Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2012)
Hong Kong and Shenzhen, China, 2-7 Jan 2012

Using Dynamic Time Warping for Sleep and Wake Discrimination

Xi Long, *Member, IEEE*, Pedro Fonseca, Jerome Foussier, *Member, IEEE*, Reinder Haakma, and Ronald M. Aarts, *Fellow, IEEE*

***Abstract*—In previous work, a Linear Discriminant (LD) classifier was used to classify sleep and wake states during single-night polysomnography recordings (PSG) of actigraphy, respiratory effort and electrocardiogram (ECG). In order to improve the sleep-wake discrimination performance and to reduce the number of modalities needed for class discrimination, this study incorporated Dynamic Time Warping (DTW) to help discriminate between sleep and wake states based on actigraphy and respiratory effort signal. DTW quantifies signal similarities manifested in the features extracted from the respiratory effort signal. Experiments were conducted on a dataset acquired from nine healthy subjects, using an LD-based classifier. Leave-one-out cross-validation shows that adding this DTW-based feature to the original actigraphy- and respiratory-based feature set results in an epoch-by-epoch Cohen's Kappa agreement coefficient of κ = 0.69 (at an overall accuracy of 95.4%), which represents a significant improvement when compared with the performance obtained without using this feature. Furthermore it is comparable to the result obtained in the previous work which used additional ECG features (κ = 0.70).**

I. Introduction

Objective assessment of sleep quality is often based on monitoring sleep and wake phases at night. Overnight polysomnography recordings (PSG) with manually annotated hypnograms are considered a "gold standard" for objectively analyzing sleep architecture and occurrence of specific sleep-related problems [1]. They are usually performed and analyzed in sleep laboratories, and are typically split into non overlapping epochs (i.e., time intervals) of 30 seconds [1].

The use of actigraphy has been shown to be very useful in sleep-wake discrimination, but there is an evident limitation of only being able to detect wake states which correspond to obvious body movements [2], [3]. Therefore, by extracting features from respiratory and cardiac activities containing relevant information about sleep stages, it may help to better discriminate between sleep and wake states at night [4]. Different sleep stages modulate autonomous nervous system functions (e.g., respiration and heart rate) differently, so that by measuring the functions it is possible to infer the sleep stages based on the corresponding data [5]. However, including more modalities brings higher cost in acquiring physiological data because, for instance, cardiac activity is notoriously difficult to capture in good conditions, especially in an unobtrusive manner, compared with body motion and respiratory activity. Thus, we examined the possibility of achieving a comparable performance when the number of modalities used is reduced, and ECG signals were excluded.

Although the actigraphic and respiratory data theoretically contains information from which the sleep or wake state can be derived, classifiers require that this information is first extracted from the original signals as "features". A number of features have been previously explored in the context of sleep-wake discrimination [2], [4], [6]. In order to provide an opportunity of improving the discrimination performance, a Dynamic Time Warping (DTW) method [7], [8] that assesses the similarity of time series was proposed to discriminate the respiratory effort patterns between a sleep and a wake state. DTW is a technique for aligning signals that searches for similarity, allowing variations in both the horizontal (e.g., scaling or shifting of the time axis) and the vertical (e.g., amplitude or offset) aspects. It may potentially provide a good shape matching between two similar series, such as two sleep respiratory effort series in this study, because it offers flexibility to compensate for signal variations. It may not find a good shape matching between two dissimilar or less similar series such as a sleep and a wakeful series, or two wakeful ones, even when the signal variations are able to be compensated. This is because the respirations of a human in a wake state are usually not as regular as in a sleep state, and they may contain unnatural respirations or even body motion artifacts. So extracting a feature from respiratory data based on DTW is promising for sleep-wake discrimination in terms of similarity in signal level. Figure 1 indicates the respiratory effort differences between sleep and wake states.

DTW has been widely used to recognize time series patterns in various areas such as speech processing [7], bioinformatics [9], and biometrics [10]. This paper aims to investigate how far a DTW-based algorithm can help to discriminate between sleep and wake states. For that purpose, the feature sets and the Linear Discriminant (LD) classifier described in the earlier work [4] are adopted in this study.

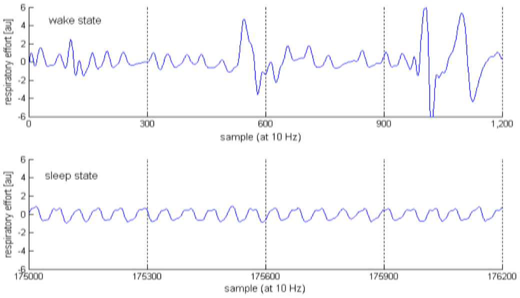

Fig. 1. Typical examples of respiratory effort series of wake (*upper*) and sleep (*lower*) with 2 minutes (i.e., 4 epochs).

Manuscript received October 14, 2011.

X. Long and R. M. Aarts are with the Eindhoven University of Technology and with the Philips Research, the Netherlands (e-mail: xi.long@philips.com). P. Fonseca and R. Haakma are with Philips Research, the Netherlands. J. Foussier is with the RWTH Aachen University, Germany.

978-1-4577-2177-9/12/$25 (C) 2012 IEEE

II. Data Acquisition

In total 9 healthy subjects (8 females) with age 31.9 ± 12.8 (mean ± std) participated in an experiment at the Sleep Health Center, Boston, USA. Actigraphy (Actiwatch, *Philips Respironics*) and full PSG (Alice 5 PSG, *Philips Respironics*) were recorded for each subject. Actiwatch is a wrist-worn device delivering activity counts derived from acceleration data, measured with a built-in accelerometer. From the PSG data only the ribcage respiratory effort signal (sampled at 10 *Hz*), as measured by inductance plethysmography was used. Sleep stages were scored by an expert according to the AASM guidelines [11]. A sleep efficiency of 91.5% ± 3.7% was calculated based on the scores of the subjects.

Before extracting features, the respiratory effort signal went through a low-pass filter for eliminating high frequency noise and then was normalized by subtracting the median peak-to-trough amplitude estimated over the entire recording to remove signal baseline. Note that the recordings from the Actiwatch were carefully synchronized with those from the PSG, using markers in both PSG and Actiwatch clocks.

III. Feature Extraction

Similar to the previous work in [4], several features were extracted from the data for each epoch of 30 seconds, but only a subset of them is considered in this study. They comprised an actigraphic feature which contains the sum activity counts over 30 seconds recorded from the Actiwatch [4], and respiratory features in both the time and frequency domains [6]. Note that all the features were normalized or smoothed, aiming to increase the robustness against inter-subject variation. However, the previous work did not consider much about series shapes and variations in the time domain. Thus, as discussed in Section I, extracting a DTW-based feature from the respiration data that quantifies the shape similarity among epochs is promising.

A. Dynamic Time Warping

DTW computes a distance between two time series with alignment in the time axis, which is called DTW-distance [8]. Consider two time series: $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ of length n and $Y = \{y_1, y_2, \ldots, y_j, \ldots, y_m\}$ of length m. The two series can be handled to form an n-by-m matrix where each element of the matrix, (i, j), corresponds to a distance function D of the squared distance between x_i and y_j: $D(i, j) = (x_i - y_j)^2$. A warping path maps the elements of X and Y through the matrix so that the total cumulative distance between them is minimized. The warping path W is denoted as $W = \{w_1, w_2, \ldots, w_k, \ldots, w_K\}$, where $w_k = (i, j)_k$ is the kth element of W. Then the DTW-distance between the two series is

$$DTW(X,Y) = \min\left\{\frac{1}{K}\sqrt{\sum_{k=1}^{K} w_k}\right\}. \quad (1)$$

The Euclidean distance between the two series is a special case of DTW-distance without alignment in the time axis when $i = j = k$ if the two series have the same length but it is known to be very sensitive to distortion in the time axis [12].

Since the nature of DTW-distance searches for an optimal warping path through all possible paths, it is combinatorially explosive. Hence, reducing the search space served by means of *conditions* helps to effectively mitigate the quadratic complexity of the DTW method [8]. Several conditions are taken into account to decrease the number of paths [8], [12]. They are: *continuity* – the steps in the matrix are confined to the points with $i_k - i_{k-1} \leq 1$ and $j_k - j_{k-1} \leq 1$; *monotonicity* – the warping path cannot go backward with respect to time; *slope constraint* – the warping paths should not have very large movements in the horizontal or vertical direction of the matrix; *boundary* – start and end points of the warping path are $(i_1, j_1) = (1, 1)$ and $(i_K, j_K) = (n, m)$, respectively; *warping band* – the path is restricted by a band of size r (i.e., $|i_k - j_k| \leq r$), which is also known as the Sakoe-Chiba Band [13]. Here we considered the *warping band* condition where r needs to be tuned due to its possible effect on discrimination accuracy.

B. DTW-Based Feature

As mentioned before, the DTW distance between two time series represents the shape dissimilarity between them. In order to help decide if an epoch of interest is a sleep or a wake state, a feature named "minimal DTW-distance" (i.e., maximal similarity) can be extracted from the respiratory effort signal of a subject based on the DTW method.

Assume that the respiratory data recorded from a subject can be split into N non-overlapping epochs $E_1, E_2, \ldots, E_N$, each of them is a time series with same length of 30 seconds. For the pth epoch E_p (p is a positive integer and $1 \leq p \leq N$) the feature value is computed as following two steps: First, compute the DTW-distance $DTW(E_p, E_q)$ between the time series of the pth epoch E_p and that of the qth epoch E_q (q is a positive integer and $1 \leq q \leq N$); Second, consider the DTW-distances between E_p and E_q for all $1 \leq q \leq N$ and $q \neq p$, the feature value is the minimal one among them. In other words, for an epoch of interest (current epoch), it searches the most similar epoch from the other remaining ones in terms of DTW-distance. Such searching aims to make sure that it can find the most similar epoch for the current one.

There are two cases: the current epoch is "sleep" (case one) and the current epoch is "wake" (case two). And the feature "minimal DTW-distance" of this epoch contains three possible situations where the minimal DTW-distance occurs: 1) between two sleep epochs, 2) between a wake and a sleep epoch, 3) between two wake epochs. Usually, the respiration curves of the wake state are less regular than that of the sleep state and they often contain many unnatural respirations. Therefore, the respiration curves between two sleep epochs should be more similar than those between a sleep and a wake epoch. Also, they should usually be more similar than the respiration curves between two wake epochs, which is caused by the inclusion of unnatural respirations in wake state. Then:

- If the current epoch is "sleep" (i.e., case one), a small value of DTW-distance is likely to be obtained by searching the similarities between this epoch and others throughout the whole signal since situation 1) may happen.
- If the current epoch is "wake" (i.e., case two), it is not

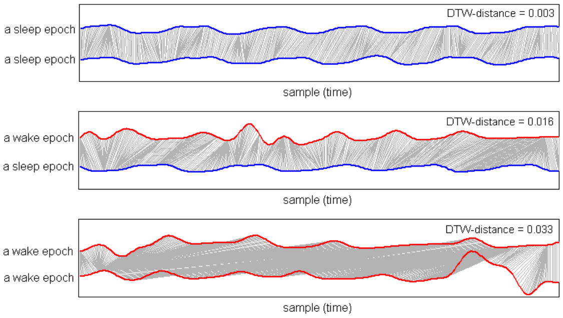

Fig. 2. Examples of two series (30 seconds each) and the alignment between them in the three situations (*upper*: two sleep epochs, *middle*: a wake and a sleep epoch, *lower*: two wake epochs) during warping process. The values of their DTW-distances are indicated.

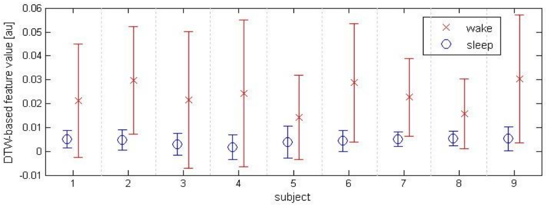

Fig. 3. Means and standard deviations of the minimal DTW-distances of sleep and wake epochs for the 9 subjects.

likely to obtain such a small feature value by doing the same process since situation 2) or 3) may happen.

Hence, it yields discrimination of sleep and wake states in feature value. This DTW-based feature also needs to be normalized as similar as the other features. Figure 2 visually describes the alignments between two epochs, as discovered by DTW, in the three situations during the DTW warping process. It shows that the feature value of DTW-distance between two sleep epochs is obviously smaller than the other two. Figure 3 clearly indicates the discrimination of sleep and wake states, whereas errors may occur due to the overlaps. As this feature considers the shape similarity between epochs allowing horizontal and vertical variations and searches the minimal DTW-distance (i.e., largest similarity) throughout the complete recording of a subject, it helps to either decrease inter-subject variation or decrease intra-subject variation.

IV. Classification

It has been demonstrated that an LD-based classifier is appropriate for the task of discriminating between sleep and wake states using actigraphy, respiratory and cardiac features [4]. In order to verify whether the new DTW-based feature described in this paper can help improve the performance in this task, the same classifier and feature set extended with the new feature was applied. An LD-based classifier distinguishes between pre-defined classes (sleep and wake), which were considered to follow the Bayesian classification rules with linear discriminants. An assumption is that the feature values are class-dependent and normally distributed per class. The classifier is obtained after a supervised learning process, applied on a training feature set that comprises annotated "examples" of features for each class, which can then be used to classify "unseen" data to one of the pre-defined classes. A more detailed description of the LD method can be found in [14].

In order to evaluate the performance of this classifier, conventional measures of specificity (proportion of correctly identified actual positives) or sensitivity (proportion of correctly identified negatives) used in binary classification are not the most adequate. The reason is that the number of epochs of one class (wake) during a recording of a whole night will naturally be much smaller than the number of epochs of the other class (sleep), in what is usually called an "unbalanced class distribution". The Cohen's Kappa coefficient $\boldsymbol{\kappa}$ [15] not only allows for a better understanding of the general performance of the classifier in correctly identifying both classes, but also allows for a better representation of the unbalanced problem when it is used as a criterion to optimize performance [4]. Moreover, several sleep statistics, typically used to assess several aspects of sleep, can be computed when discussing the performance of the classifier, such as total sleep time (TST), total wake time (TWT), sleep efficiency (SE) which is computed as the ratio of TST and total time in bed, sleep onset latency (SOL) which is the time a person takes before falling asleep, wake after sleep onset (WASO), and snooze time (ST).

V. Results and Discussion

In the experiment, a leave-one-out cross validation procedure was conducted to evaluate the performance of the sleep-wake classifier. The discrimination performance with and without using the DTW-based feature were evaluated. Cohen's Kappa was used as the evaluation criterion of discrimination performance. The use of the *warping band* condition helps reduce the quadratic complexity of computation of the DTW. The Kappa coefficient can optimally reach ~0.69, when the Sakoe-Chiba warping band r was globally optimized to be 60 samples in the warping process for all subjects. This means that the warping path is constrained by a window with a length of 120 ($2 \cdot r$) samples (i.e., 12 seconds) within an epoch of respiratory effort series containing a maximum of 300 samples. The parameter r was determined that gave the best performance based on the training set during the leave-one-out procedure.

Table I indicates the discrimination performance obtained with only the DTW-based feature, with the original feature set (from [4]), and with the feature set extended with the addition of the DTW-based feature. Here wake and sleep are defined as positive and negative class, respectively. As shown, the average $\boldsymbol{\kappa}$ is 0.51 ± 0.10 by using only the DTW-based feature. The combination of body motion and respiratory activity is useful for better discriminating between sleep and wake states. Moreover, an average $\boldsymbol{\kappa}$ of 0.69 ± 0.15 was obtained after incorporating the new DTW-based feature.

TABLE I
Discrimination Performance Comparison

Feature set	Acc.	Sens.	Spec.	$\boldsymbol{\kappa}$
Only DTW-based feature*	91.5%	69.7%	94.1%	**0.51**
Original set (without DTW)	95.0%	65.7%	97.5%	**0.65**
Extended set (with DTW)*	95.4%	68.1%	97.7%	**0.69**

*The warping path is restricted by Sakoe-Chiba band with size r = 60.

TABLE II
DISCRIMINATION RESULTS OF ALL THE SUBJECTS BASED ON THE EXTENDED FEATURE SET (WITH DTW FEATURE)

	Acc.	Sens.	Spec.	***κ****
Subject 1	96.5%	79.8%	98.5%	**0.81** (0.81)
Subject 2	94.7%	83.6%	95.8%	**0.71** (0.64)
Subject 3	90.6%	86.4%	91.3%	**0.67** (0.64)
Subject 4	93.3%	66.3%	96.9%	**0.66** (0.68)
Subject 5	95.1%	64.9%	98.6%	**0.70** (0.59)
Subject 6	98.8%	93.2%	99.1%	**0.88** (0.76)
Subject 7	99.3%	68.4%	100.0%	**0.81** (0.77)
Subject 8	94.9%	44.4%	99.9%	**0.59** (0.57)
Subject 9	95.7%	25.6%	99.7%	**0.38** (0.39)
Average	95.4%	68.1%	97.7%	**0.69** (0.65)

*The ***κ*** values without DTW-based feature are given in the brackets.

TABLE III
COMPARISON OF SLEEP STATISTICS – ABSOLUTE ERROR (MEAN ± STD)

	Absolute error (without DTW feature)	Absolute error (with DTW feature)
SE (%)	3.0 ± 2.2	2.5 ± 1.7
TST (min)	12.7 ± 9.2	10.4 ± 7.2
TWT (min)	12.7 ± 10.0	10.6 ± 7.7
SOL (min)	4.7 ± 6.9	2.4 ± 1.6
WASO (min)	8.2 ± 8.4	8.1 ± 6.4
ST (min)	0.8 ± 1.1	0.9 ± 1.2

Accordingly, an overall accuracy of 95.4% ± 2.6% was obtained. Compared with the result obtained using the original feature set, there was a significant improvement of ~0.04, in which the significance of difference was examined via a paired *t*-test where $p = 0.045$ ($p < 0.05$) and $df = 8$. An assumption of the paired *t*-test was that the two variables (i.e., Kappa coefficients with and without using the new feature) are normally distributed, this was suggested by using a *Q-Q* plot method. The results here are comparable to those obtained in the previous study when actigraphic, respiratory, and ECG data were used [4]. In that case, a ***κ*** of 0.70 and an accuracy of 96.1% were obtained. The details of the discrimination result of the 9 subjects are summarized in Table II. Compared to subject 1-7, the classifier performs less well for subject 8 and especially subject 9 where many awakenings were not correctly detected. It might result from their single-night data having a lack of wake epochs before falling asleep and/or after waking up for classifier training.

In addition, sleep statistics were estimated. For each statistic, an absolute error can be computed as the absolute difference value between the reference statistic (computed based on the manually annotated PSG data) and the estimated statistic (computed based on the results of sleep-wake discrimination). Similarly to the criterion used in [4], SOL was defined as the first epoch of a block of 17 consecutive epochs of which at least 16 were annotated as sleep. Meanwhile, ST followed a similar criterion but for wake epochs. WASO is equal to TWT excluding SOL and ST. The results are summarized and compared in Table III. These results show that as a whole, incorporating the DTW-based feature leads to obvious decrease of the absolute error in estimating SE, TST, TWT, and SOL except WASO and ST with actigraphy and respiratory signal. One reason of such exceptions may be that the respiration patterns of wake epochs during WASO and snooze periods are more similar among each other than those during sleep onset period so that discrimination ambiguity occurred.

VI. CONCLUSION

This paper studied the impact of using DTW in an automatic sleep-wake classifier. Combining a new DTW-based feature extracted from a respiratory signal with a set of existing features based on actigraphy and respiratory effort is promising in improving discrimination performance, achieving a Cohen’s Kappa coefficient of 0.69 (overall accuracy of 95.4%). The performance obtained after integrating this new feature is comparable to that obtained with a larger set of features extracted from actigraphy, respiration effort and ECG. This means that a good performance can be obtained with fewer requirements for measuring physiological signals during the night, namely those arising from the need to record ECG.

Due to the small size of dataset in this study, it is suggested to further investigate the DTW method on a larger sized dataset, such as the phase shifting effect of the series on computing the DTW-based feature, the correlation between this feature and the other respiratory features, and so on.

ACKNOWLEDGMENT

The authors would like to thank Dr. Igor Berezhnyy, Dr. Sandrine Devot, and Sam Jelfs from Philips Research Lab for the insightful comments and inspiring discussions.

REFERENCES

[1] A. Rechtschaffen and A. Kales, Eds., *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*, National Institutes of Health, Washington, DC, 1968.

[2] J. Paquet, A. Kawinska, J. Carrier, “Wake detection capacity of actigraphy during sleep,” *Sleep*, vol. 30, no. 10, pp. 1362-1369, 2007.

[3] R. J. Cole, *et al*., “Technical note automatic sleep/wake identification from wrist activity,” *Sleep*, vol. 15(5), pp. 461-469, 1992.

[4] D. Sandrine, D. Reimund, and N. Naujokat, “Sleep/wake detection based on cardiorespiratory signals and actigraphy,” *32nd Ann. Int. Conf. IEEE EMBS*, pp. 5089-5092, Nov. 2010.

[5] T. Penzel, *et al*., “Cardiovascular and respiratory dynamics during normal and pathological sleep,” *Chaos*, vol. 17(1), 015116, 2007.

[6] S. J. Redmond, *et al*., “Sleep staging using cardiorespiratory signals,” *Somnologie*, vol. 11, pp. 245-256, 2007.

[7] L. Rabiner, A. Rosenberg, and S. Levinson, “Considerations in dynamic time warping algorithms for discrete word recognition,” *IEEE Trans. Acoustics, Speech, & Signal Proc.*, vol. 26, pp. 521-527, 1978.

[8] D. J. Berndt and J. Clifford, “Using dynamic time warping to find patterns in time series,” *AAAI-94 workshop KDD*, pp. 229-248, 1994.

[9] J. Aach and G. Church, “Aligning gene expression time series with time warping algorithms,” *Bioinformatics*, vol. 17, pp. 495-508, 2001.

[10] Z. M. Kovacs-Vajna, “A fingerprint verification system based on triangular matching and dynamic time warping,” *IEEE Trans. Pattern Anal. & Machine Intelligence*, vol. 22, no. 11, pp. 1266-1276, 2000.

[11] The AASM manual for the scoring of sleep and associated events: rules, terminology & technical specifications, 2007, www.aasmnet.org.

[12] C.A.Ratanamahatana and E. Keogh, “Making time-series classification more accurate using learned constraints,” *4th SIAM (SDM04)*, 2004.

[13] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Trans. Acoustics, Speech, & Signal Proc.*, vol. AASP-26, no. 1, pp. 43-49, Feb. 1978.

[14] G. J. Mclachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Interscience, Aug. 2004.

[15] R. Bakeman and J. M. Gottman, *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press, 1986.