these bits will influence the consequent decoding, in this case, we call the state after parsing *Error State* i (ES_i), where i is the index of the internal node that the remained bits match.

The total error propagation length T is defined as the total number of symbols decoded in all desynchronization periods (in example 1, $T = 9$).

The set of states can then be written as

$$\mathcal{S} = \{ES_1, ES_2, \cdots, ES_M, SYN_2, SYN_1\} \quad (1)$$

where $M = N - 2$ is the number of internal nodes and N is the alphabet size. Every element in $\mathcal{S}$ is associated with a string of bits, specifically, ES_i is associated with the bit representation of the ith internal node, and SYN_1 and SYN_2 are associated with an empty string.

As usual, we assume that the source data are generated randomly according to the PMF. Since the bit errors may occur anywhere in the bit stream with equal probability, the sequence of states forms a Markov chain. Let π_{ij} ($1 \le i, j \le N$) be the transition probability from state i to j. Let

$$\Omega_\theta(i,j) = \{\theta \in \Theta^* : \exists c^* \in \mathcal{C}^* \text{such that } S_i\theta = c^* S_j\} \quad (2)$$

Then,

$$\pi_{ij} = \sum_{(\theta,k):\ \theta\in\Omega_\theta(i,j),\ l(c_k)=l(\theta)} \Big\{p(a_k)p_e^{d_H(\theta,c_k)} \times (1-p_e)^{l(c_k)-d_H(\theta,c_k)}\Big\} \quad (3)$$

where p_e is the crossover probability of BSC, and d_H stands for the Hamming distance.

It should be noticed that the extended transition matrix $\mathbf{\Pi} = \{\pi_{ij}\}_{N\times N}$ includes all the error states and two types of synchronization state, which is different from the error state transition matrix used in [1, 4]. Concerning the extended transition matrix, we present two Lemmas which will be used in the next sections (proof will be reported elsewhere).

Lemma 1: Suppose $\mathbf{\Pi} = \{\pi_{ij}\}_{N\times N}$ is an extended transition matrix. Then,

$$\lim_{k\to\infty} \mathbf{\Pi}^k = \mathbf{v}\mathbf{u}' = \mathbf{1}_N\mathbf{u}' \quad (4)$$

where $\mathbf{u}$ and $\mathbf{v}$ are, respectively, the *PF* left and *PF* right eigenvectors associated with $\mathbf{\Pi}$.

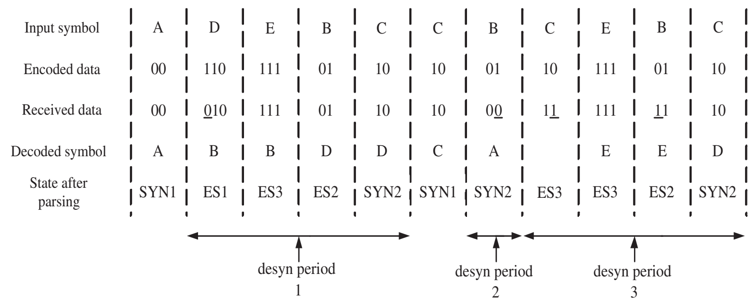

Fig. 1. The synchronization recovery process for the code shown in Table 1. The vertical lines indicate the boundary of correct code sequence. ES_1, ES_2 and ES_3, respectively, correspond to '0', '1' and '11'.

Lemma 2: Given an extended error state transition matrix $\mathbf{\Pi} = \{\pi_{ij}\}_{N\times N}$, suppose $\mathbf{u} = (u_1, u_2, \cdots, u_N)'$ is the *PF* left eigenvector, and $\mathbf{s}_1 = (\pi_{1N}, \pi_{2N}, \cdots, \pi_{NN})'$. Then,

$$\mathbf{u}'\mathbf{s}_1 = u_N \tag{5}$$

3.2. Calculation of MSER and VSER

Let $T(n)$ be the total error propagation length when the input symbol length is n. MSER is then defined as

$$\mu = \lim_{n\to\infty} \frac{MT(n)}{n} \tag{6}$$

where $MT(n)$ stands for the mean value of $T(n)$, and the mean value is averaged over all possible source data and all possible bit error positions with respect to the source distribution and the distribution of the bit error occurrence.

The VSER is accordingly defined as:

$$\sigma^2 = \lim_{n\to\infty} \frac{\sigma_T^2(n)}{n^2} \tag{7}$$

where $\sigma_T^2(n)$ denotes the variance of $T(n)$.

Theorem 1: The MSER for a VLC when the source is memoryless and the transmission channel is BSC, is given by

$$\mu = 1 - u_N \tag{8}$$

where u_N is the last component of the *PF* left eigenvector $\mathbf{u}$ associated with the extended transition matrix $\mathbf{\Pi}$.

Proof: Let $p_i(n)$ be the probability of ending up with state i after parsing the nth codewords. From the extended transition matrix we have $\mathbf{p}(n) = \mathbf{p}(n-1)\mathbf{\Pi}$, where $\mathbf{p}(j) = (p_1(j), p_2(j), \cdots, p_N(j))$. We may also write this as $\mathbf{p}(n) = \mathbf{p}(1)\mathbf{\Pi}^{n-1}$.

As $T(n)$ is a non-decreasing function with respect to the input symbol length n, we have

$$T(n+1) = T(n) + \Delta(n),\ \Delta(n) \geq 0 \tag{9}$$

From the transition process, we can find that only when the state after parsing $(n+1)$th codeword is SYN_1, $\Delta(n) = 0$, otherwise, $\Delta(n) = 1$. Hence, we can obtain the probability mass function of $\Delta(n)$

$$\begin{cases} Pr(\Delta(n) = 0) = \mathbf{p}(n)\mathbf{s}_1 \\ Pr(\Delta(n) = 1) = 1 - \mathbf{p}(n)\mathbf{s}_1 \end{cases} \tag{10}$$

where $\mathbf{s}_1 = (\pi_{1N}, \pi_{2N}, \cdots, \pi_{NN})'$ is the last column of $\mathbf{\Pi}$.

Taking expectations on Eq. 9, we have

$$\begin{cases} MT(n) = MT(n-1) + (1 - \mathbf{p}(n-1)\mathbf{s}_1) \\ \qquad\vdots \\ MT(2) = MT(1) + (1 - \mathbf{p}(1)\mathbf{s}_1) \end{cases} \tag{11}$$

Adding these equations together, we obtain

$$MT(n) = n - 1 + MT(1) - \mathbf{p}(1)(\sum_{i=0}^{n-2} \mathbf{\Pi}^i)\mathbf{s}_1 \tag{12}$$

It follows that

$$\begin{aligned} \mu &= \lim_{n\to\infty} \frac{MT(n)}{n} \\ &= 1 - \mathbf{p}(1)(\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-2} \mathbf{\Pi}^i)\mathbf{s}_1 \\ &\overset{\text{(a)}}{=} 1 - \mathbf{p}(1)\mathbf{1}_N\mathbf{u}'\mathbf{s}_1 \\ &\overset{\text{(b)}}{=} 1 - \mathbf{u}'\mathbf{s}_1 \\ &\overset{\text{(c)}}{=} 1 - u_N \end{aligned} \tag{13}$$

where (a) follows from Lemma1, (b) holds as all the probabilities sums up to unity, and (c) follows from Lemma 2. □

Lemma 3: The following limiting equality holds

$$\lim_{n\to\infty} \frac{\mathrm{Cov}(T(n), \Delta(n))}{n} = 0 \tag{14}$$

Theorem 2: The VSER for a VLC when the source is memoryless and the transmission channel is BSC, is given by

$$\sigma^2 = 0 \tag{15}$$

Proof: From Eq.9, we have

$$\sigma_T^2(n+1) = \sigma_T^2(n) + \sigma_\Delta^2(n) + 2\mathrm{Cov}(\Delta(n), T(n)) \tag{16}$$

By using a similar technique as shown in Eq. 11, we can obtain

$$\begin{aligned} \sigma_T^2(n) &= \sigma_T^2(1) + \mathbf{p}(1)\sum_{i=0}^{n-2}\mathbf{\Pi}^i\mathbf{s}_1 - \sum_{i=0}^{n-2}(\mathbf{p}(1)\mathbf{\Pi}^i\mathbf{s}_1)^2 \\ &+ 2\sum_{i=1}^{n-1}\mathrm{Cov}(\Delta(i), T(i)) \end{aligned} \tag{17}$$

Then

$$\begin{aligned} \lim_{n\to\infty}\frac{\sigma_T^2(n)}{n^2} &= \lim_{n\to\infty}\frac{\mathbf{p}(1)\sum_{i=0}^{n-2}\mathbf{\Pi}^i\mathbf{s}_1}{n^2} \\ &- \lim_{n\to\infty}\frac{\sum_{i=0}^{n-2}(\mathbf{p}(1)\mathbf{\Pi}^i\mathbf{s}_1)^2}{n^2} \\ &+ 2\lim_{n\to\infty}\frac{\sum_{i=1}^{n-1}\mathrm{Cov}(\Delta(i), T(i))}{n^2} \\ &\overset{(a)}{=} \lim_{n\to\infty}\frac{u_N - u_N^2}{n} \qquad (18) \\ &= 0 \qquad (19) \end{aligned}$$

where (a) follows from Lemma 1, 2 and 3. □

In the next Theorem, we establish a relationship between the MSER derived above and the MEPL obtained for single inversion error case [1, 3, 4].

Theorem 3: Let μ_s be the MEPL for single inversion error case. Then

$$\lim_{p_e\to 0}\frac{\mu}{p_e L_X(C)} = \mu_s \tag{20}$$

where p_e is the crossover probability of BSC.

4. EXPERIMENT RESULTS

In this section, taking the example of the five-character source shown in Table 1, we present comparisons of the theoretical results of MSER and VSER with the experimental results. By using Theorem 1, we can calculate the exact value of MSER with respect to p_e

$$\mu = \frac{4p_e^6 - 5p_e^5 - 86p_e^4 + 42p_e^3 + 284p_e^2 + 471p_e}{5p_e^4 - 20p_e^3 + 190p_e^2 + 480p_e + 55} \tag{21}$$

In Fig. 2a, the lines are obtained from Eq. 21 and the dots are obtained through experiments averaging over 10^4 samples. It can be seen that the experimental results match the theoretical results very well. In Fig. 2b, we find that VSER tends to zero as the input length n goes to infinity. In addition, when the input length n is large, the decay rate of the VSER is almost proportional to $1/n$, which is consistent with Eq. 18.

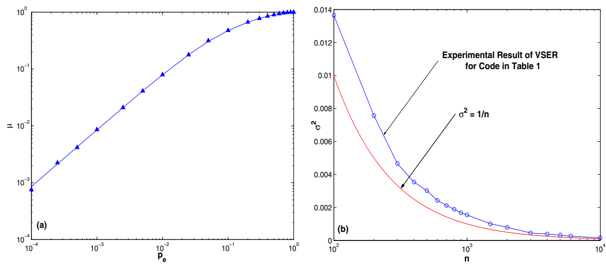

Fig. 2. (a)The MSER v.s. the crossover probability p_e (b) VSER v.s. input length n.

5. CONCLUSIONS

In this paper, we extend the analysis of the error recovery performance of VLCs to a BSC scenario. Very simple expressions for MSER and VSER are given. We also prove that MEPL for single inversion error case is a scaled value of MSER as the crossover probability p_e tends to zero.

6. REFERENCES

[1] J. Maxted and J. P. Robinson, "Error recovery of variable length codes," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 794–801, 1985.

[2] M. E. Monaco and J. M. Lawler, "Corrections and additions to 'error recovery for variable length codes'," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 454–456, 1987.

[3] P. Swaszek and P. DiCicco, "More on the error recovery for variable-length codes," *IEEE Trans. Inform. Theory*, vol. 41, pp. 2064–2071, 1995.

[4] G. Zhou and Z. Zhang, "Synchronization recovery of variable-length codes," *IEEE Trans. Inform. Theory*, vol. 48, pp. 219–227, 2002.

[5] M. Wada Y. Takishima and H. Murakami, "Error states and synchronization recovery for variable length codes," *IEEE Trans. Commun.*, vol. 42, pp. 783–792, 1994.

[6] E. Seneta, *Non-negtive matrices and markov chains*, Speinger-Verlag, New York, 1981.

[7] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Sov. Phys. –Dokl.*, vol. 10, pp. 707–710, 1966.