# An Efficient Strategy for Evaluating Similarity between Time Series based on Wavelet / Karhunen-Loève Transforms

T. Rocha, S. Paredes, P. Carvalho, J. Henriques

*Abstract* - **The present work aims to present an innovative measure able to efficiently evaluate the similarity between two physiological time series.**

**The proposed methodology combines the Haar wavelet decomposition, in which signals are represented as linear combinations of a set of orthogonal basis, with the Karhunen-Loève transform, that allows for the optimal reduction of that set of basis. The similarity measure is based on the Euclidean distance, but indirectly calculated through the linear combination coefficients of both time series. Moreover, an iterative scheme for computing the referred coefficients significantly decreases the computational complexity of the method that, due to its simplicity and fast execution, can be easily applicable in clinical applications, namely in computational demanding contexts such as telemonitoring environments.**

**This strategy has been successfully implemented and validated inside HeartCycle project, applied to blood pressure signals collected by a telemonitoring platform (TEN-HMS) in the recognition of hypertension episodes.**

## I. Introduction

The problem of similarity between biosignal time series aims at finding whether they present, or not, a similar behavior. A generalization of the similarity assessment is the similarity indexing problem, which intends to determine the subsequences of a time series that are similar to a given pattern. In this context, the main goal of the present work is the development of algorithms able to find segments of a time series (subsequences) that reflect the same dynamics as a given temporal pattern.

This work is included in the development of models for assessing the cardiovascular (CV) status of patients within the European project HeartCycle [1].These models assume that CV status i) is continually updated using measurements, parameters and symptoms, collected during daily home monitoring process, and *ii*) it may be characterized based on specific cardiovascular conditions, such as hypertension, ischemia and ventricular arrhythmias. In this context, the present work aims to develop an efficient and effective method to evaluate the similarity between a signal collected from the telemonitoring system (blood pressure, electrocardiogram, respiration rate, etc.) and a given characteristic historic pattern (e.g., sudden increase in blood pressure), in order to detect the occurrence of events that characterize the cardiovascular conditions referred above.

Significant advances have been made in the development of methods for the detection of similarities in time series and numerous approaches have been proposed. The simplest time-domain algorithms used Euclidean distance to calculate a similarity metric between time series of the same length. Others proposed dynamic time warping (DTW) for time series of different lengths [3]. Nevertheless, due to the high dimensionality of time series, most of the approaches perform dimension reduction on data. In effect, some works used discrete Fourier transform [4], singular value decomposition [5], or piecewise aggregate .approximation [6] techniques. Other authors used the principal component analysis (or Karhunen-Loève transform) [7] while others applied methods based on discrete wavelet transform (DWT) [8][9].

The present work presents an innovative measure able to efficiently evaluate the similarity between two physiological time series. The proposed methodology combines the Haar wavelet decomposition, in which signals are represented as linear combinations of a set of orthogonal basis, with the Karhunen-Loève transform, that allows for the optimal reduction of that set of basis. The similarity measure is based on the Euclidean distance, which is indirectly calculated by means of the linear combination coefficients of both time series. Furthermore, using an iterative algorithm for computing the referred coefficients, computational complexity of the method significantly decreases. Therefore due to its simplicity and fast execution characteristics, it can be easily applicable in clinical applications, such as the telemonitoring environment of the HeartCycle project. In effect, the proposed similarity assessment methodology has already been implemented successfully in several applications, namely in the context of Heartcycle project in the prediction of hypertension events, using blood pressure signals collected by a telemonitoring platform (TEN-HMS) [2].

The remainder of this paper is organized as follows. In the next section, the proposed methodology is described, while in section 3 illustrative examples of its application are presented. Finally, in section 4, some conclusions are drawn.

## II. Proposed Methodology

The proposed similarity measure and indexing scheme involve six main steps, as depicted in Figure 1.

Follows a brief description of each step identified in the scheme above.

Step 1 - Vertical shift removal: to guarantee that similarity assessments are independent of variations in the vertical position, a vertical shift removal procedure is employed.

T. Rocha and S. Paredes - Departamento de Engenharia Informática e de Sistemas, Instituto Superior de Engenharia de Coimbra, Portugal, {teresa, sparedes}@isec.pt.; P. Carvalho and J. Henriques - CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Portugal, {carvalho, jh}@dei.uc.pt.

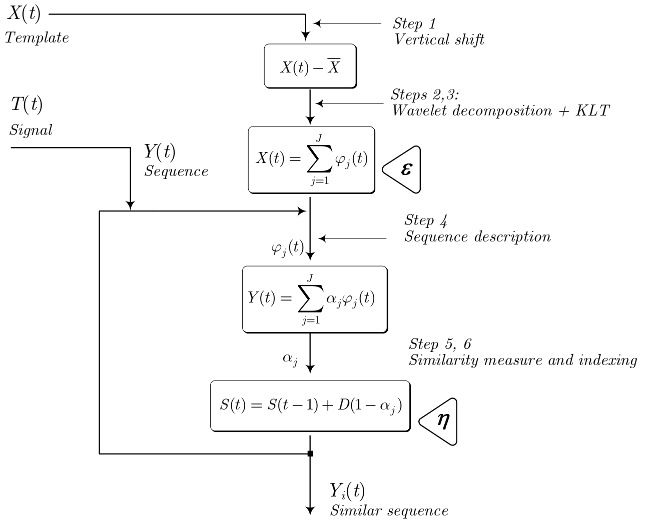


*Figure 1.* Proposed similarity measure and indexing scheme.

Step 2 - Wavelet decomposition of the template: the decomposition of the template (pattern) to be compared with the time series is achieved by means of a set of orthogonal wavelet basis.

Step 3 - Optimal dimension reduction: based on the localization property of the wavelet basis, the ones that significantly reflect the dynamical patterns of the template are chosen to compose a reduced set of basis.

Step 4 - Sequence description: a subsequence of the signal to be compared with the template is described by means of the previous reduced set of basis. It is important to refer that this description does not involve a wavelet decomposition, but a simple computation of coefficients.

Step 5 - Similarity measure: the coefficients obtained for the template and subsequence description using the reduced set of basis, are employed to derive a similarity measure. This measure allows the interpretation as a trend evolution, as well as a percentage of the amplitude difference between the time series.

Step 6 – Subsequence indexing: based on the previous similarity measure, and using the particular Haar wavelet, an efficient iterative similarity indexing algorithm is proposed.

The parameters to be selected, $\varepsilon \in \mathbb{R}^+$ and $\eta \in \mathbb{R}^+$ correspond to: $\varepsilon$: controls the approximation error by determining the number of basis to be considered in the template decomposition; $\eta$: establishes if two signals that present the same behaviour are or not similar by thresholding the difference in amplitudes of the two series under comparison.

### *A. Vertical shift removal*

The proposed scheme, assumes that similarity should be insensitive to differences in the vertical shift of the time series. Therefore, in order to eliminate this offset, the template signal $X(t) \in \mathbb{R}^{1,N}$ is in a first step modified as (1), where $\overline{X}$ is the mean value of $X(t)$.

$$X(t) = X(t) - \overline{X} \tag{1}$$

### *B. Wavelet decomposition of the template*

In a second step, a discrete wavelet transform (DWT) is applied to the template signal $X(t)$, which is decomposed in terms of an approximation of the original sequence, plus a set of details. The main trend of the input sequence is preserved in the approximation part, while the localized changes are kept in the detail parts. Assuming that the length of the signal is *N*, and considering the *L* level of decomposition, such that $L = \log_2(N)$, the original signal can be reconstructed as described by (2).

$$X(t) = c_{0,0}\phi_{0,0}(t) + \sum_j \sum_k d_{j,k}\psi_{j,k}(t) \tag{2}$$

Therefore, the signal $X(t)$ can be described as a linear combination of the basis functions, $\phi(t)$ and $\psi(t)$, respectively, approximation and detail functions. In order to simplify the computation of the similarity measure, it is assumed that the signal can be generically described as (3).

$$X(t) = \sum_{j=1}^{N} \varphi_j(t) \tag{3}$$

The basis functions $\varphi_j(t) \in \mathbb{R}^{1,N}$ are orthogonal (not necessarily orthonormal) and incorporate the respective coefficients: $\varphi_j(t) = d_j \ \psi_j(t)$ with $\varphi_1(t) = c_{0,0}\phi_{0,0}(t)$.

Using matrix notation

$$X(t) = \Gamma \ \Phi(t) \tag{4}$$

In the previous equation, $\Gamma \in \mathbb{R}^{1,N}$ is a weight vector composed of coefficients, in this case a vector of ones: $\Gamma = [1, \ 1, \ \ldots, \ 1]$. The matrix $\Phi \in \mathbb{R}^{N,N}$ is composed of the orthogonal wavelet basis.

### *C. Optimal basis reduction*

In order to reduce the number of wavelet basis, Karhunen-Loève transform (KLT) is applied. The KLT basis functions are obtained as the eigenvectors (also known as principal components) of a covariance matrix. In this particular case, the covariance matrix, $R \in \mathbb{R}^{N,N}$, is composed of the wavelet basis $\varphi_j(t)$, in the form:

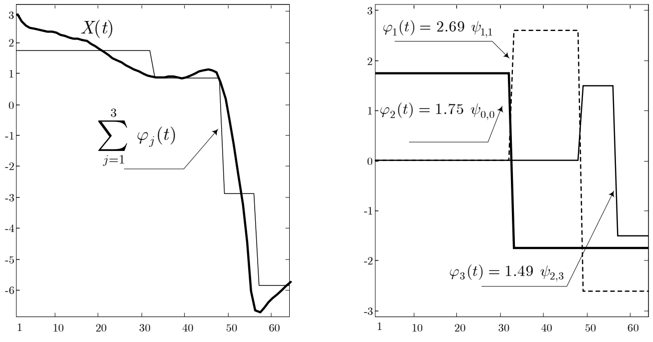
$$R = \frac{1}{N-1} \Phi \ \Phi^T \tag{5}$$

The best approximation of the signal $X(t)$ using a reduced set of basis can be achieve by means of a linear combination of the basis corresponding to the first *J* eigenvalues of the *R* matrix.

In conclusion, the original signal, $X(t)$, can be described in terms of a reduced set of *J* basis, as $\widehat{X}(t)$, equation (6).

$$\widehat{X}(t) = \sum_{j=1}^{J} \varphi_j(t) \qquad (6)$$

Figure 2 illustrates this approximation process for a signal $X(t)$ with length $N = 64$. Using Haar wavelet decomposition and establishing a predefined level of precision (threshold $\varepsilon = 0.92$) an optimal basis reduction is achieved. As can be seen in Figure 2, using only three basis it is possible to capture the main characteristics of the signal.



*a) Signal approximation* *b) Detail basis*

*Figure 2.* Signal approximation using Haar wavelet decomposition.

### *D. Similarity measure*

The proposed similarity measure is based on the Euclidean distance; however, it is indirectly computed from the coefficients of the reduced set of wavelet basis, which reflect the main dynamic patterns of the time series.

Given a signal $Y(t) \in \mathbb{R}^{1,N}$ to be compared with a template $X(t) \in \mathbb{R}^{1,N}$, the first step consists in describing it as a linear combination of the orthogonal basis functions $\varphi_j(t)$ that were used to represent the template.

$$Y(t) = \sum_{j=1}^{N} \alpha_j \ \varphi_j(t) \qquad (7)$$

The coefficients $\alpha_j \in \mathbb{R}$ always exist and are given by:

$$\alpha_j = \frac{< Y, \varphi_j >}{< \varphi_j, \varphi_j >} \qquad (8)$$

The operator $< a,b >$ denotes the dot product between the vectors *a* and *b*, i.e., $< a,b > = \sum_i a(i)\ b(i)$.

Representing the coefficients to be determined as $\Omega = [\alpha_1 \ \ \alpha_2 \ \ \dots \ \ \alpha_N]$, the signal (vector) *Y* can be written as (9), similarly to (4).

$$Y = \Omega \ \Phi(t) \qquad (9)$$

Considering trivial operations it can be shown that the coefficients are obtained by (10), where the matrix $\Phi^{\dagger}$ denotes the pseudo-inverse of matrix $\Phi$.

$$\Omega = Y \ \Phi^T \left(\Phi\Phi^T\right)^{-1} = Y\Phi^{\dagger} \qquad (10)$$

As previously referred, the proposed similarity measure between the template $X(t)$ and the sequence $Y(t)$ is based on the distance between the two vectors of coefficients $\Gamma = [1,\ 1,\ ...,\ 1]$ and $\Omega = [\alpha_1, \alpha_2, ..., \alpha_J]$.

$$D\left(X,Y\right) \ \simeq \ D(\Gamma,\Omega) \qquad (11)$$

Although several types of distances could be used, in the present work it is computed as the Euclidean distance. Finally, the distance is converted into a similarity measure, normalized in the interval [0..1], by the following operation:

$$S(X,Y) = e^{-D(\Gamma,\Omega)} \qquad (12)$$

In conclusion, using the described procedure, two time series signals are similar if equation (14) is verified, where $\eta \in \mathbb{R}^+$ (a positive scalar).

$$S(X,Y) = e^{-D(\Gamma,\Omega)} \leq \eta \qquad (13)$$

Additionally, the proposed similarity measure can be easily interpreted. In fact, regardless of its exact value, a positive coefficient ($\alpha_j > 0$) reveals that signal and template present the same behavior, i.e., the same evolution or trend. In case of a negative value ($\alpha_j < 0$), it means that the signal and template have opposite trends. Therefore, comparing the signal and template behaviors can be simply done by taking into consideration the signs of the $(\alpha_j)$ coefficients.

### *E. Similarity indexing*

The proposed indexing procedure uses a windowing scheme to calculate the similarity between the template, $X(t) \in \mathbb{R}^{1,N}$, and the signal being analyzed, $T(t) \in \mathbb{R}^{1,T}$. In fact, the similarity measure is estimated for each segment, $Y(t) \in \mathbb{R}^{1,N}$, as illustrated in Figure 5. Thus, a set of $(T-N)$ similarity measures have to be computed.

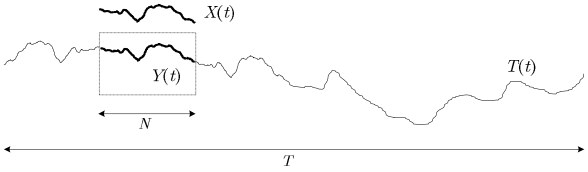


*Figure 5 - Subsequence or index similarity problem.*

For each segment, the coefficients $\Omega = [\alpha_1, \alpha_2, ..., \alpha_J]$ can be obtained using the pseudo-inverse formulation (equation (10)). Nevertheless, taking into account that the basis ($\varphi_j(t)$) are fixed and present a compact support, the similarity indexing can be computed using an iterative scheme, equation (14), which significantly decreases the computational complexity of the method.

$$\alpha_j\text{(t+1)}=\alpha_j\text{(t)}+ \\ +\kappa_j\left(-y(t+1) - y(t+N+1) + 2\ y(t+\frac{N}{2}+1)\right) \qquad (14)$$

Moreover, this procedure is independent of the type and support duration of the wavelet, since it only depends on the first, last and middle values of the segment under analysis.

## III. Illustrative Examples

The following examples illustrate the proposed similarity measure, considering the template, $X(t)$ , of the Figure 2-a).

In the first case, Figure 3, a signal $Y_1(t)$ described as follows, is considered.

$$Y_1(t) = \sum_{j=1}^{3} \alpha_j \ \varphi_j(t) = 0.29\varphi_1(t) + 0.14\varphi_2(t) + 0.49\varphi_3(t) \quad (15)$$

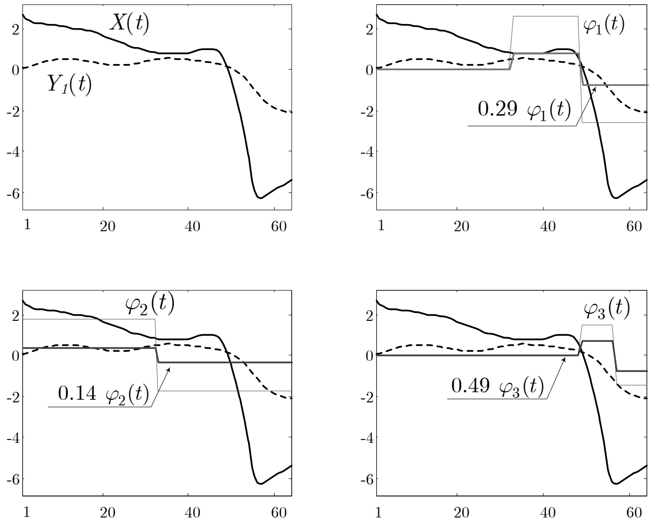


*Figure 3 – Similarity measure between time series: same behavior.*

In this case, all the coefficients $\alpha_j$ $(j = 1,2,3)$ are positive, thus having the same sign as the coefficients of the template (all equal to 1). From this simple statement, it can be concluded that template and signal present the same behavior, i.e., the same temporal trend.

In the second example, Figure 4, the same template, $X(t)$ , is compared with a second signal, $Y_2(t)$ , described as:

$$Y_2(t) = \sum_{j=1}^{3} \alpha_j \varphi_j(t) = -0.93\varphi_1(t) - 0.49\varphi_2(t) + 0.21\,\varphi_3(t) \quad (16)$$

In this case it is observed that two of the coefficients are negative and a third is positive. The existence of negative values means that signal and template coefficients have not the same sign. Thus, it may be concluded that, in global terms, they present an opposite behavior.

Summarizing, the key principle is that two signals are similar if they present the same behavior (trend), or correspondingly, if their coefficients have the same sign.

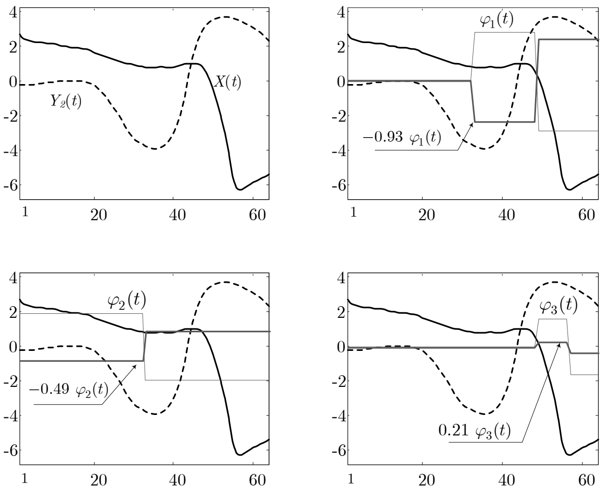


*Figure 4 – Similarity measure between time series: opposite behavior.*

## IV. Conclusions

This work proposed an innovative measure able to efficiently evaluate the similarity between two physiological time series, which combines the Haar wavelet decomposition with the Karhunen-Loève transform. Given its simplicity and fast execution characteristics, it can be easily employed in clinical applications, namely in computational demanding contexts such as the telemonitoring environment.

In particular, it has been successfully implemented and validated inside HeartCycle project applied to blood pressure signals in the recognition of hypertension episodes. In the future, the proposed similarity strategy will be applied to other biosignals and respective clinical conditions.

## References

[1] Reiter H., Maglaveras N.; HeartCycle: Compliance and effectiveness in HF and CAD closed-loop management; EMBS 2009, Minneapolis, MN, pp 299 - 302, 2009.

[2] Rocha T., Paredes S., Carvalho P., Henriques J.; A Wavelet-based Approach for Time Series Pattern Detection and Events Prediction Applied to Telemonitoring Data; Proc. of the EMBC 2011, Boston, MA, pp 6037 - 6040, 2011.

[3] Park S., Chu, W., Yoon J, Hsu C.; Efficient searches for similar subsequences of different lengths in sequence databases; Proc. of the International Conference of Data Engineering, pp 23-32, 2000.

[4] Agrawal R., Faloutsos C., Swami A.; Efficient similarity search in sequence databases; Proc. of the Intl. Conf. on Foundations of Data Organizations and Algorithms (FODO'93), pp 69–84, 1993.

[5] Wu D., Agrawal D., Abbadi A., Singh A., Smith T.; Efficient retrieval for browsing large image databases; Proc. 5th of Conf. on Information and Knowledge Management, pp 11–18, 1996.

[6] Yi B., Faloutsos C.; Fast time sequence indexing for arbitrary Lp norms; Proc. of the 26th International Conference on Very Large Data Bases, 385-394, 2000.

[7] Yang K., Shahabi C.; A PCA-based Similarity Measure for Multivariate Time Series; Proc. of the 2nd ACM international workshop on Multimedia databases, 2004.

[8] Popivanov I., Miller R.; Similarity Search Over Time-Series Data Using Wavelets; Proc. of the 18th International Conference on Data Engineering, 212, 2002.

[9] Saeed M., Mark R.; A Novel Method for the Efficient Retrieval of Similar Multiparameter Physiologic Time Series Using Wavelet-Based Symbolic Representations; Proc. of the AMIA 2006 Annual Symposium, 2006.