Physics Letters A 379 (2015) 2597–2609

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Similarity matrix analysis and divergence measures for statistical detection of unknown deterministic signals hidden in additive noise

O. Le Bot [a,b,*,1], J.I. Mars [a,b], C. Gervaise [a,b,c]

[a] Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
[b] CNRS, GIPSA-Lab, F-38000 Grenoble, France
[c] Chaire CHORUS, Foundation of Grenoble Institute of Technology, 46 Avenue Félix Viallet, 38031 Grenoble Cedex 1, France

ARTICLE INFO

Article history:
Received 8 January 2015
Received in revised form 30 April 2015
Accepted 1 June 2015
Available online 3 June 2015
Communicated by C.R. Doering

Keywords:
Similarity matrix
Divergence measures
Signal detection
Receiver Operating Characteristic (ROC)

ABSTRACT

This Letter proposes an algorithm to detect an unknown deterministic signal hidden in additive white Gaussian noise. The detector is based on recurrence analysis. It compares the distribution of the similarity matrix coefficients of the measured signal with an analytic expression of the distribution expected in the noise-only case. This comparison is achieved using divergence measures. Performance analysis based on the receiver operating characteristics shows that the proposed detector outperforms the energy detector, giving a probability of detection 10% to 50% higher, and has a similar performance to that of a sub-optimal filter detector.

1. Introduction

Deciding whether a measured data sequence is noise only or contains a short deterministic fraction within the observation time is of greatest importance in several application fields, such as radar interception, underwater acoustic signal detection, and analysis of medical signals. The general framework of a signal detector is classical, as the detector has to choose between one of the following hypotheses:

- H_0: the measured signal is noise only: $x(t) = n(t)$
- H_1: the measured signal has a deterministic part hidden in additive noise: $x(t) = s(t) + n(t)$

where $n(t)$ is white Gaussian noise (WGN), and $s(t)$ is the deterministic signal to be detected. To solve this signal detection, a statistical test is computed on the data that are measured, and then compared to a detection threshold [1].

The choice of the statistical test and the estimation of its probability density functions (PDFs) under hypotheses H_0 and H_1 depend on the amount of *a-priori* knowledge we have about the signal we want to detect and about the noise that it contains. When the waveform of the signal to detect is fully known, the optimum statistical test is known as a matched filter [1]. For the opposite situation, when the waveform of the deterministic signal is not known, classical detectors are usually based on signal energy [1] or on high-order statistics [2,3], and perform non-Gaussianity tests. Also, there are several approaches that can be used to set the detection threshold, including the Neyman–Pearson method, the Bayes' criterion, the maximum *a posteriori*, and the false discovery rate [1].

This Letter aims to present a new detection scheme using an approach that was inspired by recurrence plots [4] and is combined with divergence measures, to detect short (few tens to hundreds of samples) unknown deterministic signals in additive WGN. Recurrence plots were introduced to study the stationarity of nonlinear dynamical systems [4], and have been shown to be useful for a large set of applications, like geology [5], climatology [6], music [7] and analysis of medical signals [8], to name but a few. As recurrence plots show different patterns that depend on the dynamic of the system (*i.e.*, random, periodic, chaotic), several approaches have been presented in the literature to quantify and distinguish between these three different dynamical behaviors, and particularly for deterministic signals in random process [9–15]. A common point to all of these recurrence plot studies is their use of what is known as recurrence quantification analysis (RQA) [8,16,17] to decide whether the measured signal is noise or not. Thus, a classical detection scheme in the recurrence plot community can be summed-up as follows:

* Corresponding author.
E-mail address: lebotol@gmail.com (O. Le Bot).
[1] Also at: Pôle STIC, ENSTA Bretagne (Université Européenne de Bretagne), 2 rue Francois Verny, 29806 Brest Cedex 9, France.

http://dx.doi.org/10.1016/j.physleta.2015.06.004
0375-9601/

$$x(t) \longrightarrow SM \longrightarrow RP \longrightarrow RQA \longrightarrow Detector \quad (1)$$

where *SM* represents the similarity matrix, and *RP* the recurrence plot. However, distributions of *RQA* metrics under hypotheses H_0 and H_1 do not generally follow existing distributions, and finding analytic expressions for these latter is not straightforward [15].

Instead of using RQA, we restrict our detector to only the use of the similarity matrix, which is sometimes called the *distance matrix* or *distance plot* in the literature [15]. The similarity matrix is the intermediate matrix that is obtained before applying the recurrence threshold that leads to the recurrence plot. Thus, we avoid the choice of this recurrence threshold and our detection scheme comes down to:

$$x(t) \longrightarrow SM \longrightarrow Detector \quad (2)$$

Our detector compares the empirical distribution of the similarity matrix coefficients of a measured signal with the distribution that is expected if the measured signal is WGN. The expression of this expected distribution can be derived analytically more easily than the RQA distribution. The comparison between the empirical and the analytic distributions is carried out with a goodness-of-fit test that is based on statistical divergences [18].

Overall, the detector presented in this Letter follows the same scheme as that proposed by Michalowicz [19]. Our algorithm differs from that of Michalowicz [19] in the use of divergence measures instead of a modified version of the χ^2 test to compare the analytic and the empirical distributions of the similarity matrix coefficients. Classical χ^2 test cannot be used because the coefficients of the similarity matrix are not fully independent of each other, as demonstrated by Michalowicz [19], which can bias the result of the test by giving much more false-positive detection than expected [19]. Finally, we do not compute the similarity matrix with a Euclidean norm only, as we propose the use of Pearson's correlation coefficient and the dot-product for this purpose [20].

After a brief recall of the recurrence plot method, we describe the different steps of our detection algorithm. Strong emphasis is put on derivation of the analytic distributions of the similarity matrix coefficients under hypothesis H_0, when the Euclidean norm, Pearson's correlation coefficient, and the dot-product are used to compute the similarity matrix. Then, we discuss the choice of an appropriate divergence function to compare the analytic and empirical distributions. The third part presents the performances of our detector through the use of receiver operating characteristic (ROC) curves. Three different deterministic signals are used in this part: a periodic signal, a chaotic Rössler system and a real acoustic signal. The influence of the degrees of freedom involved in our detection scheme are also investigated, such as the choice of the similarity function or the divergence measure. The performance of the proposed detector is compared with that of an energy detector, a sub-optimal filter detector and the optimal matched-filter detector, which are commonly used in signal processing.

2. Recurrence plots

Recurrence plots were introduced to study complex systems and are aimed at visualizing the recurrence of their phase space trajectory [4]. Transforming a data sequence to a recurrence plot representation involves three main steps.

First, the phase space trajectory of the measured signal $x(i)$ ($i = 1, \ldots, N$) is reconstructed using the time delay embedding method [21,22]. Each phase space vector is given by:

$$\overrightarrow{x_m(i)} = [x(i),\ x(i+\tau),\ \ldots\ ,\ x(i+(m-1)\tau)] \quad (3)$$

where m is the embedding dimension, and τ is the delay.

The second step consists of measuring the level of similarity between two vectors of the phase space trajectory: $\overrightarrow{x_m(i)}$ and $\overrightarrow{x_m(j)}$. Calculating the similarity between all of the possible pairs of phase space vectors leads to the similarity matrix that is defined by:

$$d(i, j) = Sim\left(\overrightarrow{x_m(i)}\ ,\ \overrightarrow{x_m(j)}\right) \quad (4)$$

where $Sim(\ .,\ .)$ is the function that is chosen to study the likeness of the phase space vectors. A lot of different mathematical functions can be used for this step. Spatial distances, and particularly the Euclidean norm, are mostly used for this purpose by the recurrence plot community [23]. In this Letter, we will introduce new functions, *i.e.*, Pearson's correlation coefficient and the dot-product, which are common similarity measures in signal processing, but not in the recurrence plot community.

Finally, as the recurrence plot is obtained through the comparison of each coefficient of the similarity matrix to a threshold, the recurrence plot is a binary matrix where the coefficient of index $(i,\ j)$ is 1 if $\overrightarrow{x_m(i)}$ and $\overrightarrow{x_m(j)}$ are considered as similar, and is 0 otherwise.

3. Method

3.1. Overview of the signal detection scheme

The signal detection scheme must give an answer that allows us to decide whether a finite sequence of discrete samples contains a deterministic signal or noise only. After calculating Eq. (3) and Eq. (4), the PDF of the similarity matrix coefficients is built. This PDF is expected to fit a given theoretical PDF if the measured signal is only WGN. We use a divergence measure to compare the theoretical expected PDF under hypothesis H_0 with the empirical PDF associated with the similarity matrix of the measured signal. We recall that in probability theory, a divergence measure is a mathematical function that quantifies the distance between two probability distributions. The result of the divergence measure is a positive number D that we compare with a detection threshold λ. If D is below this threshold, this means that the distributions look alike, and consequently that the measured signal is WGN. For the opposite, *i.e.*, if D is greater than the threshold, this means that the empirical PDF differs from the theoretical noise PDF, and thus that a deterministic signal is present. The threshold λ is chosen according to the Neyman–Pearson criterion. We recall that when performing a hypothesis test between two hypothesis H_0 versus H_1, Neyman–Pearson criterion is the one that maximizes the probability of detection while guaranteeing a given probability of false alarm (Pfa). With other words, a threshold fixed by the Neyman–Pearson criterion maximizes the probability (Pd) of choosing hypothesis H_1 when H_1 is effectively true and rejects hypothesis H_0 with a probability Pfa when H_0 is effectively true. To apply this criterion, we use Monte-Carlo simulations to built the distribution of the divergence measures D between the analytic PDF expected under hypothesis H_0 and the empirical PDF of the similarity matrix coefficients of finite length WGN. All of the steps of this detection scheme are summarized in Fig. 1.

3.2. Analytical distribution of the similarity matrix coefficients in the 'noise only' case

3.2.1. Hypothesis

Under hypothesis H_0, we assume that the measured samples $x(1), x(2), \ldots, x(n)$ from a given sequence are independent Gaussian random variables with zero mean and variance σ^2.

To obtain the similarity matrix, we look at the similarity between the vectors $\overrightarrow{x_m(i)} = [x(i),\ x(i+\tau),\ \ldots\ ,\ x(i+(m-1)\tau)]$ and $\overrightarrow{x_m(j)} = [x(j),\ x(j+\tau),\ \ldots\ ,\ x(j+(m-1)\tau)]$ (with $i \neq j$), the