FACTA UNIVERSITATIS (NIŠ)
Series: Electronics and Energetics vol. 9, No. 2 (1996), 191–205

SEARCH PROCESS: THE ANALYSIS

Dragana Bajić and Dušan Drajić

Abstract. Statistical analysis of the duration of search to find a fixed L–ary sequence in a stream of random L–ary equiprobable data, as well as in a frame, is performed. The expressions obtained may be used for evaluating some of the parameters considering frame synchronisation in digital communication systems.

1. Introduction

Frame synchronisation is crucial to the proper functioning of digital communication systems. A common way to obtain synchronisation is to add the redundancy by periodical insertion of a unique sync pattern into the stream of data. The data stream is usually scrambled [5], thus data symbols may be regarded as random and equiprobable. A common procedure to obtain the synchronisation is to use the sliding window search procedure, i.e. to observe the received data through the window whose size equals the sync pattern length. If the window contents match symbol-to-symbol to the known sync pattern, correct synchronisation is supposed and verification can start. This paper deals with the statistical parameters of the search process, in random data and in frame. In previous works (i.e. [1,6]) these parameters were obtained by simulation study, while in this paper the exact formulae are derived.

2. Search process in random data

The search process in random data means that window of size N is placed over N successive digits of the observed sequence. This is the first test, at the first position ($k = 1$), and it is positive if the content of window equals to the chosen N–digit pattern. If the test is negative, the window slides one digit, and its new content (actually, only one digit is a new one) is compared to

Manuscript received Mart 14, 1996.

The authors are with Faculty of Electrical Engineering, University of Belgrade, Bul. Revolucije 73, P.O.Box 816, 11001 Belgrade, Yugoslavia.

the chosen pattern (the second position, $k = 2$). The procedure is repeated until the position where the test is positive, i.e. until the pattern is found.

Suppose that the first occurrence of the chosen N-digit pattern is at the kth position after the random starting point, as shown in Fig. 1. The searched random sequence is of length $k + N - 1$ and satisfies the following condition: the last N digits are the only N consecutive digits of the observed sequence that correspond digit–to–digit to the chosen pattern (”matching condition”).

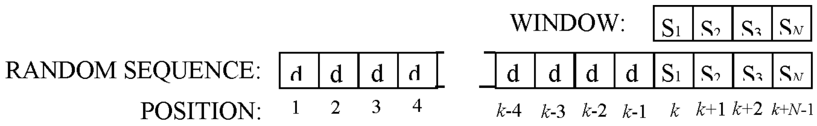

Figure 1. Search process in random data ($N = 4$).

In the classical paper [7] a formula for the expected duration of search for a fixed pattern in random data is derived as

$$T_R = \sum_{i=0}^{N} h_i \cdot L^i - N, \tag{1}$$

where N represents the number of digits in the fixed pattern (i.e. its length), each digit being equiprobable and randomly chosen from an alphabet of L letters. Bifix indicator h_i, $i = 0, \ldots, N$ is introduced with the following meaning: $h_i = 1$ if a bifix (sequence that is both prefix and suffix) of length i exists; otherwise, $h_i = 0$ and by convention $h_0 = h_N = 1$. For example, the 8–bit pattern 01011010 has two bificis: one of length 1 (0) and the other of length 3 (010), so $h_0 = h_1 = h_3 = h_8 = 1$, while $h_2 = h_4 = h_5 = h_6 = h_7 = 0$.

In order to derive the distribution function for the same process (for which (1) is the expected value), the number of sequences of length $k+N-1$ that satisfy the matching condition must be found. This number is denoted by a_k and can be, using a recursion, expressed as [2]

$$a_k = \sum_{i=1}^{\min(N,k-1)} (L \cdot h_{N+1-i} - h_{N-i}) \cdot a_{k-i}. \tag{2}$$

Obviously, $a_1 = 1$, as there is only one sequence of length N ($k = 1$) that corresponds digit–to–digit to the chosen N–digit pattern. For further discussion, consider binary bifix–free pattern 001 ($L = 2, N = 3$). The appearance of this pattern at positions $k-2$ and $k-1$ does not affect its appearance at position k (heavy shaded states in Fig. 2), so that the number of sequences of length $k+N-1$ that satisfy the "matching condition" should be twice the number of one–bit–shorter sequences (i.e. a_k should be $2 \cdot a_{k-1}$). On the other hand, each appearance of the pattern at position $k-3$ forbids one sequence with 001 at its k-th position (lighter shaded state in Fig. 2), so that the number of such sequences (a_{k-3}) should be subtracted from a_k, i.e. $a_k = 2 \cdot a_{k-1} - a_{k-3}$ (in accordance with (2)). Thus, the sequence $A = [a_1, a_2, \dots]$ is $[1, 2, 4, 7, 1220, 33, \dots]$.

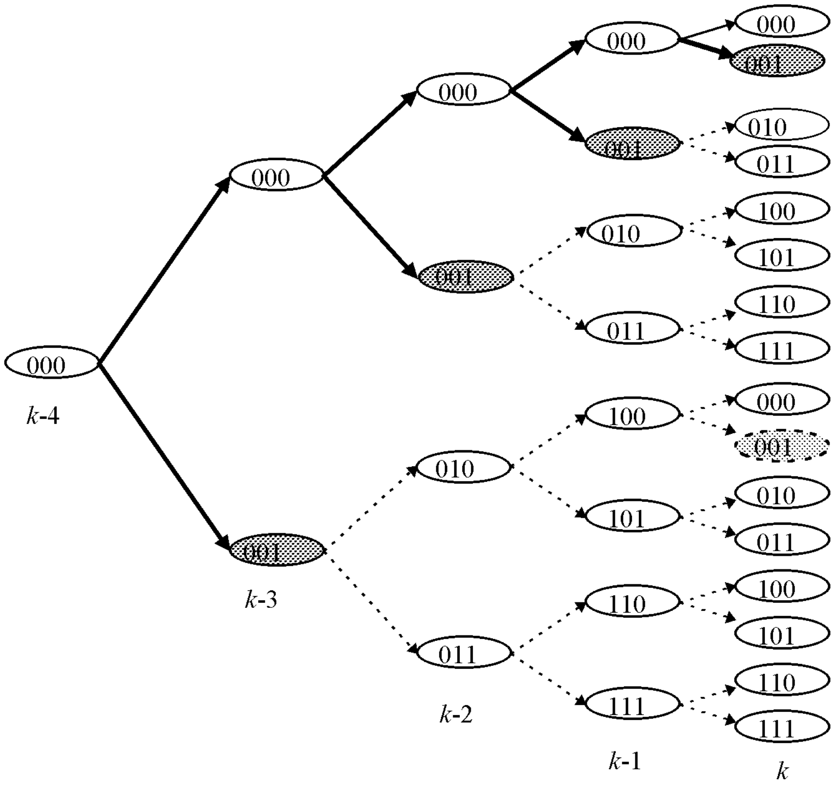

Figure 2. Tree diagram showing occurrences of pattern 001.

Consider now the pattern 010 ($h_2 = 0, h_0 = h_1 = h_3 = 1$). The relationship between the occurrences of the pattern at the specific position is shown in Fig. 3.

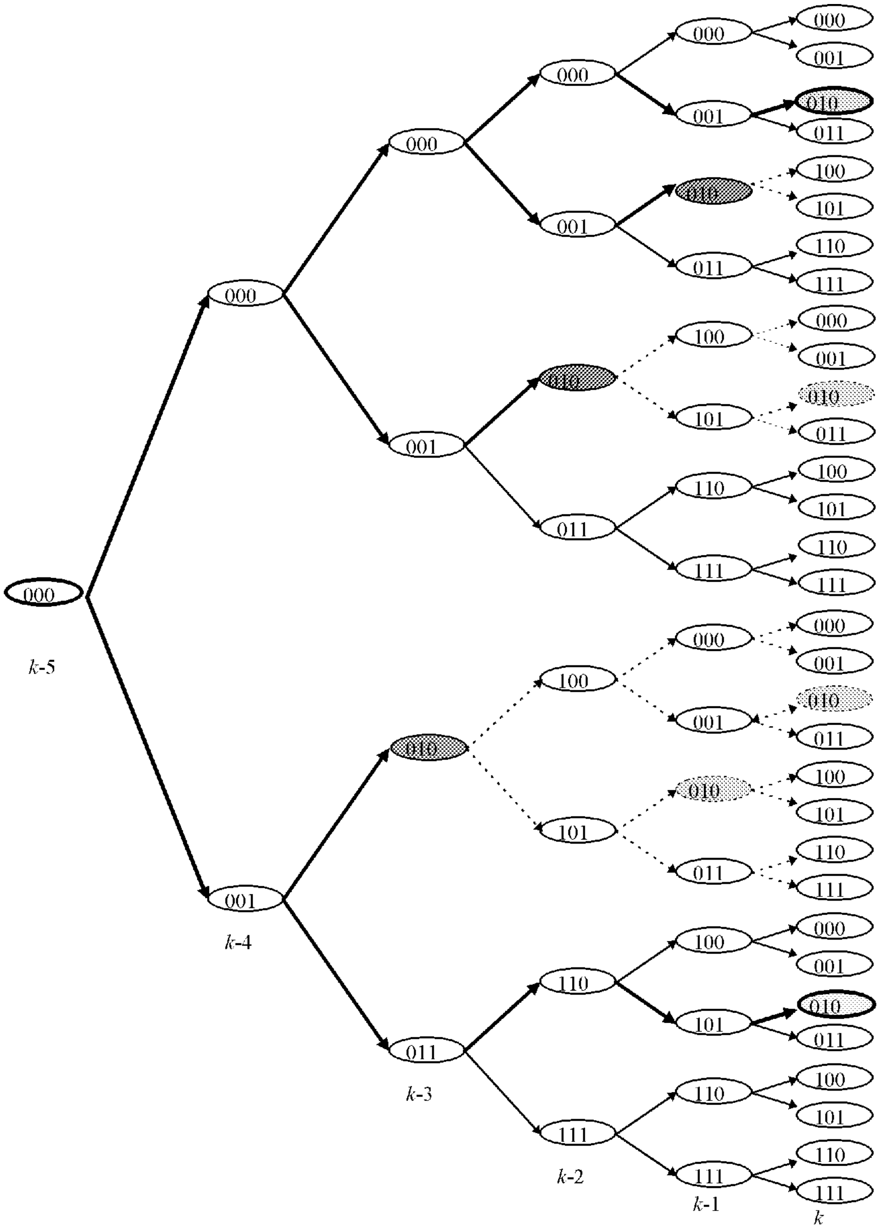

Figure 3. Tree diagram showing occurrences of pattern 010.

Some of the sequences at position $k-1$ (the number of which should be doubled in order to obtain a_k) are already forbidden by the appearance of this pattern at position $k-3$ (lighter shaded state in Fig. 3), so that their number should be added to a_{k-1} in order to balance the loss giving $a_k = 2(a_{k-1} + a_{k-3})$. On the other hand, each pattern appearing at position $k-3$ and at position $k-2$ prevents one sequence with pattern 010 at the k-th position (lighter shaded state in Fig. 3), so that the total number of prevented sequences should be subtracted, i.e. finally $a_k = 2(a_{k-1}+a_{k-3})-a_{k-2}-a_{k-3}$ (in accordance with (2) as well). Here, $A = [1, 2, 3, 5, 9, 16, 28, \dots]$. A similar line of reasoning led to (2).

The probability that the N–digit pattern will occur for the first time at the k-th position within the stream of random data (Fig. 1) equals

$$P\{k\} = a_k \cdot p^{N=k+1} = b \cdot a_k \cdot p^k, \qquad b = p^{N-1}, \tag{3}$$

as there are a_k sequences of length $k + N - 1$ satisfying the "matching condition", while the probability of each one equals p^{k+N-1} ($p = 1/L$ is the probability of a random equiprobable digit).

Expression (3) (the probability distribution function) satisfies

$$S\{x\} = \sum_{i=1}^{\infty} P\{i\} = \sum_{i=1}^{\infty} b \cdot a_i \cdot p^i = 1. \tag{4}$$

The expected duration of search for the fixed pattern in random data can be found by statistical methods, yielding, naturally, the same result as in [7], i.e.

$$T_R = \sum_{i=1}^{\infty} iP\{i\} = \sum_{i=1}^{\infty} i \cdot b \cdot a_i \cdot p^i = \sum_{i=0}^{N} h_i \cdot L^i - N. \tag{5}$$

Variance is evaluated as

$$\sigma^2 = (T_R + N) \cdot (T_R + N + 1) - 2\sum_{i=0}^{N} i \cdot h_i \cdot L^i. \tag{6}$$

The evaluation of (4), (5) and (6) is given in the Appendix A.

The probability distribution functions for 6 different five–bit patterns is plotted in Fig. 4. The simulation study results are also plotted for the bifix–free binary pattern 00101, for which $h_i = 0$, $i = 1, \dots, 4$ and the all–zero pattern for which $h_i = 1$, $i = 0, \dots 5$; the expected duration of search and their variance for binary bifix–free [8] and the all–zero patterns vs. pattern length are plotted in Fig. 5. In both figures, thick lines represent the results according to formulae (3), (5) and (6), while the dashed lines represent the simulation study, simulation being performed over the sample of 100000 searches (Fig. 4) and 10000 searches (Fig. 5). Better concordance of calculated and simulated data in Fig. 5. is a consequence of the fact that the number of simulation runs was 10000 for each of the simulation points. On the other hand, Fig. 4. shows the probability distribution function, so each of the simulation runs results in one of 100 possible values, therefore the number of simulation runs for a single point was, on average, 100000/100 = 1000.

It may be noticed that the p.d.f. is similar to the exponential distribution. Moreover, the values of variance are approximately the squares of mean values of the search time, also a characteristic of the exponential distribution. Fig. 6 represents the relative error for 6 different 5–bit patterns contrasted to the exponential function:

$$\varepsilon_R = \frac{\mid P\{k\} - \dfrac{1}{T_R} e^{-\dfrac{k}{T_R}} \mid}{\dfrac{1}{T_R} e^{-\dfrac{k}{T_R}}}. \tag{7}$$

For $k > 5$ the values of relative error are less than a few percents. Better fitting is obtained for patterns with periodical structure (00000 and 01010).

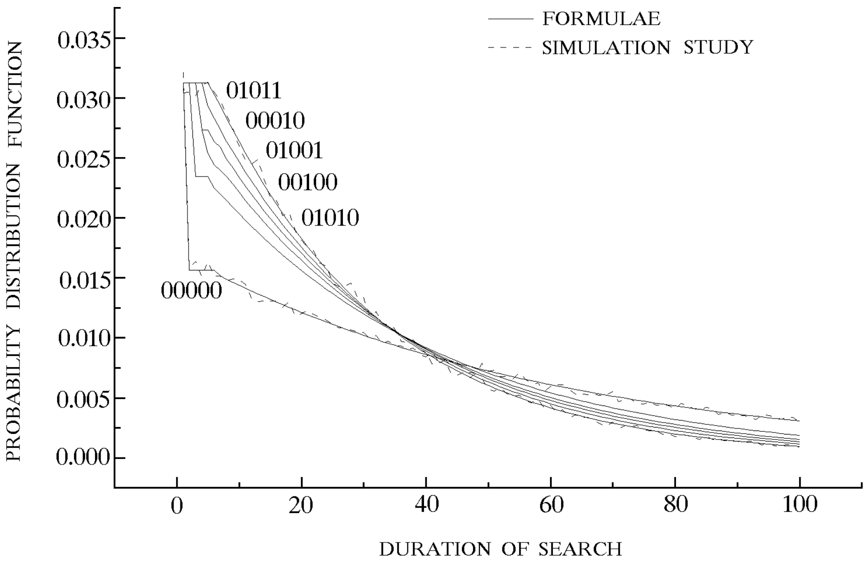

Figure 4. Probability distribution function for different 5–bit patterns.

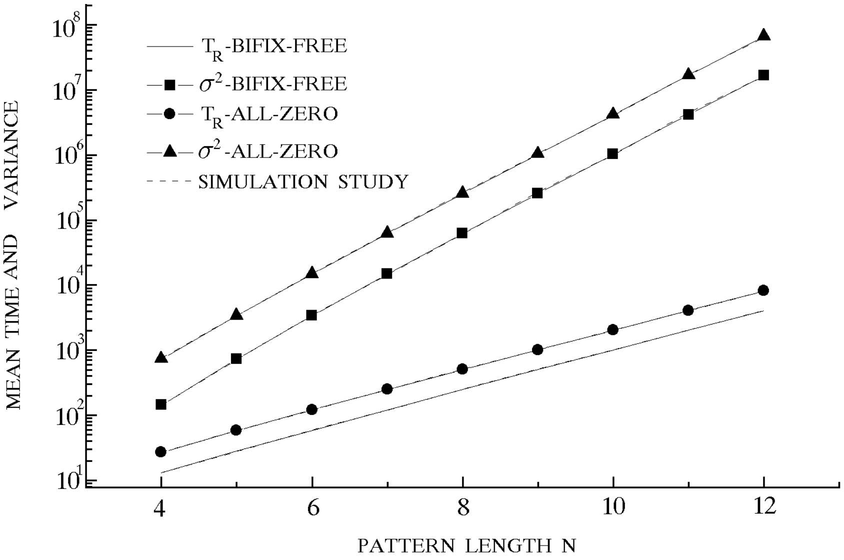

Figure 5. Mean search time and variance for bifix–free and all–zero patterns.

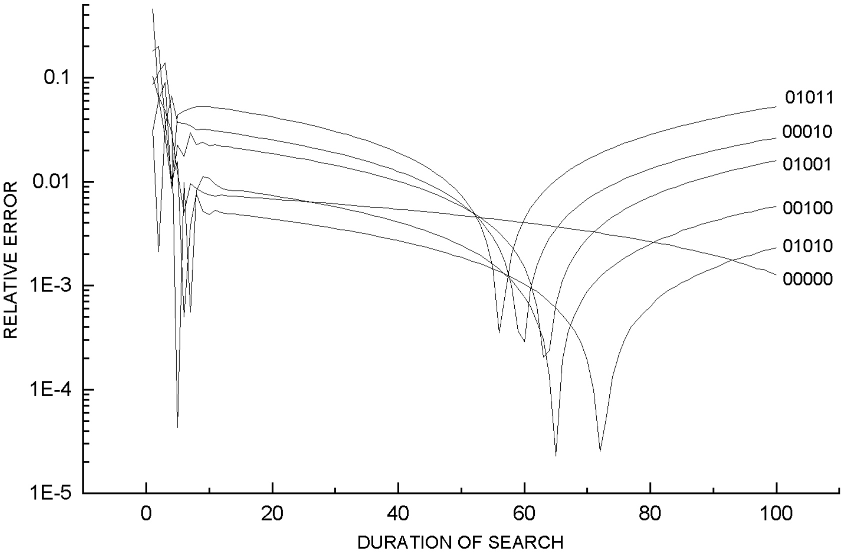

Figure 6. Relative error of p.d.f. in respect to exponential distribution.

3. Search process in frame

While the search process in random data is of theoretical interest, for practical systems the more interesting problem is the search process in the frame, where the sync pattern of length N is periodically inserted into the stream of equiprobable random (scrambled!) data, its period being F digits (frame length). In spite of the interest in problem, the analyses so far were based on simulation studies [1,6].

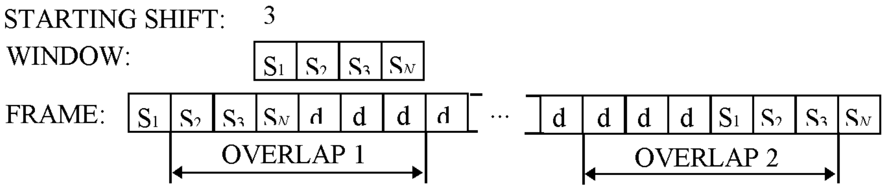

Figure 7. Search process in frame ($N = 4, S = 3$).

The search process starts at a random position, S digits shifted from the real sync pattern position (Fig. 7). The p.d.f. of event that the pattern will be found for the first time at exactly k-th test, $P\{k\}$, depends on the starting position S, the frame length F and the pattern structure described by bifix indicators h_i, $i = 0, \ldots, N$.

If the starting position S is within the data region, then the search process is equivalent to the search process in random data and Eq. (3) describes its p.d.f. At the moment when the sliding window reaches the second overlap region, the possibility of pattern simulation depends on the existence of bifix of corresponding length, so the formula for p.d.f. has to be multiplied by the corresponding bifix–indicator. The same apply if S is within the second overlap region.

If the starting position S is within the first overlap region, the situation is more complicated. Let k_m denotes the first position within the 1-st overlap region where sync pattern can be simulated. For example, if the search for the binary sync pattern 0101 ($h_2 = 1$) starts at $S = 1$, the first test will be negative (sequence $101d$, d-data, cannot simulate sync pattern). The 2-nd test might be positive ($a_2 = h_2 = 1$) if the first 2 data bits equal to 10, so $k_m = 2$. The third test is negative and $a_3 = h_1 - h_3 = 0$. The fourth test ($h_0 = 1$) might be positive, if the first 4 data bits equal to 1010 – but this could never happen, as the first two data bits 10 were already "used" for the simulation at the 2-nd position, thus $a_4 = h_0 - h_2 = 0$. Therefore, the complete formula for p.d.f. of the search process in frame can be expressed as

a) $1 \leq S \leq N-1$ (the 1-st overlap region)

$$P\{k\} = \begin{cases} a_k \cdot p^{S+k-1}, & 1 \leq k \leq F-N+1-S \\ h_{N+S+k-1-F} \cdot a_k \cdot p^{F-N}, & F-N+1-S \leq k \leq F-S+1 \end{cases}$$

$$a_k = \begin{cases} h_{N-S+1-k}, & 1 \leq k \leq k_m \\ h_{N-S+1-k} - h_{N-(k-k_m)}, & k_m < k \leq N-S+1 \\ \sum\limits_{j=1}^{\min(k-1,N)} (L \cdot h_{N+1-j} - h_{N-j}) \cdot a_{k-j}, & N-S+1 < k \leq F-S+1 \end{cases}$$

$$k_m = \begin{cases} 1, & h_{N-S} = 1 \\ k, \quad \sum\limits_{j=1}^{k-1} h_{N-S+1-j} = 0, \quad h_{N-S+1-k} = 1, & k \leq N-S+1 \end{cases}$$

b) $N \leq S \leq F-1$ (including the 2-nd overlap region)

$$P\{k\} = \begin{cases} a_k \cdot p^{N+k-1}, & 1 \leq k \leq F-N+1-S \\ h_{N+S+k-1-F} \cdot a_k \cdot p^{F-S}, & F-N+1-S < k \leq F-S+1 \end{cases}$$

$$a_k = \begin{cases} 1, & k = 1 \\ \sum\limits_{j=1}^{\min(k-1,N)} (L \cdot h_{N+1-j} - h_{N-j}) \cdot a_{k-j}, & 1 < k \leq F-S+1 \end{cases} \tag{8}$$

where a_k, as in Eq. (3), denotes the number of sequences of length $k+N-1$ that satisfy the matching condition.

The evaluated p.d.f. is actually the probability of sync pattern simulation, except $P\{F-S+1\}$, which is the probability that the sync pattern will be found at its real position without previous simulations ("non-simulation probability", P_{NS}). This probability, for $S=1$, is plotted in Fig. 8, including the frame lengths of classical European PCM systems [4]. The superiority of bifix-free patterns (already known to prevent the simulation of sync pattern in the overlap regions [1,6]) is obvious for longer sync–patterns (further comments are given in Appendix B).

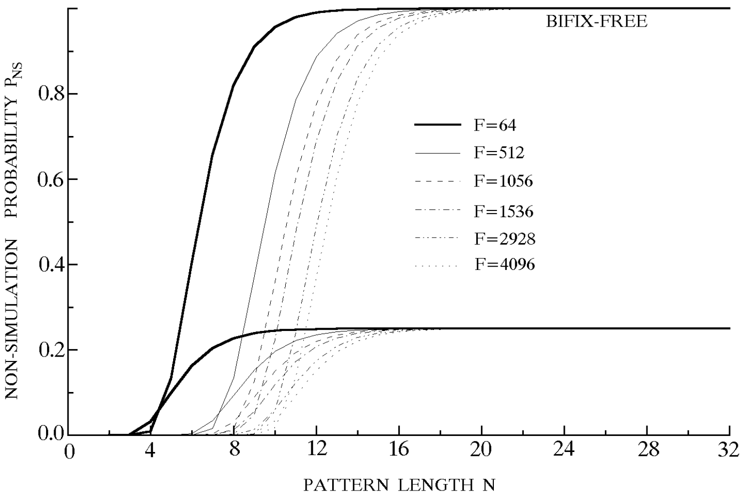

Figure 8. Non-simulation probability for different pattern lengths ($S = 1$).

The dependence of P_{NS} upon the starting search shift S is shown in Fig. 9 ($F = 512, N = 7$ – primary PCM). Bifix–free and all–zero patterns are drawn in solid lines, dashed lines representing other different 7–bit patterns (similar figure, obtained by simulation study, is given in [1]). The all–zero pattern is better than bifix–free up to the $S = 340 = T$ (turning point). The relative turning point ($T \cdot 100/F[\%]$) vs. pattern length N is plotted in Fig. 10. It may be concluded that for each N, there exists a maximum frame length ("turning length") for which P_{NS} of a bifix–free pattern is greater or equal to P_{NS} of the all–zero pattern (for the worst case, $S = 1$, and therefore for all other values of S). The turning length is plotted in Fig. 11 (dashed line) and for $N = 7$, 10 and 12 (European PCM) equals to 342, 2819 and 11333 respectively, leading to the conclusion that, for primary PCM ($F = 512 > 342$) the choice of $N = 7$ was improper (the existence of better structures is already shown in [1,3]). The same figure shows the maximum frame length for which P_{NS} is still less than 0.99, 0.9 and 0.5 which might be useful for further implementation.

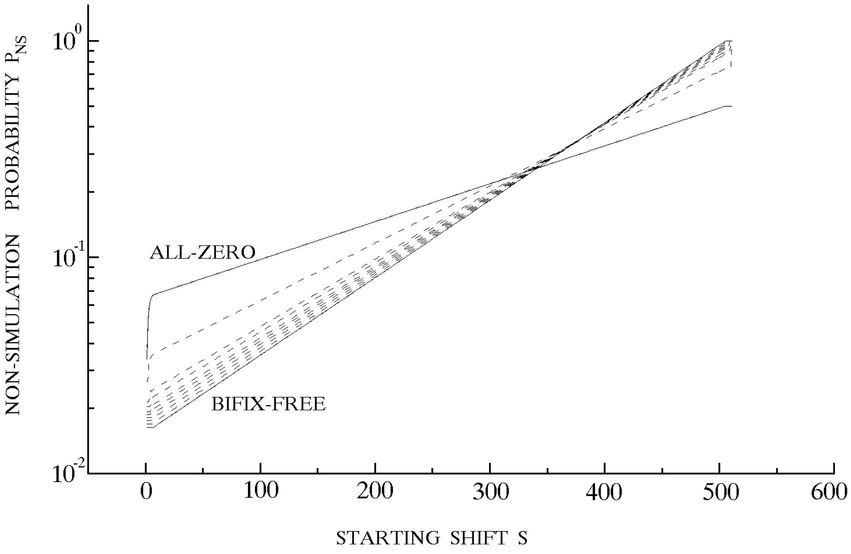

Figure 9. Non–simulation probability for different 7–bits patterns.

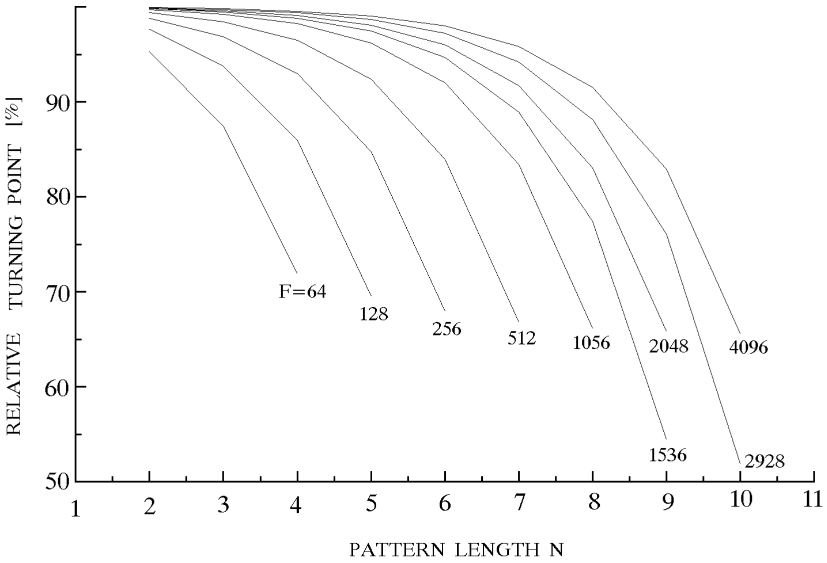

Figure 10. Relative turning point vs. pattern length.

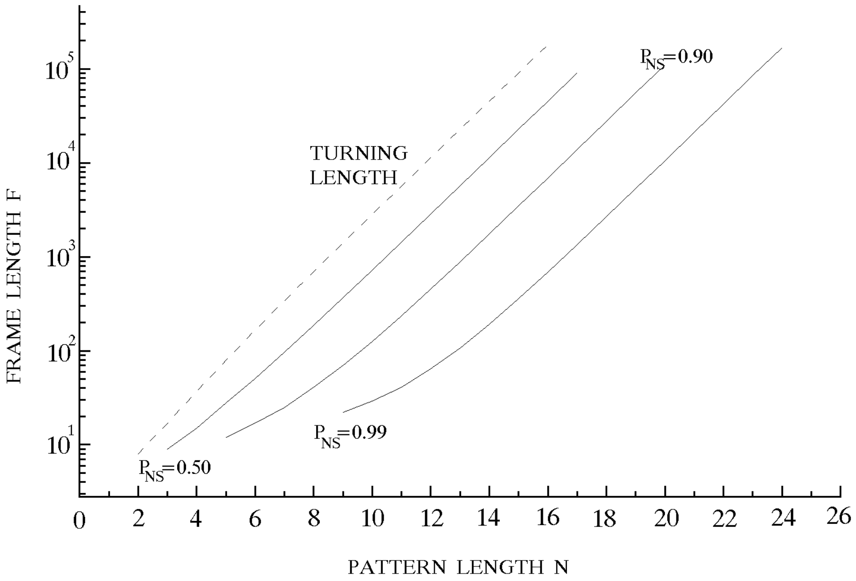

Figure 11. Frame lengths for which P_{NS} is less than specific value.

4. Conclusion

The derived formulae could be a useful tool for the design and optimisation of new systems (for which the simulation study is tedious or even useless, due to its length) in order to achieve the desired P_{NS}. It might also be used for further researches considering the system reframe times.

Acknowledgments

The authors would like to thank the anonymous reviewers for their careful review and valuable comments.

REFERENCES

1. M. N. Al–Subbagh and E. V. Jones: *Optimum patterns for frame alignment.* IEE Proc. part F - Commun. Radar & Signal Processing, Vol. 135, No. 6, December 1988, pp. 594–603.
2. D. Bajić and D. Drajić: *Duration of search for a fixed pattern in random data: distribution function and variance.* Electronics Letters, Vol. 31, No. 8, April 1995, pp. 631–632.
3. D. Bajić and D. Drajić: *On the Optimal Frame Synchronisation Markers for Channels with Memory.* IEEE Trans. on Comm., Vol. 43, No. 2/3/4, 1995, pp. 1326–1328.

4. CCITT RECOMMENDATION: *Blue Book III.4.*, Genève 1988

5. D. W. CHOI: *Frame Alignment in a Digital Carrier System – A Tutorial.* IEEE Comm. Magazine, Vol. 28, No. 2, February 1990, pp. 47–54.

6. H. HÄBERLE: *Frame synchronizing PCM systems.* Electrical Communications, Vol. 44, No. 4, 1969, pp. 280–287.

7. T. NIELSEN: *On the Expected Duration of a Search for a Fixed Pattern in Random Data.* IEEE Trans. on Information Theory, Vol. IT-19, No. 5, September 1973, pp. 702–704.

8. T. NIELSEN: *A Note on Bifix–free Sequences.* IEEE Trans. on Information Theory, Vol. IT-19, No. 5, September 1973, pp. 704–706.

Appendix A

The following sums are necessary for further evaluations (bearing in mind that $L \cdot p = 1,\ b = p^{N-1},\ a_1 = 1$ and $h_0 = h_N = 1$)

$$\begin{aligned}
X &= \sum_{i=1}^{N}(Lh_{N+1-i} - h_{N-i})p^i \\
&= (Lh_N - h_{N-1})\frac{1}{L} + (Lh_{N-1} - h_{N-2})\frac{1}{L^2} + \ldots + (Lh_1 - h_0)\frac{1}{L^N} \\
&= h_N - \frac{h_{N-1}}{L} + \frac{h_{N-1}}{L} - \frac{h_{N-2}}{L^2} + \frac{h_{N-2}}{L^2} - \ldots - \frac{h_1}{L^{N-1}} + \frac{h_1}{L^{N-1}} - \frac{h_0}{L^N} \\
&= h_N - \frac{h_0}{L^N} \\
&= 1 - p^N
\end{aligned}$$

$$\begin{aligned}
Y &= \sum_{i=1}^{N} i(Lh_{N+1-i} - h_{N-i})p^i \\
&= h_N - \frac{h_{N-1}}{L} + \frac{2h_{N-1}}{L} - \frac{2h_{N-2}}{L^2} - \ldots - \frac{(N-1)h_1}{L^{N-1}} + \frac{Nh_1}{L^{N-1}} - \frac{Nh_0}{L^N} \\
&= \sum_{i=0}^{N-1} h_{N-i}p^i - Np^N
\end{aligned}$$

$$(A1)$$

$$\begin{aligned}
Z &= \sum_{i=1}^{N} i^2(Lh_{N+1-i} - h_{N-i})p^i \\
&= h_N - \frac{h_{N-1}}{L} + \frac{4h_{N-1}}{L} - \frac{4h_{N-2}}{L^2} + \frac{9h_{N-2}}{L^2} - \ldots - \frac{N^2h_0}{L^N} \\
&= \sum_{i=0}^{N-1} (2i+1)h_{N-i}p^i - N^2p^N
\end{aligned}$$

Using (2), the following can be written:

$$\begin{aligned} ba_1p &= ba_1p \\ ba_2p^2 &= (Lh_N - h_{N-1})ba_1p^2 \\ &\vdots \\ ba_ip^i &= (Lh_N - h_{N-1})ba_{i-1}p^i + \ldots + (Lh_1 - h_0)ba_{i-N}p^i \\ &\vdots \end{aligned} \qquad (A2)$$

Left sides of equations $(A2)$ represent the terms of the infinite summation (4), so that their sum equals to S, being equal to the sum of the right sides of $(A2)$ as well,

$$S = ba_1p + XS = p^N + (1-p^N)S \qquad \Rightarrow S = 1. \quad Q.E.D. \qquad (A3)$$

For the expected value, similar summation can be performed, yielding

$$T_R = ba_1p + XT_R + YS \Rightarrow T_R = \sum_{i=0}^{N} h_iL^i - N. \quad Q.E.D. \qquad (A4)$$

The similar procedure, applied to the evaluation of the second moment $E\{k^2\}$, gives

$$\begin{aligned} &E\{k^2\} = ba_1p + XE\{k^2\} + 2YT_R + ZS \\ &\Rightarrow E\{k^2\} = T_R^2 + (T_R + N)(T_R + N - 1) - 2\sum_{i=1}^{N} ih_ip^i. \end{aligned} \qquad (A5)$$

Subtraction of T_R^2 from $(A5)$ yields (6).

Appendix B

If the starting shift S equals 1 (the first position after the real sync pattern position), for a bifix–free pattern the following may be written

$$\begin{aligned} P\{1\} &= P\{2\} = \ldots = P\{N-1\} \\ &= P\{F-N+1\} = \ldots = P\{F-1\} = 0. \end{aligned} \qquad (B1)$$

As the pattern will certainly be found at position F (if not earlier), it may be written

$$\sum_{k=1}^{F} P\{k\} = 1 \Rightarrow P\{F\} = P_{NS} = 1 - \sum_{k=N}^{F-N} P\{k\}. \qquad (B2)$$

Each of the $P\{k\}$ has factor p^N, so, regardless to a_k, the sum $\sum_{k=N}^{F-N} P\{k\}$ can be neglected for bigger N (longer sync patterns), therefore, in these cases, $P_{NS} \approx 1$.

On the other hand, for the case of all–zero patterns,

$$P\{1\} = 0.5, \qquad P\{2\} = \ldots = P\{N-1\} = 0. \tag{$B3$}$$

(B3) Sum $Q = \sum_{k=N}^{F-2} P\{k\}$ can be neglected for bigger N for the same reasons as the sum in Eq. $(B2)$.

It is interesting to evaluate the probability that the sync pattern will be simulated one bit before its real position. This probability may be evaluated as follows

$$\begin{aligned} P\{F-1\} &= Pr\{\text{not simulated earlier}\} \cdot 0.5 \\ &= (1 - P\{1\} - Q) \cdot 0.5 = 0.25 - 0.5Q. \end{aligned} \tag{$B4$}$$

At last,

$$\begin{aligned} P\{F\} = P_{NS} &= 1 - P\{1\} - Q - P\{F-1\} \\ &= 1 - 0.5 - Q - 0.25 + 0.5Q = 0.25 - 0.5Q < 0.25. \end{aligned} \tag{$B5$}$$

For bigger N, $P_{NS} \approx 0.25$.