THE REDUNDANCY OF A SIMPLE SYNCHRONIZATION METHOD FOR VARIABLE LENGTH CODES

YU. M. SHTARKOV
Institute for Space Research
Academy of Sciences of the USSR
Moscow, USSR.

Summary Prefix insertion prior to the coded representation of every fixed length block of data provides a simple synchronization method for variable length coding. Unlike fixed length coding where the prefix appears with a set period, the appearance time of each prefix word in the variable length coded representation is a random variable. At the receiver a synchronization decision is made whenever a pattern within a threshold Hamming distance of the prefix is received. In this paper an expression is found for small synchronization error probabilities. This egression depends on the coded block length only through its average value $\bar{L}$. The optimal value for the recognition threshold is found. The necessary and sufficient condition for an arbitrarily small synchronization error probability is shown to be that the prefix length grows as log $\bar{L}$. The results are discussed from the viewpoint of data compression and source encoding.

Introduction It is well-known that when no synchronization channel is provided, synchronization can be realized serially by inserting a special prefix word A of length S, prior to each coded data block. At the receiver frame synchronization is achieved by comparing the received sequence with A for every bit shift in the incoming sequence. Whenever the Hamming distance in the comparison is less than or equal to a threshold T a synchronization decision is made.

A synchronization error occurs whenever channel errors change a prefix in more than T places or when the prefix pattern occurs within Hamming distance T in the data itself. For fixed length coding, the probability of synchronization error P can be decreased significantly by using the fact that the prefix appearance at the receiver is periodic. However, after data compression, a variable-length code block results so that the time of appearance of each prefix is a random variable. It is known for fixed-length coding that the redundancy of the method is negligible for moderate block lengths at arbitrarily small synchronization error probabilities. The main problem treated in this paper is the redundancy for variable-length coding when no periodic structure is available in the received sequence.

The only other paper which analyzes a similar, practical, synchronization scheme is that of Timor [1],who assumed the existence of a parallel synchronization channel.

Probability of Synchronization Error A synchronization error for any block is said to occur if the prefix word before and/or after the coded block is not recognized and/or a data pattern is mistakenly taken to be a prefix.

Let

P_1 = Prob[missing a prefix]
P_2 = Prob[finding sync in the data]

Then the probability of error is

$$P = 1 - (1-P_1)^2(1-P_2) \approx 2P_1 + P_2, \tag{1}$$

the approximation holding for small P_1, P_2 and consequently, P. Suppose signalling is q-ary, errors are symbol-to-symbol independent, the probability of correct reception is $1-\epsilon_2$ and the probability of transition to any other symbol is $\epsilon/(q-1)$. Obviously

$$P_1 = \sum_{i=T+1}^{S} \binom{S}{i} \epsilon^i (1-\epsilon)^{S-i} = 1 - \sum_{i=0}^{T} \binom{S}{i} \epsilon^i (1-\epsilon)^{S-i}. \tag{2}$$

Let

P_0 = Prob[finding sync in any one code segment of length S]

$$= \sum_{i=S-T}^{S} \binom{S}{i} \left(\frac{1}{q}\right)^i \left(1 - \frac{1}{q}\right)^{S-i} = \sum_{i=0}^{T} \binom{S}{i} \left(\frac{1}{q}\right)^{S-i} \left(1 - \frac{1}{q}\right)^i . \tag{3}$$

Assuming that the prefix word is designed so that no false sync recognition can occur when the pattern includes more than half of the prefix on either end, there are L possible shifts where an error event is possible. Each shift is <u>not</u> independent, though, because of overlap. Nonetheless, an upper bound on P_2 is obtained by ignoring the probability of joint false sync recognition events, i.e.,

$$P_2 \leq LP_0 \tag{4a}$$

Furthermore, there are at least L/S nonoverlapping patterns where sync can be falsely recognized so that

$$P_2 \geq P_2^* = 1 - (1-P_0)^{L/S} \approx \frac{L}{S} P_0 \tag{4b}$$

the latter approximation being valid for small values of P_2^*. Thus

$$\frac{1}{S} LP_0 \leq P_2 \leq LP_0,$$

or for some a, depending on L, $1/2S \leq \alpha \leq 1/2$:

$$P_2 = 2\alpha LP_0. \tag{5}$$

Then from (1-5)

$$P = 2\left\{1 - \sum_{i=0}^{T} \binom{S}{i} \left[\varepsilon^i(1-\varepsilon)^{S-i} - \alpha L \left(\frac{1}{q}\right)^{S-i} \left(1 - \frac{1}{q}\right)^i\right]\right\} \tag{6}$$

Note that L, the coded block length is a random variable. Taking the expectation of (6) and noting that by the law of the mean, $\overline{\alpha L} = \alpha^* \bar{L}$, where α^* also satisfies $1/2S \leq \alpha^* \leq 1/2$:

$$\overline{P} = 2\left[1 - \sum_{i=0}^{T} \binom{S}{i} f_\varepsilon(i)\right] \tag{7}$$

where

$$f_\varepsilon(i) = \varepsilon^i(1-\varepsilon)^{S-i} - \alpha^* \overline{L} \left(\frac{1}{q}\right)^{S-i} \left(1 - \frac{1}{q}\right)^i \tag{8}$$

Thus we have proven:

Theorem 1. To a good approximation the synchronization error probability depends on the length of the codeword only through its average value as shown in equations (7) and (8).

The Optimal Threshold T_0 In the subsequent material, we use q for the base of the logarithm $\lfloor . \rfloor$ to denote the integer part of the contents. The optimal value of the threshold T is denoted T_0 and minimizes $\overline{P}$ in (7-8) for fixed S and $a^*\bar{L}$. We will prove the following:

Theorem 2. If

$$\varepsilon < 1 - \frac{1}{q}, \tag{9}$$

then

$$T_0 = \max\left\{0, \left\lfloor \frac{[1 + \log(1-\varepsilon)]S - \log \alpha^* \overline{L}}{\log(1-\varepsilon) + \log(q-1) - \log \varepsilon} \right\rfloor\right\}. \tag{10}$$

Comment: Due to (9) the denominator in (10) is positive.

Proof: From (9), $(q-1)(1-\epsilon)/\epsilon < 1$. Using this we develop the recursive relationshi p

$$f_\epsilon(i+1) = \frac{\epsilon}{1-\epsilon}\left\{\epsilon^i(1-\epsilon)^{S-i} - \frac{(1-\epsilon)(q-1)}{\epsilon}\ \alpha^{*\bar{L}}\left(\frac{1}{q}\right)^{S-i}\left(1-\frac{1}{q}\right)^i\right\}$$

$$< \frac{\epsilon}{1-\epsilon} f_\epsilon(i), \tag{11}$$

so that $f_\epsilon(i)$ has at most one zero crossing from the positive side. If $f_\epsilon(0) < (0)$, it is obvious from (7-8) that the optimal threshold is $T_0 = 0$. If $f_\epsilon(0) > 0$, T_0 is the largest integer for which $f_\epsilon(T_0) > 0$, i.e.,

$$f_\epsilon(T_0) > 0, \quad f_\epsilon(T_0 + 1) \leq 0, \tag{12}$$

from which, with (8), (10) follows. The theorem is proved.

The Asymptotic Behavior of S as $\bar{L} \to \infty$. Suppose that it is desired to achieve $P < \delta < 1/2$, δ an arbitrarily small constant, for fixed ϵ, by choosing S and $\bar{L}$ with $T = T_0$, its optimal value. In general, the solution is difficult. However, asymptotically in $\bar{L}$, we have the following theorem:

Theorem 3. As $\bar{L} \to \infty$, for an arbitrarily small synchronization error probability, it is necessary and sufficient that

$$S = \theta \log \alpha^{*\bar{L}} \tag{13}$$

where θ is a constant satisfying

$$\theta \geq \theta_1 = \frac{1}{1-H_q(\epsilon) - \epsilon \log(q-1)}$$

$$H_q(\epsilon) = -\epsilon \log\epsilon - (1-\epsilon)\log(1-\epsilon). \tag{14}$$

Proof: Using (13) in (10) we obtain

$$T_0 \approx \frac{1 + \log(1-\epsilon) - \theta^{-1}}{\log(1-\epsilon) + \log(q-1) - \log \epsilon} \quad S = \gamma S \tag{15}$$

Consider (2) and (3) for the two types of error probabilities. Using well-known results for the tails of a binomial distribution [2,ch.7], P_1 and P_0 go to zero with $L \to \infty$ if y is such that

$$\epsilon < \gamma < 1 - \frac{1}{q} \tag{16}$$

If (16) does not hold, either P_1 or P_0 goes to 1/2 or 1. Thus (16) is necessary and sufficient. From (15), $\gamma > \epsilon$ implies (14). $\gamma < 1 - 1/q$ in (15) can be rearranged to give

$$\theta^{-1} > -\frac{1}{q}\log\frac{1}{q} - (1-\frac{1}{q})\log(1-\frac{1}{q}) + \frac{1}{q}\log(1-\epsilon) + (1-\frac{1}{q})\log\epsilon \qquad (17)$$

In accordance with the standard properties of entropy [2, pg.71], the right hand side of (17) is nonpositive and thus (17) is satisfied for every $\theta > 0$. Then (14) is only necessary and sufficient condition. The theorem is proved.

Remark 1. From (16) it follows that (14) is a sufficient condition even if the optimal value of T_0 in (10) is not used. If (10) is used as well (14) is necessary.

Remark 2. It is simple to show that P_1 an P_2 go to zero exponentially at the same rate up to a constant multiplier as $\bar{L} \to \infty$, i.e., both are asymptotically of the form $B\exp(-\beta\bar{L})/\sqrt{\bar{L}}$ with different values of B but the same value of β. This also follows from the optimality of T_0.

Remark 3. The value of α^* does not influence the asymptotic behavior of S. That is, from (13) and thebounds on α^*, $1/2S \le \alpha^* \le 1/2$, for any θ (which does not depend on α^*):

$$\theta[\log\bar{L} - \log 2S] \cong \theta[\log\bar{L} - \log\log\bar{L}] \le S \le \theta[\log\bar{L} - \log 2],\ \theta > \theta_1, \qquad (18)$$

so that S grows as $\theta \log \bar{L}$ as $\bar{L} \to \infty$.

Connection with Data Compression and Source Encoding In most cases of practical interest, variable-length coding is used for data compression (but Timor [1] listed other situations such as in the varying of an onboard spacecraft experiment). Then, from the last section, the use of the simple method of synchronization described in this paper results in a decrease in data compression ratio by a factor of less than

$$K = \frac{\bar{L}}{\bar{L} + \theta\log\bar{L}} \approx 1 - \theta\frac{\log\bar{L}}{\bar{L}} \qquad (19)$$

over data without synchronization. Obviously, for reasonable values of $\bar{L}$, K is essentially one. Note that from (14) for reasonable values of ϵ, $\theta \approx 1$. In addition, for fixed ϵ, increasing q decreases θ.

An important question regarding the results of this paper is the following. As $\bar{L} \to \infty$ on a noisy channel most codewords will be in error in spite of good synchronization. Therefore, why should one be interested in the asymptotic behavior of S? There are two answers to this question.

First, for adaptive telemetering the errors have a local effect only which does not depend on $\bar{L}$. The second answer involves the use of error-correcting codes. Suppose that after inserting the data prefix we divide the message into blocks of length k and use a (k+r,k) error-correcting code. If $\bar{L} \approx k$, a small decoding error probability will keep most of the coded blocks error-free. In this event, most of the prefix words will be error-free as well. On the average, the number of errors is less than ϵ. However, when a decoding error does occur, there will be a large number of errors in the prefix word and in most cases the prefix will not be recognized. Therefore,

$$\bar{P} < \text{Prob[decoding error]} + P_2. \tag{20}$$

In any event, we can choose a (k+r,k) code in such a way that the probability of a decoding error goes to zero in the same way as P_2 does with $\bar{L}$ (as noted above, $\bar{L} \approx k$. Thus, most of the codewords will be eror-free and $P \to 0$ as before.

We pointed out earlier that for the optimal value T T_0, the values P_i go to zero with $\bar{L}$ as $\bar{L}$ as $B_i \exp\{-\beta\bar{L}\}/\sqrt{\bar{L}}$, $i = 1,2$. The difference between B_1 and B_2 results in a different average received codeword rate from that transmitted. This could be adjusted by a slight variation in T about T_0.

As a final point, we will discuss the relationship of these results to source encoding. If we encode source output blocks of length n, then the shortest average coded blocklength satisfies

$$nH \leq \bar{L} \leq nH + O(n), \tag{21}$$

where H is the stationary source entropy (or alternately, the rate distortion function). For the variable length code we must choose a synchronization prefix of length

$$S = \theta \log \alpha^* \bar{L} \approx \theta(\log n + \log H + \log \alpha^*), \ \theta > \theta_1. \tag{22}$$

The per output symbol synchronization redundancy is then

$$\rho_n = \frac{\theta \log n}{n} + O\left(\frac{1}{n}\right). \tag{23}$$

It is interesting to note that the same redundancy results when we don't know the source statistics (see, for example, [3]) and use universal coding methods. If we don't know the statistics and use this synchronization method, then we can put H* = maximum possible source entropy in (21) and (22). But this appears only in the 0(1/n) term in (23).

In actuality the redundancy of this synchronization method is really less than in (23), because it is not necessary to satisfy Kraft's inequality [2, p. 69]. It is not known what the

exact effect of this point is on $\bar{L}$ and ρ_n in (23), but calculations for some examples have shown only a slight difference.

Conclusion It is useful to mention some possible improvements on the methods given in this paper. For example, if during the encoding procedure a prefix pattern appears in the coded block, some of the code symbols can be changed so that wrong synchronization will not occur at the receiver. The resulting distortion in the reproduction will be slight compared with the distortion which would result from a synchronization error. This method was first proposed by Butman [4]. It is tightly connected with the idea of fixed rate source encoding [2, p. 101].

The second improvement is connected with the application of error-correcting codes. From (20) we know that the only problem involves decreasing $P_2(P_1 = 0)$. Then we can use the Gilbert prefix comma-free codes [5] or the constructive method for such codewords by the Artom method [6]. This method is similar to that of Butman [4] for a noiseless channel.

References

[1] U. Timor, "Frame synchronization in time-multiplexed PCM telemetry with variable frame length," IEEE Trans. on Communications, vol. COM-20, N5 1005-1008; Oct. 1972.

[2] R. M. Fano, "Transmission of information," The MIT Press and John Wiley and Sons, Inc., New York, London; 1961.

[3] Yu. M. Shtarkov, "Encoding of finite lengths' messages on the output of source with unknown statistics," Proc. of V Conf. on coding theory and information transmission, Moscow-Gorkiy, vol. I, pp. 147-152; 1972.

[4] S. Butman, "Synchronization of PCM channels by the method of word stuffing," IEEE Trans. on Communication Technology, vol. COM-16, no.2 252-254; Apr. 1968.

[5] E. N. Gilbert, "Synchronization of binary messages," IEEE Trans. on Information Theory, vol. IT-6, 470-477; 1960.

[6] A. Artom, "Choice of prefix in self-synchronizing codes," IEEE Trans. on Communications, vol COM-20, no. 2, 253-254; Apr. 1972.