Matching of quasi-periodic time series patterns by exchange of block-sorting signatures

Bachir Boucheham *

University of Skikda, 20 Août 1955, Department of Informatics, BP 26, Route El-Hadaek, Skikda DZ21000, Algeria

Received 15 June 2007; received in revised form 4 November 2007
Available online 21 November 2007

Communicated by R.C. Guido

Abstract

We propose a novel method for quasi-periodic time series patterns matching, through signature exchange between the two patterns. The signature is obtained through sorting of the time series on magnitude. The advantage is that the difficult task of comparing the two patterns can be easily performed as a result of this exchange: The original time series is compared (point to point matching) to the reconstructed time series obtained through the reverse process, using the other time series signature. The matching is such that periods in one time series are put into correspondence with periods in the other time series, even if the time series is of different basic patterns and/or different lengths. The method is simple to implement and requires no parameters. It was compared to the very appreciated DTW algorithm on execution time, space and accuracy. Due to the quasi-periodic nature of the electrocardiogram, the tests were performed on ECG traces, selected from the Massachusetts Institute of Technology – Beth Israel Hospital (MITBIH) public database. Results show that the proposed method outperforms DTW on all aspects. This suggests that our method could be a good alternative to the classical DTW technique for quasi-periodic signals comparison. Specific applications are foreseen for our method: Novelty detection and person identification using ECG.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Pattern matching; Time series; Block-sorting signature; Shape exchange; DTW

1. Introduction

Time series are measurements of different sorts of phenomena taken at regular intervals. For instance, we can cite natural phenomena (e.g. river levels), physiological phenomena (e.g. electrocardiogram), economical phenomena (e.g. stock exchange index) or artificial phenomena (e.g. traffic density in a computer network) as examples of useful time series. Analysis and interpretation of such time series are of great importance for experts in their respective fields. Common applications in that respect include Creation and querying of specialized databases, compression and storage or transmission of time series, and interpretation of time series patterns, including event and abnormal behavior detection. This study is exclusively dedicated to quasi-periodic time series. These are time series that are concatenations of quasi-similar patterns called pseudo periods (periods for short). Our interest in quasi-periodic time series is due to the fact that many phenomena are of quasi-periodic nature. We particularly focus the study on the electrocardiogram (ECG): A quasi-periodic physiological time series, intensively used in medical care. A typical electrocardiogram is a sequence of quasi-regular periods reflecting heart cycles. Each cycle is itself a concatenation of three consecutive basic patterns: P wave, QRS complex and T wave. Fig. 1 presents a two period ECG trace with its main patterns.

In previous works, we developed algorithms and methods for ECG baseline wandering detection and correction (Boucheham et al., 2005), QRS detection (Boucheham et al., 2003), short-term compression (Boucheham et al.,...
In the current work, we focus on resolving the pattern matching problem for this type of time series. This problem can be stated this way: Given two ECG time series, say $X$ and $Y$, find a method to decide whether they are similar. This is an important problem, the solution of which could have impact on many interesting ECG applications: Clustering, data-mining, novelty detection, automatic diagnosis, person identification, and perhaps other applications. We stress the point that, although this work is centered on ECG time series, the study is also suitable to other quasi-periodic time series, since we do use no priori knowledge on the nature of the data. For this reason, we indifferentely refer to the processed data by ECG or (quasi-periodic) time series.

In (Boucheham, 2007), we noticed that sorting of ECG signals on magnitude yields a kind of signature of the time series, in the sense that the obtained trace was a global and a stable characteristic of the used time series, in each test. This observation was behind the idea to use this signature to compare quasi-periodic time series patterns. This last operation would allow quantifying the similarity between two given time series patterns. However, it is desirable to derive a detailed matching using this signature. This desire is motivated by the observation that human experts can visually tell in all cases if two time series are similar or not. In case of dissimilarity, they can even indicate the regions where there is disagreement between the two patterns. In a sense, when comparing time series visually, we somehow try to compare periods in each time series. Therefore, ideally, periods in one time series should be put into correspondence with periods in the other time series, even if the number of periods, lengths or the sources of the two times series were different. We show that this operation is possible, no matter the number of periods in each pattern and no matter if the two patterns come from different sources or not. The comparison operation is made possible through exchange of the two time series respective signatures, while using the respective time indexes for each series. For this reason, we use the acronym SEA (Shape Exchange Algorithm) for our method. We also show that the very appreciated DTW method, as far as time series comparison is concerned, is unable to deal with all situations, whereas our method never fails. In addition, we show that the proposed SEA method is more accurate, faster and less memory consuming than DTW. We also illustrate some specific applications we foresee for our method.

The rest of this work is organized as follows. In Section 2, the problem of similarity measure for time series comparison is stated in a more detailed manner and related works are cited and commented. In particular, the DTW method is presented in brief in this section. In Section 3, the main materials and methods for the proposed solution are presented. Especially, the proposed SEA method is presented in this section. In Section 4, a thorough comparison SEA vs. DTW on ECG time series is presented and applications using the SEA method are illustrated. In Section 5, results of the conducted tests are discussed and concluding thoughts and future works are announced.

2. Similarity measure for time series matching

2.1. Problem statement

Given two time series $X = (x_i)$, $i = 1:n$ and $Y = (y_j)$, $j = 1:m$, the problem of comparing $X$ and $Y$ consists in proposing a similarity measure $d(X, Y)$ capable of quantifying the degree of resemblance between $X$ and $Y$. In other words, quantity $d(X, Y)$ should allow concluding if the two patterns ‘fit’ and how good the fit is. One of the earliest pattern matching techniques is the Euclidian distance. This measure is defined only for equal lengths time series ($n = m$). In this case, the Euclidian distance between $X$ and $Y$ is defined by Eq. (1)

$$\text{Euclidian}(X, Y) = \sqrt{\sum_{i=1}^{n} [x_i - y_i]^2}$$  \hspace{1cm} (1)

The advantage of the Euclidian distance is its simplicity. It also has an advantageous $O(n)$ computational complexity. However, this measure is not efficient in direct data comparison due to its sensitivity to small distortions in the time axis (Keogh and Pazzani, 2000; Keogh and Ratanamahatana, 2005). For instance, Fig. 2a illustrates the inefficiency of this measure on a small ECG time series. It is easy to derive from the figure that the time series S1 and S2 are similar. However, due to a shift in the time axis of one time series with respect to the other, the Euclidian distance would rather induce a very large dissimilarity measure because it assumes that the $i$th point in one series is aligned with the $i$th point in the other series. More generally, the problem with the Euclidian distance is its inability to deal with four basic difficulties encountered in time series matching: Noise (Fig. 3a), offset translation (Fig. 3b), amplitude scaling (Fig. 3c) and time axis scaling (Fig. 3d). Many methods have been developed to deal with these problems. Obviously, time scaling is the most difficult aspect in time series pattern matching.
2.2. Existing methods

Many authors addressed the problem of time series similarity measure, in a variety of fields in science and technology. These applications include financial data analysis (e.g. Agrawal et al., 1993), moving object tracking (Chen et al., 2003) and music retrieval (e.g. Zhu and Shasha, 2003), especially in the data-mining paradigm. The Discrete Fourier Transform (DFT) was one of the first techniques used to compare time series (Agrawal et al., 1993). This technique consists in selecting the first $k$ coefficients of the DFT of the two given time series and then in using the Euclidian distance for the comparison. This frequency domain method proved to be successful in many applications. However, the focus in this study is on time domain approaches, which generally lend themselves more easily to intuitive interpretation. One of these methods consists in transforming first the time series raw data to character strings over a limited discrete alphabet, then, in applying string matching algorithms (e.g. Shatkay and Zdonik, 1996). The main weaknesses of this approach are due to errors induced by quantification and to its sensitivity to noise. In another approach, Bozkaya et al. proposed a modified edit distance referred to by Longest Common Sub-Sequence (LCSS) (Bozkaya et al., 1997). This method requires a threshold value on the edit distance be set, which is difficult without a priori knowledge on the data, or through a learning process. Another approach consists in using histograms to compare time series. Although histograms are global data descriptors, their weakness consists in ignoring the temporal order of the data. Consequently, one histogram is in fact shared by a huge number of different time series. Chen and Özsu (2003) proposed a solution to this problem through multi-scale histograms by dividing the time series into many segments. But the problem is still persistent at these levels. In addition, sectioning of the time series may lead to fragmentation of important semantic information.

One of the most appreciated pattern matching techniques for time series is the Dynamic Time Warping method (DTW). Furthermore, this technique is reported to be ‘exceptionally difficult to beat in time series pattern matching’ (Xi et al., 2006). Particularly, the DTW method was successfully applied to speech recognition (e.g. Sakoe and Chiba, 1978; Berndt and Clifford, 1996). It was first used in the data-mining paradigm by Berndt and Clifford (Berndt and Clifford, 1994). The main advantage of the DTW technique is its high ability to take into account time axis scaling. However, this technique suffers mainly from high temporal complexity. In addition, for the specific case of quasi-periodic time series, we will illustrate the limitations of DTW with respect to our proposed method. DTW is presented in brief in the next sub-section.
2.3. Dynamic time warping (DTW)

To align two given time series \( X = (x_i), \ i = 1:n \) and \( Y = (y_j), \ j = 1:m \), using DTW, a \( n \times m \) cost matrix \( D \) is constructed, where \( D(i,j) \) is some conventional metric. For instance, the Euclidian distance is mostly used to compute matrix \( D \) (Eq. (2))

\[
D(i, j) = (x_i - y_j)^2
\]

(2)

Matrix \( D \) can be seen as the alignment cost of points \( x_i \) with points \( y_j, \ i = 1:n, j = 1:m \).

The second step in DTW is the determination of a warping path \( W \), a continuous sequence of elements \( w_s = (i,j), \ s = 1:K \), where \( K = |W| \), with the following constraints:

a. \( \max(n, m) \leq K \leq n + m - 1 \). This condition set bounds on the cardinality of \( W \).

b. \( w_1 = (1, 1) \) and \( w_K = (n, m) \). This condition forces the endpoints in each time series to match.

c. Let \( w_s = (u,v) \) and \( w_{s-1} = (q,r), \ s = 2:K \), then \( u - q \leq 1 \) and \( v - r \leq 1 \). This condition restricts the next of the current element to one of its adjacent neighbors.

d. Let \( w_s = (u,v) \) and \( w_{s-1} = (q,r), \ s = 2:K \), then \( u - q \geq 0 \) and \( v - r \geq 0 \). This completes condition (c) by restricting the next element to north, east or diagonal neighbor.

Even with all these restrictions, there exit many warping paths, but evidently, with different costs. Let \( f(W) \) be the cost of warping path \( W = (w_s), s = 1:K \). We would then be interested in the path with the minimum cost (Eq. (3))

\[
W_{\text{min}}(X,Y) = \text{Arg}\min f(W)
\]

(3)

where

\[
f(W) = \frac{1}{K} \sqrt{\sum_{s=1}^{K} \delta_s}
\]

(4)

In Eq. (4), \( \delta_s \) is the warping cost of element \( w_s = (u,v), \ s = 1:K \); thus \( \delta_s = D(u,v) \). This problem is formulated in a dynamic programming approach. Let \( \gamma(i,j) \) be the cumulative cost of warping path \( W = (w_s), s = 1:K \), at cell \( w_s = (i,j) \). Then, the minimum warping cost at cell \( w_s \) is given by Eq. (5)

\[
\gamma(i,j) = D(i,j) + \min(\gamma(i-1,j-1),
\gamma(i,j-1), \gamma(i-1,j)), \ i \geq 2, \ j \geq 2
\]

(5)

with \( \gamma(1,1) = 0 \). The minimum warping cost is then \( \gamma(n,m) \). An illustrating example of the warping path (of the case shown in Fig. 2b), realized by DTW, is given in Fig. 4. The matching between the two time series is given in Fig. 5, upper plots, and the errors due to the matching are reported in Fig. 5, lower plots.

Analysis of the DTW method shows that it has an \( O(nm) \) time and space complexities. It is clear that when
\( m \) and \( n \) are big enough, these complexities would be of great inconvenience for the DTW usage. A more detailed presentation of DTW can be found in (Kruskall and Liberman, 1983). Fig. 2b illustrates the nice work performed by this method to overcome the time axis shift problem and Fig. 6 shows an illustration of DTW in overcoming the time axis scaling problem. Note that, in this last case, the two time series are of different lengths.

Even though DTW is a very appreciated method in the context of time series pattern matching, in addition to its main negative aspect (high temporal complexity), in our view, this technique is unable to deal with some situations that are specific to quasi-periodic time series. For illustration, Fig. 7 shows two ECG time series taken from the same record and the work performed this time by DTW. It is clear that, intuitively, one can state that the two patterns are similar and perhaps belong to the same record. However, Fig. 7 clearly illustrates one aspect that DTW cannot deal with, as far as quasi-periodic time series are concerned. We refer to this aspect by the phase shift problem. In Fig. 7, DTW has only matched pieces of the time series that appear in the same sequential order (the QRS complex in this case). However, patterns that do not appear in the same order in both traces, and although present, could not be correctly matched by DTW. The reason for this weak aspect is due to the matching strategy used by DTW. It matches time series in the temporal order. But since quasi-periodic signals are concatenations of quasi-similar patterns, then, as long as the two time series to match are longer than one period, there should be a way to decide whether they are similar. The situation is better seen through the reconstruction of each trace using the other trace matched-magnitude at each temporal index. This illustration is reported in Fig. 8 for the case shown in Fig. 7, where the upper plots report the original time series \( (X \) and \( Y) \), the middle plots report the original time series \( (X \) and \( Y), \) discontinuous lines) versus the reconstructed time series \( (X_{\text{Rec}} \) and \( Y_{\text{Rec}}, \) continuous lines). The matching errors are reported in the lower two plots \( (X - X_{\text{Rec}} \) and \( Y - Y_{\text{Rec}}) \). The plots clearly show that the matching is good at the QRS region, whereas it is not acceptable elsewhere. Particularly, the clinically significant \( P \) and \( T \) patterns have not been correctly matched. In another context (time series clustering), Oates et al. (2001) noticed that the output of DTW contains mistakes. This led them to propose a hybrid similarity model based on combining DTW and HMMs. However, their method deals with general time series. For the specific case of quasi-periodic time series and, due to their quasi-similar basic patterns, no matter the lengths and the onsets/offsets of given segments \( X \) and \( Y), \) methods should be developed to answer questions like Is time series \( X \) of the same nature than time series \( Y \)

---

**Fig. 6.** Illustration of DTW pattern matching: A remarkable solution to the time scaling problem. Although the two time series are ECG traces coming from two different sources and are of different lengths, DTW correctly matched the subpatterns.

**Fig. 7.** Phase shift problem: A problem that DTW cannot deal with. Notice that, although the series are similar quasi-periodic traces (ECGs), only pieces of the two traces that appear in the same (sequential) order, the QRS complex here, have been correctly matched. The other pieces \( (P \) and \( T \) waves here), and although present in both traces, have just been matched with sequentially equivalent parts.
(categorization by clustering)? Or, are time series $X$ and $Y$ similar enough (to conclude that they come from the same source, e.g. a person’s heart, in the case of ECG)? Such methods should use global and stable descriptors of the time series to be able to answer these questions.

3. Materials and methods

3.1. Distance between vectors of measures

Let $X = (x_i)_{i=1:N}$, $Y = (y_i)_{i=1:N}$ be two vectors of real measures. For the purpose of comparison of $X$ and $Y$, we use two of the most appreciated similarity–dissimilarity criteria: The normalized percent root difference (PRD) (Eq. (6)) and the correlation measure (Eq. (7))

$$\text{PRD}(X, Y) = \sqrt{\frac{\sum_{i=1}^{N} (x_i - y_i)^2}{\max(\sum_{i=1}^{N} (x_i - \bar{x})^2, \sum_{i=1}^{N} (y_i - \bar{y})^2)}} \times 100\%$$

$$\text{Corr}(X, Y) = \frac{\text{cov}(X, Y)^2}{\text{var}(X) \cdot \text{var}(Y)}$$

with $\text{cov}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$

and $\text{var}(X) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$

In Eq. (6), the numerator is the root of the sum of squared errors (SSE). In other words, this is the Euclidian distance between $X$ and $Y$. The denominator under the root is a normalization weight to unsure PRD values within the range 0%, for a perfect match between $X$ and $Y$ and 100% for a perfect mismatch. In Eq. (8), cov is the covariance of $X$ and $Y$, whereas in Eq. (9) var is the variance of $X$. In Eqs. (6), (8) and (9), $\bar{x}$ (respectively, $\bar{y}$) is the mean value of $X$ (respectively, $Y$). Note that the correlation factor is defined for equal lengths vectors, and that $-1 \leq \text{Corr}(X, Y) \leq +1$ always holds. Note also that the more $X$ and $Y$ are correlated (thus similar), the closer to +1 is $\text{Corr}(X, Y)$.

3.2. Block-sorting as a signature: The SEA method

3.2.1. SEA main steps

The proposed SEA method for quasi-periodic time series patterns matching is composed of two main steps:

Step 1: Signature establishment: Usually, a time series is specified by its magnitude components only, since it is usually sampled at regular intervals. The time indexes are in fact also components of the time series. Therefore, time series $X = (x_i)$, $i = 1:N$ is rewritten $X = (x_i, i)$, $i = 1:N$. In this notation, $i$ is the temporal index of magnitude value $x_i$. Sorting of $(x_i, i)_{i=1:N}$ on the first coordinate aims to consider the time series, not on the temporal order, but rather on the magnitude order. The outcome of this block-sorting operation is a set of time indexes and a trace of the sorted magnitudes. The obtained trace for $X$ is referred to in this study by signature ($X$). As stated above, this last trace possesses two important properties to be considered as a signature. First, it is a global descriptor, in the sense that it is the outcome of sorting the whole data. Second, it is stable, in the sense that sorting of segments of the same time series at different onsets/offsets always yields quasi-similar traces. What is new and important is that we argue that efficient comparison of the original time series can be performed based on direct data, using the time series signatures. This is a legitimate claim, since each signature is a characteristic of its corresponding time series. The comparison is explained in the next step.

Fig. 8. Illustration of DTW inability to deal with the phase shift problem. Upper plots: original $X$ and $Y$ time series, middle: reconstruction of each trace with the other trace matched-magnitude (continuous line: reconstructed trace, discontinuous line: original trace), lower plots: Matching errors. This illustration corresponds to the case shown in Fig. 7.
Step 2: **Shape exchange and comparison:** One can just proceed to compare the two signatures, through the Euclidian distance, for instance. This type of comparison can only tell how good the similarity between the time series patterns is. However, the signature ignores the temporal dimension of the time series. Therefore, as a global evaluation criterion of the similarity, the signature is unable to locate in time the existing dissimilarity between the time series patterns. In the second step of SEA, we propose to perform a direct matching of the time series patterns that can yield a detailed map of similarities/dissimilarities (point to point matching). This step consists in the exchange of the signatures between the two time series. In other words, time series \( X \) will ‘wear’ the shape of time series \( Y \) through its signature and vice versa. Since the signature is a characteristic of the time series, it is legitimate to expect from two similar time series to yield similar signatures. Therefore, it is also legitimate in this case to expect low distortions when reconstructing the two time series, following their respective signatures exchange. On the opposite side, it would just be legitimate to expect important distortions in the case of two non-similar time series when they are reconstructed, following their respective signatures exchange. The comparison is then performed for each time series (e.g. \( X \)) with its reconstructed one as a result of the exchange operation (e.g. \( X_{\text{Rec}} \)), using the PRD and Corr factors as objective criteria and visual inspection as a subjective criterion. In the following, we present the SEA method in a more formal way.

### 3.2.2. SEA algorithm

Let \( X = (V_{1...N}, P_{1...N}) \), and \( Y = (W_{1...M}, Q_{1...M}) \) be the time series to match, where

- \( V_{1...N} \) is the set of magnitude values of time series \( X \),
- \( P_{1...N} \) is the set of time indexes of time series \( X \),
- \( W_{1...M} \) is the set of magnitude values of time series \( Y \),
- \( Q_{1...M} \) is the set of time indexes of time series \( Y \).

Let also Sort-on-Magnitude be a procedure that sorts any input time series on the magnitude coordinate, and Sort-on-Temporal-Index the inverse procedure of Sort-on-Magnitude.

The SEA algorithm’s important steps are the following:

(a) **Sorting on Magnitude:**
- \( X' = (S1, P') \leftarrow \text{Sort-on-Magnitude}(X) \), \( S1 = \text{signature}(X) \), \( P' \): temporal index of \( S1 \).
- \( Y' = (S2, Q') \leftarrow \text{Sort-on-Magnitude}(Y) \), \( S2 = \text{signature}(Y) \), \( Q' \): temporal index of \( S2 \).

(b) **Normalization:** If \( N \neq M \), then \( X' \) and \( Y' \) are normalized as described below (Section 3.3).

(c) **Signature Exchange:**
- \( X'' = (S2, P') \): \( X'' \) uses magnitudes of \( Y \) and time indexes of \( X \).
- \( Y'' = (S1, Q') \): \( Y'' \) uses magnitudes of \( X \) and time indexes of \( Y \).

As can be noticed, there has been exchange of the signatures between the two time series, whereas the time indexes have been maintained.

(d) **Reconstruction and matching:**
Let \( X_{\text{Rec}} \) (respectively, \( Y_{\text{Rec}} \)) be the reconstructed time series as a result of reordering \( X'' \) and \( Y'' \) on their respective temporal index. Formally

\[
X_{\text{Rec}} = \text{Sort-on-Temporal-Index}(X''); \quad Y_{\text{Rec}} = \text{Sort-on-Temporal-Index}(Y'');
\]

### 3.3. Time series of different lengths

Using the same notations above, and assuming that \( |X| = n \neq |Y| = m \), the comparison is performed by first applying a linear mapping between the two signatures \( S1 \) and \( S2 \) and the corresponding time indexes \( P \) and \( Q \). In a sense, this is as though we force the longer time series to shrink to the length of the shorter time series and the shorter series to expand to the length of the longer series.

### 3.4. SEA time and space complexities

Analysis of the SEA algorithm shows that its main procedures are Sort-on-Magnitude, Sort-on-Temporal-Index, the PRD and the Corr procedures. It is known that sorting of an array of length \( k \) can be performed in a \( O(k \cdot \log_2(k)) \) to \( O(k^2) \), depending on the used sorting technique. The PRD and the Corr factors can both be computed in a linear time \( O(k) \). We tried more than one sorting procedure and did not notice significant change in the execution time. Thus, we chose the Sort-by-Selection procedure, which possesses a theoretical \( O(k^2) \) temporal complexity. The overall SEA temporal complexity is then \( O(2(n^2 + m^2) + 2(n + m)) \), which is of quadratic order. For the space complexity, the main data structure is basically a couple of vectors \( X(n) \), \( X_{\text{Rec}}(n) \), \( Y(m) \) and \( Y_{\text{Rec}}(m) \). The space complexity is then \( O(2(n + m)) \), which is of linear order.

### 4. SEA vs. DTW

In this section, we proceed to the comparison of SEA vs. DTW. The comparison is performed on the numerical aspect through the PRD (Eq. (6)) and the correlation factor Corr (Eq. (7)). The formulas are applied for both traces \((X, Y)\) with their respective reconstructed traces \((X_{\text{Rec}}\) and \(Y_{\text{Rec}}\)) and for both methods (SEA and DTW). Hence, we compute for each method: PRD\((X, X_{\text{Rec}}\)), Corr\((X, X_{\text{Rec}}\)), PRD\((Y, Y_{\text{Rec}}\)) and Corr\((Y, Y_{\text{Rec}}\)). We also report the execution time for each method. We begin the comparison by the introductory examples in Section 2.
4.1. Illustrative examples

Fig. 9a illustrates the work performed by our method for the case shown in Fig. 2. The obtained numerical results SEA vs. DTW for this case are reported in Table 1. The similarity (Corr) and dissimilarity (PRD) factors indicate that the two methods did perform quite the same, with a small advantage for the SEA method. For the execution time, the SEA reported value is about 1/2 that of the DTW. For the purpose of further comparison with DTW, we also solved the case shown in Figs. 7 and 8 by SEA. Recall that DTW failed to match the two similar traces in these figures. For this case, the computed numerical results for both methods are reported in Table 2. For this case, all numerical values are in favor of SEA: The SEA Corr is significantly greater than that of DTW, the PRD is significantly smaller than that of the DTW and the SEA execution time is significantly lower than that of the DTW. Note here that, while in the first experiment, the length of both $X$ and $Y$ was set to 50 samples, the ratio of the DTW execution time to that of SEA was about 2:1. In Table 2, where the size of $X$ and $Y$ was set to 300 samples, this ratio was of 10:1 order. Needless to say here that on the quality of reconstruction criterion based on visual inspection, the SEA is advantageous over DTW. Clearly, for this last case, only the SEA method could do a good match. Fig. 10a illustrates how the SEA method performed the matching operation for this case. The plot shows that the SEA method traverses the time series, not in the temporal order, like in the DTW, but rather in the magnitude order. Fig. 10b reports the match (original vs. reconstructed, upper plots) and the errors of reconstruction (lower plots). These plots are to be compared with the ones shown in Fig. 8. It can be checked that the matching is almost perfect to the point it is difficult to distinguish the original time series (discontinuous line) from the reconstructed one (continuous line). The error plots also confirm the high quality of the reconstruction, even at high frequency regions where it is still acceptable.

4.2. Thorough comparison

In this sub-section, the proposed SEA method is compared to the DTW algorithm in a deeper way. On the theoretical aspect, the DTW and SEA analyses show that both SEA and DTW temporal complexities are of quadratic order. For the space complexity, there is a clear advantage for the SEA method (linear complexity) over the DTW method (quadratic complexity). On the experimental level,

---

**Table 1**

<table>
<thead>
<tr>
<th>Method</th>
<th>PRD($X, X_{rec}$) 0...100%</th>
<th>PRD($Y, Y_{rec}$) 0...100%</th>
<th>Corr($X, X_{rec}$) 0...1</th>
<th>Corr($Y, Y_{rec}$) 0...1</th>
<th>Execution time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEA</td>
<td>1.88</td>
<td>1.55</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.031</td>
</tr>
<tr>
<td>DTW</td>
<td>2.37</td>
<td>2.56</td>
<td>0.9999</td>
<td>0.9998</td>
<td>0.063</td>
</tr>
</tbody>
</table>

**Table 2**

<table>
<thead>
<tr>
<th>Method</th>
<th>PRD($X, X_{rec}$) 0...100%</th>
<th>PRD($Y, Y_{rec}$) 0...100%</th>
<th>Corr($X, X_{rec}$) 0...1</th>
<th>Corr($Y, Y_{rec}$) 0...1</th>
<th>Execution time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEA</td>
<td>13.31</td>
<td>9.84</td>
<td>0.995</td>
<td>0.995</td>
<td>0.125</td>
</tr>
<tr>
<td>DTW</td>
<td>45.65</td>
<td>27.44</td>
<td>0.924</td>
<td>0.965</td>
<td>1.188</td>
</tr>
</tbody>
</table>
we further compared the two methods on the same above used criteria: PRD, Corr factor, execution time and visual inspection. For this purpose, we chose a reference ECG time series (rec. 100 in the MIT-BIH database) and applied the following methodology. In each application, we took a segment \( X \) from the very beginning of this record and a second segment \( Y \) from the very end of this record. In the successive applications, the lengths of both \( X \) and \( Y \) were set, respectively, to 500, 1000, 1500, 2000 and 3000 samples. For each application and method, we report the obtained numerical values for the PRD, Corr and execution time. We indicate that all applications have been performed on a machine with the following specifications: Intel Celeron 2.4 GHz processor with 256 Mbytes main memory, running under Windows XP. We also mention that both DTW and SEA have been implemented in the Matlab 7 programming environment.

The obtained results for the PRD, Corr and execution time have been reported in the form of graphical plots this time. Fig. 11, upper plot, shows the PRD values for both methods as a function of the time series size. This plot clearly shows that the SEA PRD is always lower than that of DTW. The middle plot of the same figure shows that the Corr factor of SEA is close to the +1 value (perfect match) and is always better than the DTW Corr in all cases. Fig. 11, lower plot, reports the ratio DTW-execution-time over SEA-execution-time. The plot shows that this ratio is, respectively, 16:1, 30:1, 48:1, 64:1 and 97:1. Clearly, even though the two methods share the same theoretical temporal complexity, the advantage is by far in favor of the SEA method. Knowing that we use the same record data, this figure confirms the superiority of SEA over DTW in the quality of matching, based on the numerical aspect. For the quality of matching comparison, based on visual inspection, Fig. 12 reports the case of the 500 samples time series for \( X \) and \( Y \). Fig. 12a (respectively, Fig. 12b) reports in the upper plots the original time series \( X \) and \( Y \), in the middle plots the reconstructed time series \( X_{\text{Rec}} \) and \( Y_{\text{Rec}} \) and in the lower plots, the error of the matching for the DTW method (respectively, SEA method). It can be seen that the middle plot of Fig. 12a (DTW) shows important distortions (mismatch). These distortions are confirmed by the computed errors (lower plots, Fig. 12a). On the other hand, Fig. 12b (SEA), middle plot, shows a near perfect reconstruction, to the point it is difficult to distinguish between the original and the reconstructed time series. This fact is confirmed by the computed error (lower plots, Fig. 12b).

4.3. Specific applications

In this sub-section, we illustrate some specific cases that SEA can solve and deserve to be reported. The applications are concerned with the capability of SEA to deal with matching time series of different lengths in various situations. For this purpose, we report three cases: Same time series (same person ECG) taken at different onsets/offsets (Fig. 13), same annotation time series taken from different records (different persons ECGs, Fig. 14) and different annotations time series taken from the same record (same
Fig. 11. Comparison results SEA vs. DTW for increasing size segments from rec. 100 (MITBIH). Upper plot: PRD% (0…100), middle plot: Correlation: 0…1, lower plot: Ratio DTW-execution-time over SEA-execution-time. The legend is common for the two upper plots.

Fig. 12. SEA vs. DTW for a 500 samples time series (rec.100–MITBIH). (a) DTW. (b) SEA. Upper: original time series, middle: original vs. reconstructed times series, lower: error of reconstruction.
person ECG, Fig. 15). We report especially the mean PRD and Corr for the three applications shown in Table 3. This table shows that for the first application, the PRD is rather low whereas the correlation is rather high, which suggests that the two ECGs would belong to the same record. In other words, the two ECGs would belong to the same person. For the second application, it is all the way around: The PRD is rather high whereas the correlation is rather low, which suggests that the two ECGs would belong to two different records. In other words, the two ECGs would belong to two different persons. These two applications give an idea of what the proposed SEA method can do regarding the resolution of ‘The person identification using ECG problem’. The third application shows a high correlation, but the PRD gives an ambiguous value to decide. However, the plots shown in Fig. 15 explain the situation. The two ECGs belong to the same record, but the left is a normal beat ECG trace, whereas the right ECG shows one abnormal beat (second from the left), the shape of which is different from the normal (other) beats. The medical experts annotation for this beat is: Premature Ventricular Contraction (PVC). The difference between the two ECGs has been computed by the SEA method in a rather good manner, since the novelty (abnormal beat, PVC) has been isolated as the major event in $Y-Y_{Rec}$. This illustration also gives an idea of what the proposed SEA method can do in resolving ‘The novelty detection in time series problem’. Note also that, even though the used time series are of different lengths, patterns and number of periods, the SEA method correctly matched the sub-patterns in all applications, in the sense that the significant patterns ($P$ wave, QRS complex and $T$ wave) in one time series have been mapped onto their respective correspondents in the other time series. This is an important property. It enables computation of the difference between the two time series on a point-to-point basis. Detailed comparison can then be
established. For instance, the residual error shown in Fig. 13, lower left plot, can tell that the QRS patterns in the left time series ($X$) are shorter than those in the right time series ($Y$). In Fig. 14, it is the contrary; those of $X$ are taller than those of $Y$.

However, these illustrations must be considered with some caution. The SEA method is basically a quasi-periodic time series similarity measure, on the basis of which, and perhaps other criteria, decisions could be taken for specific applications. For instance, the case of ECGs that are similar to the human eye, but which belong to two different persons, is particularly challenging for the task of ‘Person identification using ECG’. Indeed, for similar time series, the SEA method will always induce a high correlation factor (Corr) and a low distortion measure (PRD). Evidently, the SEA results should be combined to other criteria for accurate decision. For instance, it is believed that the ECG trace is unique to each person. Thus, even though ECG traces look globally similar to the human eye, features within each trace show generally differences that can be exploited for the discrimination task. To illustrate these ideas, we localized two couples of similar ECG traces in the MIT-BIH database. These couples are, respectively, records (101,103) and records (115,123). The tests performed consisted in applying the SEA method to each record by taking 2000 samples from its beginning and 2000 samples from its end, then in applying the SEA method to the first 2000 samples from each couple of similar records. The results of these tests are reported in Table 4. The applications concerning the comparison of the similar traces are reported in Fig. 16, upper plots, for the couple (115–123) and Fig. 16, lower plots, for the couple (101–103). In this figure and in Table 4, $B$ stands for ‘beginning of record’ and $E$ stands for ‘end of record’. Table 4 shows that the SEA method assigned high correlation factors (Corr) and low distortion measures (PRD) for segments belonging to the same person. In these cases, one should notice that the correlation factor is higher than 0.99. For the comparison of similar segments (115–123) and (101–103), the SEA assigned this time also high correlation values, but lower than those of ‘same-record’ segments. For the PRD, both couples received a value significantly higher than that of the ‘same-record’ segments. The visual inspection of the plots shown in Fig. 16 reveals that the quality of reconstruction (matching) in both tests is lower than that obtained when comparing segments belonging to the same person. This observation is confirmed by the error plots. Results suggest then that the SEA method could discriminate between similar ECGs belonging to different persons,

![Fig. 15. Illustration of novelty detection in electrocardiogram traces (rec. 119). The left trace ($X$) is a normal beat ECG. For the right ECG ($Y$), all beats (periods) are normal, except the second. This is a PVC (premature ventricular contraction) beat. Note the characteristic QRS and ST segment (depression) of that particular beat. The error $Y_{\text{Rec}}^– Y$ clearly shows that SEA method could notice the difference between the two original time series $X$ and $Y$.](image)

Table 3

| Specific applications main results |
|-----------------|-----------------|
| Application | Mean PRD% | Mean correlation |
| 1 | 19.35 | 0.992 |
| 2 | 64.58 | 0.858 |
| 3 | 37.62 | 0.952 |

Applications 1 and 2 illustrate capability of person identification by ECG using SEA. In 1, the ECGs belong to the same person (low PRD and high correlation). In 2, ECGs belong to different persons (high PRD, low correlation). Application 3 illustrates novelty detection. This result is better shown in Fig. 15 (error). The correlation also tells that the ECGs could belong to the same person.

Table 4

| Application of the SEA method to similar ECGs belonging to different persons |
|-----------------|-----------------|
| ECG segments | Mean PRD% | Mean Correlation |
| 115B–115E | 13.19 | 0.9926 |
| 123B–123E | 16.34 | 0.9900 |
| 115B–123B | 24.83 | 0.9700 |
| 101B–101E | 19.28 | 0.9946 |
| 103B–103E | 14.18 | 0.9980 |
| 101B–103B | 33.79 | 0.9877 |

The first column specifies the couple of used segments (e.g. 115B: 2000 samples from the beginning of record 115, 115E: 2000 samples from the end of record 115).
in some cases. However, we believe that other criteria should be combined to the SEA method results for more accurate decisions.

5. Discussion and conclusion

The different applications showed that both DTW and SEA are useful in quasi-periodic time series matching. However, results showed that only the SEA method could deal with this problem without any specific requirements on the alignment of the two time series to match. The test applications also showed that the SEA method is more efficient than DTW on the basis of quality of matching, expressed by the PRD and Corr factors and visual inspection. The applications also illustrated that the SEA method outperforms the DTW method on the basis of the execution time under the same programming environment. Another interesting aspect of the proposed method is that no parameters are required to match (compare) two time series using SEA. The proposed method’s superior performance is due to a novel strategy in traversing the time series. Whereas the classical DTW uses a temporal order strategy in processing the time series samples, SEA is based on exchange of block-sorting signatures. In this respect, it

Fig. 16. Application of the SEA method to similar ECGs belonging to different persons. Upper plots: 115B–123B, lower plots: 101B–103B. For each case, we plotted the original traces (upper), the original vs. the reconstructed trace (middle) and the difference between the original and the reconstructed trace (lower).
matches the samples on the magnitude order. On the basis of these observations, it can be argued that the SEA method could be a suitable alternative to DTW for quasi-periodic time series matching.

We also mention that the current study did not target specific applications using the SEA method. However, we foresee at least two important applications: Person identification and Novelty detection. The person identification using ECG application is a promising new trend in the field of pattern recognition and biometrics, e.g. (Biel et al., 2001; Israel et al., 2005). It comes, if not as an alternative, then as a complementary tool to the existing biometrics (finger prints, face recognition, Iris recognition). The Novelty detection application is a well-known problem within the time series data-mining community, e.g. (Dasgupta and Forrest, 1996; Ma and Perkins, 2003). Its main applications are automatic diagnosis, monitoring, fault detection, etc., with respect to processes and systems generating the time series. For both applications, we illustrated one example using SEA. These applications deserve to be developed in future works.

Acknowledgements

The author would like to thank the anonymous referees for their valuable critics and suggestions. Their comments greatly contributed to the quality enhancement of this work. This work is partially supported by the Algerian Ministry of Higher Education and Scientific Research under grants CNEPRU No. J2101/02/03/05 and CNEPRU No. 0201620070005.

References


