A Wavelet Scheme for Reconstruction of Missing Sections in Time Series Signals

TR Rocha[1], SP Paredes[1], JH Henriques[2]

[1] Departamento de Engenharia Informática e de Sistemas, Instituto Superior de Engenharia de Coimbra, Coimbra, Portugal

[2] CISUC, Centro de Informática e Sistemas, Universidade de Coimbra, Coimbra, Portugal

Abstract

This work proposes a wavelet scheme to reconstruct missing data in physiologic signals that have been removed from multi-parameter recordings of patients in intensive care units.

According to the proposed strategy, the missing data section is estimated based on two other sections. If the signal to be reconstructed is an ECG, the two sections are obtained from the two other ECG derivations available in the record. Otherwise, if the incomplete signal has only one derivation, the two sections are obtained from the signal itself, by means of a pattern matching procedure. In both cases the removed section of the signal is estimated based on a strategy that combines wavelet decomposition with autoregressive models.

Applied to all records of set A, set B and Set C, this strategy provided results of (47.75% / 56.25%), (46.52% /54.30%) and (38.33% / 47.33%) for scores 1 and 2, respectively.

1. Introduction

Time series analysis, in particular search of similar temporal patterns and time series prediction, plays an important role in the clinical context. The development of automatic systems able to estimate future values of a signal based on its past values, is of central importance. In a real time scenario these systems might reveal underlying relationships between temporal patterns and the onset time of significant clinical conditions.

Several approaches for pattern detection in time series have been proposed. The simplest time-domain algorithms use Euclidean distance [1] to calculate similarity metric between time series. However, they assume that time series have the same length and the same time scale. Others [2] proposed the warping distance which was more adequate for a variety of problems, since it deals with changes and shifts in time scale. Instead of using a metric, [3] developed a similarity model based on the identification of landmarks (points of great importance) in each time series. High dimensionality of time series makes their direct indexing very time consuming, thus, most of the approaches for similarity search, perform dimension reduction on time series data. This reduction is accomplished by applying transformations to data. In effect, some works [1] used Discrete Fourier Transform (DFT). In turn, [4] employed the Singular Value Decomposition transform, while [5] used Piecewise Aggregate Approximation (PAA). Works based on Discrete Wavelet Transform (DWT) have also been proposed [6].

The basic idea behind prediction involves the development of models that estimate future values of a signal based on its past values. Linear autoregressive models, like autoregressive (AR) and moving average (MA) mappings, are well known examples of such [7]. Additionally, a large number of non-linear regressive mappings have been proposed for prediction tasks, namely fuzzy systems, neural networks and phase space reconstruction techniques [8]. Several types of transform also have been applied for time series forecasting, such as principal component analysis [9], independent component analysis [10], Fourier transform [11] and Wavelet transform (WT) [12], [13]. Among these, WT seems to be ideal for time series forecasting since time information is preserved in the transformed variables. Moreover, WT is a very effective technique for local representation of the signals in both time and frequency domains [14]. Namely, several approaches have been introduced for time-series filtering and prediction based on wavelet transform together with autoregressive models [15], Kalman filters [16] and neural networks [17].

This work proposes a strategy that combines wavelet decomposition with autoregressive models to reconstruct missing data in physiologic signals. Two main phases are considered. In the first, the section of the signal (template) localized immediately before the prediction window is described by means of an autoregressive model, based on two other sections. In the case of the ECG, these two sections are obtained from the other ECG derivations. Otherwise, if the incomplete signal has only one derivation, the two auxiliary sections are obtained from the entire signal, by means of a pattern matching procedure. In a second phase, the missing section is estimated using the computed autoregressive models, this time based on future sections data. In the case of the ECG, future data consists of the future values available in other ECG derivations. Otherwise, future data is obtained by considering the segments localized immediately after the auxiliary sections.

The paper is organized as follows. In section 2 the proposed methodology is described. In section 3 the results using PhysioNet - Computing in Cardiology challenge 2010 datasets are presented and discussed. Finally, in section 4, some conclusions are drawn.

2. Methods

In the proposed methodology the input consists of a record with 6, 7, or 8 signals acquired from bedside ICU patient monitors, that include electrocardiogram (ECG), blood pressures (ABP/ART), respiration (RESP), fingertip plethysmograms (PLET), central venous pressure (CVP) and pulmonary artery pressure (PAP).

The first step is to identify the target signal, that is, which signal of the input is the one whose final *M*-second segment is missing. Two situations are considered: the signal is identified as an ECG or as not an ECG. In both circumstances, the next step consists of modeling the section (S1), localized immediately before the missing section (P1), based on two other sections (S2 and S3). In the case the target signal is not an ECG, two auxiliary sections *S2* and *S3* are selected from within the same signal, using a pattern matching algorithm based on wavelet coefficients correlation. The two segments that best match section *S1*, are chosen to be the auxiliary sections (Figure 1a).

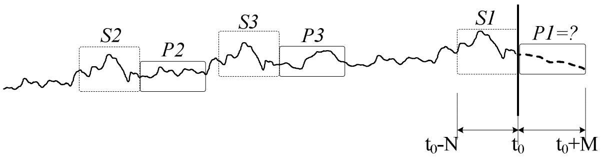

Figure 1a) Prediction of a non ECG section.

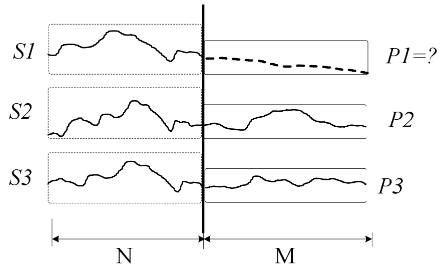

Figure 1 b) Reconstruction of an ECG section.

If the missing section belongs to an ECG, the two sections (S2 and S3) correspond to the same localization of S1 in the other ECG derivations (Figure 1b).

In this work a prediction approach based on wavelet decomposition is developed. Separate AR prediction models for the decomposed signals are built, being the various predictions merged in a final model. As claimed by several authors [18] the decomposition of time series into different components allows a better separation of the general trend terms and periodic terms, thus proving better prediction results. In effect, having defined the three sections *S1*, *S2* and *S3*, Wavelet transform of level *L*, is then applied to all of them. The next step is to model *S1* as a function of the *S2* and *S3*, using, for this finality, (*L+1*) AR models of order *P*.

The last phase is the responsible for the reconstruction of *P1*. Once again, the procedure to take depends on if the signal is, or is not, an ECG. If so, two sections *P2* and *P3* are selected from the two other ECG signals, in the localization correspondent to the prediction window (Figure 1b). Otherwise, sections *P2* and *P3* are obtained from the *M* second of signal immediately following sections *S2* and *S3* of the previous step (Figure 1a). Using AR models' parameters determined before, section *P1* is finally obtained using Wavelet reconstruction technique applied to new (*L+1*) AR models.

2.1. Pattern matching

In order to reduce time computation, Wavelet transformations are applied to the original data and a few coefficients of the transformed data are then indexed. Using Wavelet transform, a signal *X(t)* can be decomposed by cascade algorithm as shown in (1):

$$\begin{aligned} X(t) &= A_1(t) + D_1(t) \\ &= A_2(t) + D_2(t) + D_1(t) \\ &= A_L(t) + D_L(t) + D_{L-1}(t) + \ldots + D_1(t) \end{aligned} \tag{1}$$

where $D_i(k)$ and $A_i(k)$ are the detail and approximation signals at level *i*, respectively.

Considering *Y(k)* as the entire signal and *S1(k)* as the template to search within it, the proposed pattern matching method uses approximation coefficients of level *L* to find the two sections that maximize correlation between *Y* and *S1*. For the effect, the two scoring algorithms used by the challenge are employed. In fact, *S2* and *S3* sections correspond to the ones that have the two higher sums of scores *Q1* and *Q2*. The original indexes corresponding to the beginning of *S2* and *S3* in *Y(k)* are easily obtained from the identified indexes at level *L*.

2.2. Data reconstruction/prediction

Firstly, sections *S1*, *S2* and *S3* are decomposed in approximation and detail coefficients by means of Wavelet transform of level *L*, according to (1). After that, section *S1* is modelled as a function of *S2* and *S3*, by means of (*L+1*) AR models of order *P*, in the following way:

$$A_L S1(k) = a_0 A_L S2(k) + a_1 A_L S2(k-1) + \ldots + a_P A_L S2(k-P) + \\ b_0 A_L S3(k) + b_1 A_L S3(k-1) + \ldots + b_P A_L S3(k-P) \quad (2)$$

$$D_j S1(k) = c_{0j} D_j S2(k) + c_{j1} D_j S2(k-1) + \ldots + c_{Pj} D_j S2(k-P) + \\ d_{0j} D_j S3(k) + d_{1j} D_j S3(k-1) + \ldots + d_{Pj} D_j S3(k-P) \quad (3)$$

Parameters a_i, b_i, c_{ij} and d_{ij} ($i = 1,\ldots,P$; $j = 1,\ldots,L$), are determined by using a common least squares method.

Finally, the estimation of the missing data *P1,* is performed using de same parameters in new (*L+1*) AR models that depend on *P2* and *P3*. In fact, given $A_L P1$, $D_L P1$, …, $D_1 P1$, signal *P1* is straightforward obtained by the following equation:

$$P1(k) = A_L P1(k) + D_L P1(k) + \ldots + D_1 P1(k) \quad (4)$$

The Haar basis wavelet was chosen for its simplicity and mainly because it avoids aliasing problems, which are relevant in the forecasting context.

3. Results

3.1. Training and test data sets

In this year's challenge three data sets, A, B and C, of 100 records each, were made available to the participants. Each record contained 6, 7, or 8 ten-minute signals acquired from bedside ICU patient monitors. One of them had the final 30-second segment replaced by a gap, being the goal of the challenge to reconstruct this missing segment in each record. Set A was to use as a training set. Set B had target signals withheld until the conclusion of the challenge, and, although it was possible to obtain scores for Set B reconstructions, they were not included in the final rankings. Finally, set C target signals which were also withheld, were the ones that determined the final rankings and the winners of the challenge.

All the implementations done in this work regarding database access, wavelet transform and AR models implementation, were carried out using Matlab.

3.2. Parameters

Regarding the pattern searching wavelet transform based algorithm, the level of decomposition considered was *L*=4. Therefore, the approximation coefficients of level 4 were used to compute the correlation scores between the signal and the template S1, in order to identify the two auxiliary sections S2 and S3. Since the morphology of the signals in the record is different, the size (N) of templates (S1) was also variable (Table 1).

With respect to the reconstruction/prediction the level of decomposition was 5 and 3, respectively for the ECG and for the other signals. The order (P) of the regressive models, also varies from signal to signal (Table 1). The parameters corresponding to the AR models were computed using the template signals, according to (2) and (3).

Table 1. Size of templates and order of AR models.
(*fs* – sampling frequency = 125 Hz)

	ECG	ABP	ART	CVP	PAP	PLE	RES
N	120×fs	6×fs	6×fs	2×fs	5×fs	3×fs	30×fs
P	120	30	20	20	20	20	10

3.3. Discussion of results

In the case of ECG signals the reconstruction process achieved significant results. Figure 2 depicts the particular case of an ECG (set A, record #1). Actually, when considering only the scores for the ECG signals an average of 88.6% for score 1 and of 93.4% for score 2 was achieved (in set A).

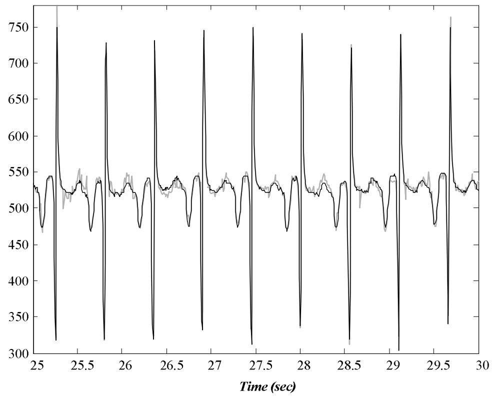

Figure 2 – Reconstruction of an ECG signal (set A, record #1).

As can be observed in Table 2, considering all the signals, the prediction strategy did not achieve acceptable results.

Table 2. Scores 1 and 2 for datasets A, B and C.

Dataset	Score1	Score2
A	47.75	56.25
B	46.52	54.30
C	38.33	47.33

Figure 3 and Figure 4 show two typical prediction examples that can be used to justify these values. In the first case (Figure 3), when the signal to be predicted follows a regular pattern the strategy is able to cope with it. Otherwise, if it presents an unexpected variation (Figure 4, approximately starting at second 13 and ending at second 15), the prediction scheme has not the capacity to deal with it. Another important difficulty of the prediction strategy is illustrated in Figure 4. In this example, although the original and predicted signals present a similar shape, actually there are lagged in time. As a consequence, both challenge scores were zero.

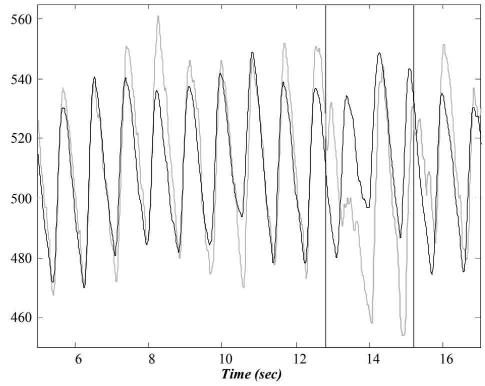

Figure 3 – Reconstruction of a PLET signal (set A, record #69).

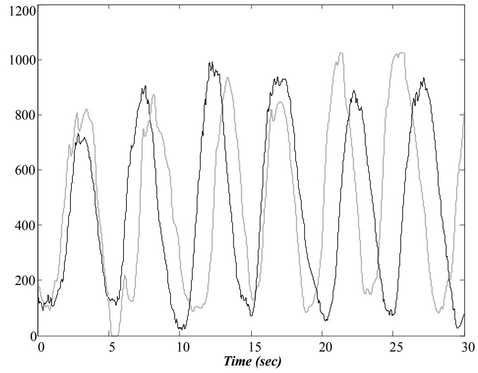

Figure 4 – Reconstruction of a RESP signal (set A, record #30).

Although these results were not very promising, the experiments performed have showed the potential of the methodology, especially in the reconstruction of the ECG. For the prediction of the other signals, in particular for non regular signals, some aspects should be improved. A future direction of work should be the development of strategies able to deal with the lag of signals, as illustrated in Figure 4.

4. Conclusions

This work proposed a methodology that combines wavelet decomposition with autoregressive models, to reconstruct/predict physiologic signals that have been eliminated from multi-parameter recordings.

In the case of the ECG reconstruction the strategy showed its potential to be applied in the clinical practice. However, for the other signals the strategy revealed some limitations that will be focused in future research.

Acknowledgements

This work was supported by HeartCycle, a project partly funded by the European Community's Seventh Framework Programme, FP7-216695, and by CISUC, Center for Informatics and Systems of University of Coimbra, Portugal.

References

[1] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases. Proc. of the 4th International Conference on Foundations of Data Organization and Algorithms, 1993; 730: 69-84.

[2] 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, 2000; 23-32.

[3] Perng C, Wang H, Zang S, Parker DS. Landmarks: a new model for similarity-based pattern querying in time series databases. Proc. of the International Conference of Data Engineering, 2000; 33-44.

[4] Wu D, Agrawal D, Abbadi A, Singh A, Smith T. Efficient retrieval for browsing large image databases. Proc. of the fifth international conference on Information and knowledge management, 1996; 11–18.

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

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

[7] Gang D, Shi Z, Yang L. Time series prediction using wavelet process neural network. Chinese Physics B, 2008; 17:6.

[8] Camilleri M. Forecasting Using Non-Linear Techniques Time Series Analysis: An Overview of Techniques and Main Issues. Computer Science Annual Research workshop, 2004; 19-28.

[9] Hiden H, Willis M, Tham M, Montague G. Non-linear principal components analysis using genetic programming. Computers and Chemical Engineering, 1999; 23: 413-425.

[10] Roberts S, Roussos E, Choudrey R. Hierarchy, priors and wavelets: structure and signal modeling using ICA. Signal Processing, 2004; 84: 283-297.

[11] Schoukens J, Pintelon R. Identification of Linear Systems: a Practical Guideline to Accurate Modeling. Pergamon Press, 1991.

[12] Yao S, Song Y, Zhang L, Cheng X. Wavelet transform and neural networks for short-term electrical load. Energy Conversion and Management, 2000; 41: 1975-1988.

[13] Chong T. Financial Time Series Forecasting Using Improved Wavelet Neural Network. Master's Thesis, 2009.

[14] Yevgeniy B, Lamonova M, Vynokurova O. An adaptive learning algorithm for a wavelet neural network. Expert Systems, 2005; 22: 235-140.

[15] Soltani S, Boichu D, Simar P, Canu S. The long term-memory prediction by multi-scale decomposition. Signal Processing, 2000; 80: 2195-2205.

[16] Cristi R, Tummula M. Multi-rate, multi-resolution, recursive Kalman Filtering. Signal Processing, 2000; 80: 1945-1958.

[17] Bashir Z, Hawary M. Short term load forecasting by using wavelet neural networks. Proc. of the Canadian Conference on Electrical and Computer Engineering, 2000; 163-166.

[18] Gan X, Liu Y, Austin F. A Prediction Method for Time Series Based on Wavelet Neural Networks. Computational Intelligence and Security, Lecture Notes in Computer Science, 2005; 3801/2005: 902-908.

Address for correspondence: Teresa Rocha, teresa@isec.pt
Instituto Superior de Engenharia de Coimbra, DEIS,
Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal