Practical Frame Synchronization for Data with Unknown Distribution on AWGN Channels

Marco Chiani, *Senior Member, IEEE*, and Maria G. Martini, *Member, IEEE*

Abstract— We study practical frame synchronization based on sequential search of sync markers: these synchronizers can be used for frames of both known or unknown lengths. First, according to the hypothesis testing theory, a new simple metric is derived for the Additive White Gaussian Noise (AWGN) channel, based on the generalized likelihood ratio test, thus without any assumption on the statistical distribution of the data symbols. An analytical performance evaluation is then given, with closed form expression for the characteristic function of the decision variable. It is shown that a remarkable gain is obtained with respect to correlation based tests.

Index Terms— Frame synchronization, hypothesis testing, detection, GLRT, synchronization patterns.

I. INTRODUCTION

FRAME synchronization is a critical point in digital communication. Often, synchronization markers (or sync words, SW's) are inserted in the bitstream to facilitate synchronization at the receiver end, after channel impairments have occurred [1]–[3]. In some applications these SW's delimit blocks of data with variable lengths, as in the case of some source encoders employing variable length coding, such as, e.g., MPEG-4. A performance evaluation of frame synchronization in the general case of unknown packet lengths is thus of interest.

The derivation of the optimal detection procedure over AWGN channels has been given in [4] for packets of fixed, known lengths and data with known distribution. The extension of the optimal approach [4] to non binary modulation schemes and to frames of known but not necessarily constant lengths can be found in [5], with performance evaluated by simulation.

However, a simpler and common engineering approach for frame synchronization consists in a sequential detection algorithm: in this case, starting from a given position, the correlation between the received (continuous valued) samples and the SW symbols is computed, and compared with some threshold. If the threshold is exceeded the frame synchronizer declares a SW presence, otherwise the search continues [2], [3]. Note that detection algorithms based on sequential search of the SW can be also used for frames of unknown lengths.

In this letter, we derive the suitable metric for sequential frame synchronization algorithm according to the generalized

likelihood ratio test (GLRT) theory; then, we provide the analytical evaluation of the performance in terms of probability of detection and of false alarm, shown also as Receiver Operating Characteristics (ROC).

By comparison among the ROC's it is shown that the new metric provides a high gain with respect to tests based on correlation.

II. PROBLEM STATEMENT

We consider binary antipodal signaling, where the ith transmitted bit $b_i \in \{0, 1\}$ gives rise, after transmission through the AWGN channel, matched filter reception and ideal sampling, to a sample $r_i = (-1)^{b_i} + n_i$, where n_i are independent identically distributed (i.i.d.) Gaussian random variables (r.v.'s), each with zero mean and variance σ^2. The transmitted bits (0 or 1), or, equivalently, the corresponding antipodal symbols ($+1$ or -1) can either represent data or the SW sequence. More precisely, the data stream composed of i.i.d. symbols $d_i \in \{-1, +1\}$ is subdivided into frames of unknown lengths, delimited by N binary symbols SW's, $(c_1, ..., c_N)$, with $c_i \in \{-1, +1\}$. We do not make here any further assumption about the data symbols d_i, that are therefore to be considered with unknown probability distribution.
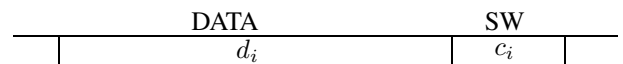
DATA		SW	
	d_i	c_i	

Fig. 1. Frame Structure

We assume that the synchronizer works as follows: starting from a position k, it observes a vector of N subsequent samples; based on the vector it decides if the SW is in position k; if not, it moves to the position $k+1$, repeating the steps. In this paper, we study the problem of deciding at each position k of the bitstream whether a sync word is present or not. The relation between this problem and other performance metrics such as, e.g., the probability of acquisition in one pass, is addressed in [2].

We treat the case as an hypothesis testing problem. After observing N subsequent samples, the synchronizer must choose between two possible alternatives:

$$\mathcal{H}_0 \quad : \quad r_i = d_i + n_i, \quad i = 1, \dots, N \quad (1)$$
$$\mathcal{H}_1 \quad : \quad r_i = c_i + n_i, \quad i = 1, \dots, N,$$

the first hypothesis, $\mathcal{H}_0$, representing the case where the observed samples are due to data in AWGN, the second, $\mathcal{H}_1$, corresponding to the observation of the sync word in

AWGN. In the design and analytical performance evaluation of the frame synchronizer we here neglect the case of "mixed data", i.e. when both data and SW symbols are present in the evaluation window. In fact, if the SW is properly chosen with low aperiodic autocorrelation properties (e.g. Barker sequences [1] [2]), its symbols should mimic random data, with, moreover, the additional property that some configurations can be avoided. Simulation results are shown later for the case of "mixed data" to support this approximation. In particular we will show that if the SW's are long enough, the case $\mathcal{H}_0$ of purely random data represents generally a worst case in terms of probability of false alarm respect to the "mixed data" case.

III. DERIVATION OF THE GLRT FOR FRAME SYNCHRONIZATION

Since we are not making any assumption about the probability distribution of data symbols d_i, we resort to the generalized likelihood ratio test (GLRT) [6] for the design of the detection strategy; the likelihood ratio test would require in fact the knowledge of the data distribution. This approach leads in principle to a two steps procedure: first, we estimate the unknown data vector[1] $\mathbf{d} = (d_1, \ldots, d_N)$ in the hypothesis $\mathcal{H}_0$; then, we use the estimate $\widehat{\mathbf{d}}$ as if it were the correct value of the transmitted data. In our case the maximum likelihood estimate of d_i is $\widehat{d}_i = \text{sign}(r_i)$. By indicating with $\mathbf{r} = (r_1, \ldots, r_N)$ the vector of received samples and with $f_X(.)$ the probability density function (p.d.f.) of a generic r.v. X, the GLRT is

$$\Lambda_g'(\mathbf{r}) = \frac{f_{\mathbf{R}|\mathcal{H}_1}(\mathbf{r})}{f_{\mathbf{R}|\mathcal{H}_0, \mathbf{d}}(\mathbf{r}|\widehat{\mathbf{d}})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda', \quad (2)$$

where $\mathbf{R} = (R_1, \ldots, R_N)$ is the vector of r.v.'s corresponding to the vector $\mathbf{r}$ of received samples and λ' is the selected threshold. From (1), the p.d.f.'s in (2) are Gaussian: by indicating with $p_G(v; \mu, \sigma^2) = 1/\sqrt{2\pi\sigma^2} \exp\{-(v-\mu)^2/(2\sigma^2)\}$ the Gaussian p.d.f. with argument v, mean μ and variance σ^2, we have $f_{\mathbf{R}|\mathcal{H}_1}(\mathbf{r}) = \prod_{i=1}^N p_G(r_i; c_i, \sigma^2)$ and $f_{\mathbf{R}|\mathcal{H}_0, \mathbf{d}}(\mathbf{r}|\mathbf{d}) = \prod_{i=1}^N p_G(r_i; d_i, \sigma^2)$. So, by substituting these expressions in (2) with $\widehat{d}_i = \text{sign}(r_i)$, the logarithmic version of (2) gives the GLRT as

$$\Lambda_g = \Lambda_g(\mathbf{r}) = \sum_{i=1}^N \left(|r_i| - r_i c_i\right) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \lambda, \quad (3)$$

where $\lambda \propto \ln(\lambda')$. Note that (3) can be interpreted as twice the sum of the absolute value of the samples r_i whose signs differ from that of the SW symbols c_i.

The probability of false alarm or emulation, P_{EM}, is given by

$$P_{EM} = \Pr\{\Lambda_g < \lambda | \mathcal{H}_0\} = F_{\Lambda_g | \mathcal{H}_0}(\lambda), \quad (4)$$

where λ is the selected threshold in (3) and $F_{\Lambda_g | \mathcal{H}_l}(.)$ is the cumulative distribution function (c.d.f.) of Λ_g under the hypothesis $\mathcal{H}_l$, $l = 0, 1$. Similarly, the probability of missed detection, P_{MD}, is given by

$$P_{MD} = \Pr\{\Lambda_g > \lambda | \mathcal{H}_1\} = 1 - F_{\Lambda_g | \mathcal{H}_1}(\lambda), \quad (5)$$

and the probability of correct detection is $P_D = 1 - P_{MD}$.

[1]In the following we use capital letters to indicate random variables and bold for vectors.

IV. ANALYTICAL EVALUATION OF THE GLRT PERFORMANCE

We first observe that, since $|c_i| = 1$, we can equivalently write the metric in (3) as

$$\Lambda_g = \sum_{i=1}^N |r_i c_i| - r_i c_i = \sum_{i=1}^N z_i, \quad (6)$$

where $z_i = |\psi_i| - \psi_i$ with $\psi_i = r_i c_i$. Hence,

$$z_i = \begin{cases} 0 & \text{if } \psi_i \geq 0, \\ -2\psi_i & \text{otherwise.} \end{cases} \quad (7)$$

From (7), indicating with $f_{\Psi_i | \mathcal{H}_l}(.)$ the p.d.f. of Ψ_i under the hypothesis $\mathcal{H}_l$, $l = 0, 1$ and by using the transformation of random variables, the p.d.f. of Z_i is

$$f_{Z_i | \mathcal{H}_l}(z) = \frac{1}{2} f_{\Psi_i | \mathcal{H}_l}\left(-\frac{z}{2}\right) u(z) + \Pr\{\Psi_i \geq 0 | \mathcal{H}_l\} \delta(z) \quad (8)$$

where $u(\cdot)$ is the unitary step function and $\delta(\cdot)$ is the Dirac delta function.

We now recall that the characteristic function (ch.f.) of the sum of independent random variables is the product of the ch.f.'s of the single terms. As we are interested in the distribution of $\Lambda_g(\mathbf{R})$ defined in (6) that is the sum of N independent r.v.'s each with p.d.f. (8), it is convenient to evaluate the ch.f. of the r.v. Z_i, here defined as $\Phi_{Z_i | \mathcal{H}_l}(\nu) \triangleq \mathbb{E}\left[e^{j2\pi\nu Z_i}\right]$, j being the imaginary unit. By using (8) we obtain

$$\Phi_{Z_i | \mathcal{H}_l}(\nu) = \mathbb{E}\left[e^{j2\pi\nu Z_i}\right] \quad (9)$$
$$= \Pr\{\Psi_i \geq 0 | \mathcal{H}_l\} + \frac{1}{2} \int_0^\infty f_{\Psi_i | \mathcal{H}_l}\left(-\frac{z}{2}\right) e^{j2\pi\nu z} dz.$$

Then, the ch.f. of $\Lambda_g(\mathbf{R})$ is simply

$$\Phi_{\Lambda_g | \mathcal{H}_l}(\nu) = \prod_{i=1}^N \Phi_{Z_i | \mathcal{H}_l}(\nu). \quad (10)$$

We now specialize (9) to the two hypotheses $\mathcal{H}_0, \mathcal{H}_1$.

A. Case $\mathcal{H}_0$

We first consider the case where the sync word is not present and the metric is thus applied to random data. Hence, the conditional r.v. $R_i | d_i$ has p.d.f. $f_{R_i | d_i}(r) = p_G(r; d_i, \sigma^2)$. Denoting $\Pr\{d_i = 1\} = \alpha$ and $\Pr\{d_i = -1\} = \beta = 1 - \alpha$ we have therefore $f_{R_i}(r) = \alpha p_G(r; 1, \sigma^2) + \beta p_G(r; -1, \sigma^2)$. For the r.v. $\Psi_i = R_i c_i$, from basic probability theory we get

$$f_{\Psi_i | \mathcal{H}_0}(\psi) = \alpha p_G(r; c_i, \sigma^2) + \beta p_G(r; -c_i, \sigma^2), \quad (11)$$

and $\Pr\{\Psi_i \geq 0 | \mathcal{H}_0\} = K_0 \triangleq \alpha + (1/2 - \alpha)\text{erfc}\left(c_i/\sqrt{2\sigma^2}\right)$. Substituting in (9) and by using the identity [7, eq. (7.4.2)], we can express the ch.f. in closed form as:

$$\Phi_{Z_i | \mathcal{H}_0}(\nu) = \frac{1}{2} e^{-4\pi\nu(jc_i + 2\pi\nu\sigma^2)} \cdot \left[\alpha \, \text{erfc}\left(\frac{c_i - j4\pi\nu\sigma^2}{\sqrt{2}\sigma}\right) \right.$$
$$\left. + \beta \, e^{j8c_i\pi\nu}\text{erfc}\left(\frac{-c_i - j4\pi\nu\sigma^2}{\sqrt{2}\sigma}\right)\right] + K_0. \quad (12)$$

Note that for equiprobable data symbols $\alpha = \beta = K_0 = 1/2$ and, as expected due to the symmetries of the r.v.'s involved, (12) is the same for $c_i = 1$ and $c_i = -1$. In other words, for
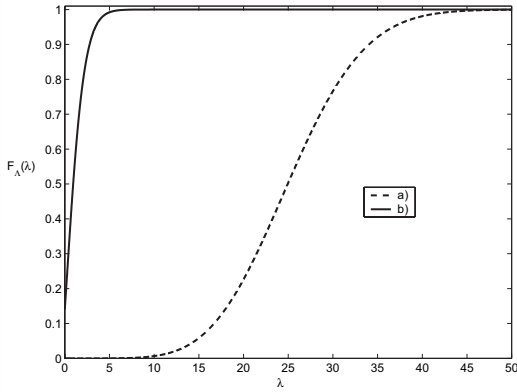
Fig. 2.   Probability of emulation (a) and of correct detection (b) as a function of the threshold λ. Sync word with length $N = 24$, $\sigma^2 = -3dB$.

$\mathcal{H}_0$ and equiprobable data symbols the particular SW pattern does not play any role. On the other side, SW patterns must be properly designed for the mixed data case, as previously discussed.

B. Case $\mathcal{H}_1$

We here consider the case where the sync word is present. Under this hypothesis, the generic R_i is Gaussian distributed, with mean c_i and variance σ^2. Thus, whatever c_i is, the r.v. Ψ_i has p.d.f.

$$f_{\Psi_i|\mathcal{H}_1}(\psi) = p_G(\psi; 1, \sigma^2), \qquad (13)$$

and consequently $\Pr\{\Psi_i \geq 0|\mathcal{H}_1\} = K_1 \triangleq \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2}\sigma}\right)$. Hence, from (9) and (13) we get:

$$\Phi_{Z_i|\mathcal{H}_1}(\nu) = K_1 + \frac{1}{2}e^{-4\pi\nu(j+2\pi\nu\sigma^2)}\text{erfc}\left(\frac{1 - j4\pi\nu\sigma^2}{\sqrt{2}\sigma}\right). \quad (14)$$

In conclusion, expressions (12) and (14) substituted in (10) give in closed form the ch.f. of the decision variable. From it, the p.d.f. and c.d.f. can be simply obtained by means of a single integral, that can be efficiently evaluated for instance by Fast Fourier Transform (FFT) methods.

As an example of application of the previous analysis, we report in Fig. 2 the cumulative distribution functions of the metric under the hypothesis $\mathcal{H}_0$, i.e. the probability of emulation, and under the hypothesis $\mathcal{H}_1$, i.e., the probability of correct detection $P_D = 1 - P_{MD}$.

In the figure we assumed equiprobable data symbols, i.e., $\alpha = \beta = 1/2$, a SW of $N = 24$ bits and additive Gaussian noise with $\sigma^2 = -3dB$.

A representation of the results in the same conditions is given in terms of ROC in Fig. 3, where the probabilities of detection, P_D, and of false alarm, P_{EM}, are reported on the axes. Moving along each curve, different values for the threshold λ are considered. For comparison, curves for the correlation based tests with unquantized (soft) or binary quantized (hard) received samples are also reported [8], together with simulation points. Moreover, we report simulation results for the "mixed data" case, where the test is applied to a vector composed of random data in the first (last) j positions,
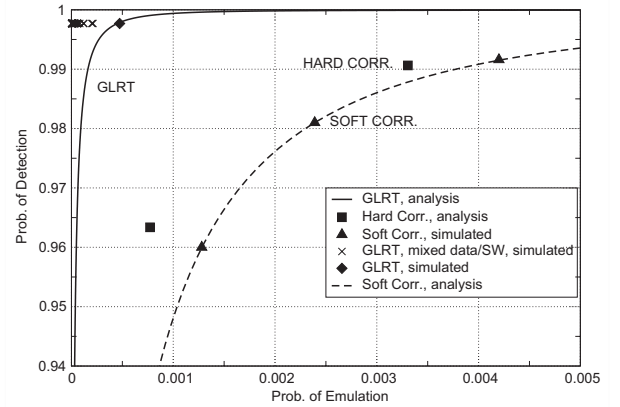

Fig. 3.   Receiver Operating Characteristic (ROC) curves for GLRT, hard decisions correlation and soft correlation synchronizers, $N = 24$, $\sigma^2 = -3dB$. Symbols indicate simulation results. The SW pattern for the "mixed data" case (in octal) is 76571440.

followed (preceded) by the shifted versions of the SW for the remaining $N - j$ positions, with $j = 1, \ldots, N - 1$. The SW pattern chosen for the simulation in the "mixed data" case is (in octal) 76571440 [2]. These and other unreported simulation results show that "mixed data" is generally better than pure random data in terms of P_{EM}, justifying the approach used in Section II.

Regarding the performance of the proposed test, we can observe that the new metric greatly outperforms both the correlation metrics, for any value of P_{EM}.

V. CONCLUSIONS

We studied sequential frame synchronization for data with unknown distribution. According to the hypothesis testing approach, the generalized likelihood ratio test has been derived. Moreover, the analytical performance evaluation of the resulting GLRT synchronizer has been provided, showing that a high gain can be obtained with respect to correlation based synchronizers. Results have been also validated through simulation.

REFERENCES

[1] R. H. Barker, "Group synchronization of binary digital systems," in *Communication Theory*. London, England: W. Jackson, 1953.
[2] R. A. Scholtz, "Frame synchronization techniques," *IEEE Trans. Commun.*, vol. 28, pp. 1204–1213, Aug. 1980.
[3] C. Georghiades and E. Serpedin, *Communication Handbook*. CRC Press, 2002, ch. 18.
[4] J. L. Massey, "Optimum frame synchronization," *IEEE Trans. Commun.*, vol. 20, pp. 115–119, Apr. 1972.
[5] A. Kopansky and M. Bystrom, "Detection of aperiodically embedded syncronization patterns," in *Proc. IEEE International Symp. Inf. Theory (ISIT 2001)*, p. 135, June 2001.
[6] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*. New York: John Wiley & Sons, 1968.
[7] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington, D.C.: United States Department of Commerce, 1970.
[8] M. Chiani and M. G. Martini, "Optimum synchronization of frames with unknown, variable lengths on Gaussian channels," in *Proc. IEEE Global Telecomm. Conf.*, pp. 4087–4091, Nov. 2004.