35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Respiration Amplitude Analysis for REM and NREM Sleep Classification

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

***Abstract*— In previous work, single-night polysomnography recordings (PSG) of respiratory effort and electrocardiogram (ECG) signals combined with actigraphy were used to classify sleep and wake states. In this study, we aim at classifying rapid-eye-movement (REM) and non-REM (NREM) sleep states. Besides the existing features used for sleep and wake classification, we propose a set of new features based on respiration amplitude. This choice is motivated by the observation that the breathing pattern has a more regular amplitude during NREM sleep than during REM sleep. Experiments were conducted with a data set of 14 healthy subjects using a linear discriminant (LD) classifier. Leave-one-subject-out cross-validations show that adding the new features into the existing feature set results in an increase in Cohen's Kappa coefficient to a value of κ = 0.59 (overall accuracy of 87.6%) compared to that obtained without using these features (κ of 0.54 and overall accuracy of 86.4%). In addition, we compared the results to those reported in some other studies with different features and signal modalities.**

I. Introduction

Over-night polysomnography (PSG) is currently considered as "gold standard" for objectively assessing sleep architecture and occurrence of sleep-related disorders [1], [2]. PSG recordings are typically collected in sleep laboratories and are usually split into non-overlapping time intervals (or epochs) of 30 seconds [1]. According to the American Association of Sleep Medicine (AASM), sleep can be divided into different states: wake, rapid-eye-movement (REM), and non-REM (NREM) which is further subdivided in sleep stages N1, N2 and N3 [3].

Several automatic wake-REM-NREM classifiers have been proposed which use multiple physiological signals including actigraphy, electrocardiogram (ECG), and respiratory effort [2], [4]. These signals contain information from which different sleep states can be derived [2], [5]. Using information extracted from these signals in so-called "features", most proposed systems employ a single classifier and use a set of fixed features to classify different sleep stages. However, this may not be the most appropriate approach since the best set of features which characterizes a given stage may not be necessarily the same which characterizes another. This is due to differences in the expression of autonomic nervous activity associated with different sleep states [5]. Thus we adopt a "hierarchical" scheme containing two levels. On a first level we classify epochs as sleep and wake, and on a second level, as REM and NREM. Our previous studies have addressed the problem of sleep and wake classification using multiple signal modalities including actigraphy, respiratory effort, and ECG signals [6], [7], [8]. The main reason of using these three types of modalities is that, to some extent, they are possible to be unobtrusively acquired [2], [4], [6]. As a follow-up, this study explores REM and NREM classification based on the idealistic assumption that all the sleep and wake stages can be correctly classified. Although experiments have shown that in practice this assumption does not hold, it is important to know to what extent a classifier and the existing feature set can discriminate between REM and NREM sleep states. Therefore, in this paper, we focus exclusively on the problem of REM and NREM classification instead of simultaneously considering all sleep stages.

Regarding REM and NREM classification, the same features used for sleep and wake classification based on actigraphy, respiratory effort and ECG data were first considered (see Section III-A). In addition, it has been shown that the amplitude of the breathing effort signal is more regular during NREM sleep than during REM sleep [9]; and also the tidal volume decreases and the respiratory variability increases when the state changes from NREM to REM [10]. Thus, we propose a new set of features that represent information about the respiration amplitude.

The problem of REM and NREM discrimination is not new. Some studies have reported relative success in the discrimination between REM and NREM sleep states in a hierarchical approach based on heart rate variability (HRV) derived from ECG signals [11] and on the combination of peripheral arterial tone, pulse rate, pulse oximetry, and actigraphy [12]. In order to evaluate the impact of the new feature set and the overall REM and NREM classification performance, we will compare the results of our classifier with those reported in literature.

II. Data Set

Fourteen single-night PSG recordings of healthy subjects (4 males and 10 females, with age 30.6 $\pm$ 10.7 y and BMI 24.4 $\pm$ 3.4 kg/m^2) were included in our data set. A subject is considered "healthy" if he or she has a Pittsburg Sleep Quality Index (PSQI) of less than 6. Among the subjects, nine were monitored in the Sleep Health Center, Boston, USA during 2009 and five in the Philips Experience Lab,

X. Long, P. Fonseca, and R. M. Aarts are with Department of Electrical Engineering, Eindhoven University of Technology, Den Dolech 2, 5612 AZ, Eindhoven, The Netherlands and with the Philips Research, High Tech Campus 34, 5656 AE, Eindhoven, The Netherlands x.long@tue.nl.

R. Haakma is with the Philips Research, High Tech Campus 34, 5656 AE, Eindhoven, The Netherlands reinder.haakma@philips.com.

J. Foussier is with the Philips Chair for Medical Information Technology, RWTH Aachen University, Pauwelsstrasse 20, 52074 Aachen, Germany foussier@hia.rwth-aachen.de.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

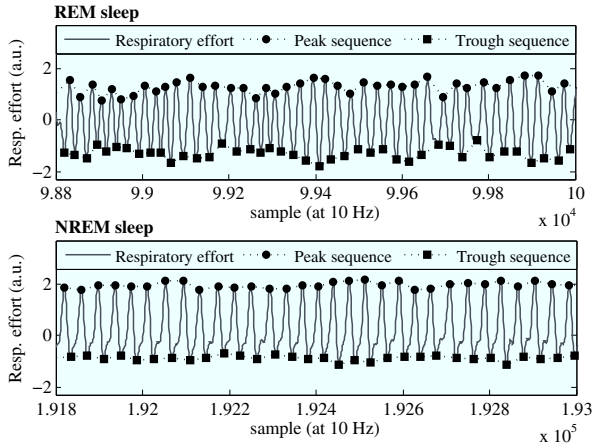

Fig. 1. A typical example of a 2-min respiratory effort signal (i.e., 4 epochs) in REM (*top*) and NREM (*bottom*) sleep states. The peaks and troughs are represented by filled circles and squares, respectively.

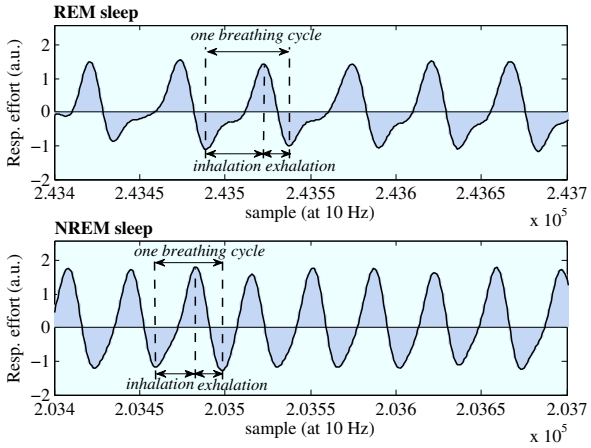

Fig. 2. A typical example of a 30-s respiratory effort signal (i.e., one epoch) in REM (*top*) and NREM (*bottom*) sleep states. The areas between the curves and the baseline are filled in gray. The inhalation and exhalation periods of one breathing cycle are indicated.

Eindhoven, the Netherlands during 2010. Actigraphy and full PSG were recorded for each subject and sleep stages were scored by an expert according to the AASM guidelines [3]. From the PSG recordings, the thoracic respiratory effort signal and the ECG data were used. Compared to our previous work where we analyzed 15 subjects [8], one was excluded due to technical problems with the recordings.

III. Feature Set Description

A. Existing Features

The existing pool of 65 features that has been used in previous studies for sleep and wake classification was first considered [6], [7]. It consists of activity counts derived from actigraphy [6], respiratory and ECG (HRV) features in time and frequency domain [2], and non-linear features based on detrended fluctuation analysis [13], sample entropy [14], and dynamic warping [7].

B. Respiration Amplitude Analysis

As mentioned in Section I, the amplitude of the respiratory effort signal is different during REM and NREM. In order to represent these differences, a number of features were implemented with the ultimate goal of improving REM and NREM classification performance.

Fig. 1 illustrates two short segments of a normalized respiratory effort signal during REM and during NREM sleep (the normalization will be explained later). It can be observed that the envelopes formed by the peak and trough sequences of the REM signal, when compared with the NREM signal: i) are more 'irregular'; ii) have generally a lower absolute mean (or median); and iii) have larger variance.

In addition, we also considered the respiratory effort 'area' comprised between the respiratory effort amplitude and its mean value (zero in the example below). It is assumed that this area, to a certain degree, correlates with breathing volume. Respiratory effort has often been used as a surrogate of tidal volume since it is obtained by measuring volume changes of chest and/or abdominal (e.g., respiratory inductance plethysmography, or RIP) [15]. It has been reported that the tidal volume, minute ventilation, and inspiratory flow rate in REM sleep are significantly lower than those in NREM sleep [10]. These are illustrated in Figure 2, suggesting a difference between REM and NREM in terms of respiratory effort area.

TABLE I
Description of RAB Features

Feature	Description
$resp\_p\_sdmedian$ $resp\_t\_sdmedian$	Standardized median (median divided by standard deviation) of peaks/troughs
$resp\_p\_ApEn$ $resp\_t\_ApEn$	Peak/trough sequence 'regularity', measured by approximate entropy [14]
$resp\_p\_to\_t\_diff$	Median of peak-to-trough differences
$resp\_cyc\_area$ $resp\_inh\_area$ $resp\_exh\_area$	Median of respiratory effort area during breathing cycles, inhalation or exhalation segments
$resp\_cyc\_aot$ $resp\_inh\_aot$ $resp\_exh\_aot$	Median of respiratory effort area over time (*aot*) measured during breathing cycles, inhalation or exhalation segments
$resp\_time\_ratio$ $resp\_aot\_ratio$	Ratio between the medians of respiratory effort time/*aot* during inhalation and exhalation segments

C. Respiration-Amplitude-Based Features

Based on these observations we implemented 13 new features. These respiration-amplitude-based (RAB) features are listed and described in Table I. All these features use a window of 3 epochs centered on each epoch of interest.

To extract the RAB features, it is important to precisely detect and locate the peaks and troughs and the transition points between inhalation and exhalation from the raw respiratory effort signal. This was achieved with some steps comprising: 1) high-frequency noise filtering; 2) baseline removal; 3) turning point detection based on sign change of signal slope; and 4) correction of falsely detected peaks/troughs. The signal resulting from step 2) was further normalized by dividing the median peak-to-trough amplitude estimated over the entire recording before further feature extraction.

D. Signal Calibration

The respiration amplitude might be affected by body movements during sleep because the sensor position may shift and/or stretch. This would lead to an uneven comparison of the signal amplitude before and after body movement, possibly yielding errors while computing the features. To avoid this issue, we normalized each signal segment between two epochs with body movements to zero mean and unit variance. These epochs were identified by comparing their activity counts with a threshold. A threshold of 2 was experimentally found to be an adequate value for this purpose.

IV. REM and NREM Classification

A. Discriminative Power

In order to evaluate the discriminative power (i.e., class separability) of the features in classifying REM and NREM, a Mahalanobis distance (MD) metric [16] was employed. Given a single feature f, the inter-class MD between the two classes (labeled as *REM* and *NREM*) is expressed as

$$D_{M\_f} = \frac{|\mu_{R\_f} - \mu_{N\_f}|}{\sigma_f}, \tag{1}$$

where μ_{R_f} and μ_{N_f} are the population means of *REM* and *NREM*, respectively; and σ_f is the standard deviation of the feature f. This equation is also called the absolute standardized distance of means, a simplified version of a MD with a single dimension. A larger D_{M_f} reflects a higher discriminative power in separating the two classes.

B. Classifier

A linear discriminant (LD) classifier has been used for sleep and wake classification in previous studies [6], [7], [8]. Likewise, we adopted a similar classifier for REM and NREM classification. To assess the performance of classification, conventional measures of sensitivity (proportion of correctly identified actual REM epochs) and specificity (proportion of correctly identified actual NREM epochs) used in a binary classification were first considered. However, since the relative epoch count for the *REM* class (17.9%) during a whole-night recording is usually much smaller than for the *NREM* class (82.1%), in what is called imbalanced class distribution, these measures may not be the most appropriate criteria. The Cohen's Kappa coefficient of agreement κ is considered a better criterion for this problem [17]. It does not only allow 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 imbalanced problem when used as a criterion to optimize performance [2], [17]. The classifier was evaluated using a leave-one-subject-out cross-validation (LOSOCV) procedure. In the following, *REM* and *NREM* were considered the positive and the negative class, respectively.

C. Feature Selection

Before classification, we applied a Sequential Forward Selection (SFS) algorithm to select features that optimizing the final classification performance, as measured by κ. This step was performed on each training set of the cross-validation. The selected features were then used for REM and NREM classification on the testing set of every LOSOCV iteration. These tests were conducted with feature sets comprising the existing pool of 65 features, the 13 new RAB features, and the combination of the existing features and the RAB features (in a total of 78 features), which are denoted F-EXST, F-RAB, and F-COMB, respectively. Note that because feature selection is performed on each iteration of the LOSOCV, the selected features for each iteration can be different.

V. Results and Discussion

Table II indicates the mean MD values for the RAB features and the maximal Spearman correlation C_{max} between this RAB feature and all the other existing features. It also indicates the number of times δ each RAB feature has been selected by the SFS algorithm from all the 78 features on the 14 iterations of the LOSOCV procedure, on a minimum of 0 (never selected) and a maximum of 14 (always selected). For comparison, we also indicate the feature $resp\_std\_5\_epochs$ (i.e., standard deviation of respiratory frequency over 5 epochs) in the table since this feature is with the highest MD value among the original set of features F-EXST. It can be observed that the first four features were frequently selected during cross-validation. Although the features $resp\_p\_ApEn$, $resp\_t\_ApEn$, $resp\_p\_to\_t\_diff$, $resp\_time\_ratio$, and $resp\_aot\_ratio$ have smaller MDs compared with some of the other RAB features, they were selected more often. This is because, as shown in the table, these features are less correlated to the other features and therefore add discriminatory information. Besides this, some RAB features were selected few times, which might be because: 1) they were not very discriminative for corresponding training sets; or 2) they correlated a lot with other features that had already been selected before during the SFS feature selection process (see Table II).

TABLE II
SUMMARY OF SOME STATISTICS OF THE RAB FEATURES

Feature	D_M	C_{max}*	δ**
$resp\_p\_sdmedian$	0.79	0.60	13
$resp\_t\_sdmedian$	0.76	0.57	14
$resp\_p\_ApEn$	0.34	0.35	12
$resp\_t\_ApEn$	0.24	0.29	10
$resp\_p\_to\_t\_diff$	0.68	0.19	8
$resp\_cyc\_area$	0.81	0.59	3
$resp\_inh\_area$	0.75	0.61	4
$resp\_exh\_area$	0.84	0.54	4
$resp\_cyc\_aot$	0.78	0.59	5
$resp\_inh\_aot$	0.81	0.55	2
$resp\_exh\_aot$	0.73	0.59	5
$resp\_time\_ratio$	0.44	0.12	4
$resp\_aot\_ratio$	0.44	0.12	5
$resp\_std\_5\_epochs$	1.14	—	14

*The maximal Spearman correlation between the RAB feature and each of the other existing features.
**Indicates the number of iterations for which a feature was selected based on SFS algorithm for classification across the 14 iterations of LOSOCV. All 78 features were considered in each iteration.

TABLE III

COMPARISON OF REM AND NREM CLASSIFICATION PERFORMANCE

Feature Set	# Epochs	# Subjects	Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	κ
F-EXST*	10,429	14	LD	86.4	63.5	91.4	0.54
F-RAB*	10,429	14	LD	81.4	40.0	90.6	0.32
F-COMB*	10,429	14	LD	87.6	68.2	91.7	0.59
ECG features† [11]	∼20,000	24	HMM§	79.3	70.2	85.1	0.55
Watch-PAT features†,‡ [12]	142,919	227	ARDA [18]	88.5	68.1	91.8	0.59

Note: The table indicates the pooled results over subjects.

*On average, for the feature sets F-RAB, F-EXST and F-COMB, 6.2, 34.4 and 40.8 features were selected (based on SFS) over all iterations of LOSOCV.

†The results were re-computed based on the corresponding reported confusion matrix.

‡The features were extracted from actigraphy, pulse rate, oxyhemoglobin saturation, and finger arterial pulse wave volume.

§Hidden Markov Model.

Table III compares the classification performance obtained with the existing features (F-EXST), with the RAB features (F-RAB), and with the combination of them (F-COMB). It shows that adding the RAB to the existing features led to a clear increase in Cohen's Kappa coefficient κ (from 0.54 to 0.59). Additionally, Table III also compares the results of our classifier with other results reported in literature. One of the first observations is that our classifier slightly outperforms that achieved with only ECG features [11] (κ = 0.55). Moreover, our results are comparable to those obtained with the Watch-PAT feature set [12] (κ = 0.59), which used features derived from the peripheral arterial tone and oxyhemoglobin information besides to heart rate (i.e., pulse rate) and actigraphy. Note that 227 subjects were recruited in that study, which is much more than our study. Therefore, it is suggested to further test our classifier based on a larger data set with more subjects.

Although the addition of the RAB features results in an overall classification performance improvement, the variance remains high (average κ and average sensitivity over all subjects are 0.59 $\pm$ 0.21 and 68.4 $\pm$ 21.4%, respectively). This can be explained by the large physiological differences between subjects in the way sleep stages are expressed on respiratory and cardiac features. This naturally leads to difficulties in further improving classification performance. Hence, it is worth further investigating how to reduce the between-subject variation of the features.

Furthermore, it was assumed that the respiratory effort area can accurately represent breathing tidal volume when extracting some RAB features. However, this is not always a reasonable assumption, particularly for subjects who change their posture during sleep [19]. In those cases such features can be inaccurate, thus harming classification performance. This challenge should be further studied.

Finally, due to the small size of data set used in this study (only with 14 subjects), we applied LOSOCV to evaluate the classifier instead of dividing the data into training and testing sets. Thus, again, it is suggested to evaluate our classifier on a larger-sized data set with separated training set and testing set in the future.

ACKNOWLEDGMENT

The authors would like to thank Dr. Tim Leufkens from Philips Research Lab for the insightful comments.

REFERENCES

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

[2] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53(3), pp. 485–496, Mar 2006.

[3] *The AASM manual for the scoring of sleep and associated events: rules, terminology & technical specifications*, The American Association of Sleep Medicine, 2007, www.aasmnet.org.

[4] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti, "Sleep staging based on signals acquired through bed sensor," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14(3), pp. 776–785, May 2010.

[5] J. Trinder, *et al.*, "Autonomic activity during human sleep as a function of time and sleep stage," *J. Sleep Res.*, vol. 10(4), pp. 253–264, 2001.

[6] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/wake detection based on cardiorespiratory signals and actigraphy," in *Proc. 32nd Ann. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Buenos Aires, Argentina, Aug. 2010, pp. 5089–5092.

[7] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, "Using dynamic time warping for sleep and wake discrimination," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Inf. (BHI'12)*, Hong Kong and Shenzhen, China, 2012, pp. 886–889.

[8] X. Long, P. Fonseca, R. Haakma, R. M. Aarts, and J. Foussier, "Time-frequency analysis of heart rate variability for sleep and wake," in *Proc. 12th IEEE Int. Conf. BioInf. and BioEng. (BIBE)*, Larnaca, Cyprus, 2012, pp. 85–90.

[9] *Basic of Sleep Guide*, 2nd ed.: Sleep Research Society, 2011, p. 142.

[10] N. J. Douglas, *et al.*, "Respiration during sleep in normal man," *Thorax*, vol. 37(11), pp. 840–844, Nov. 1982.

[11] M. O. Mendez, *et al.*, "Sleep staging from heart rate variability: time-varying spectral features and Hidden Markov Models," *Int. J. Biomed. Eng. Tech.*, vol. 3, pp. 246–263, 2010.

[12] J. Hedner, *et al.*, "Sleep staging based on autonomic signals: a multi-center validation study," *J. Clin. Sleep Med.*, vol. 7(3), pp. 301–306, Jun. 2011.

[13] S. Telser, M. Staudacher, Y. Ploner, A. Amann, H. Hinterhuber, and M. Ritsch-Marte, "Can one detect sleep stage transitions for on-line sleep scoring by monitoring the heart rate variability," *Somnologie*, vol. 8(2), pp. 33–41, May 2004.

[14] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Am. J. Physiol. Heart Circ. Physiol.*, vol. 278, no. 6, pp. H2039–H2049, Jun. 2000.

[15] R. P. Millman, H. Knight, L. R. Kline, E. T. Shore, D. C. Chung, and A. I. Pack, "Changes in compartmental ventilation in association with eye movements during REM sleep," *J. Appl. Physiol.*, vol. 65, no. 3, pp. 1196–1202, Sep. 1988.

[16] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proc. Nat. Inst. Sci. India*, Calcutta, 1936, pp. 49–55.

[17] R. Bakeman and J. M. Gottman, *Observing Interaction: An Introduction to Sequential Analysis*, 2nd ed.: Cambridge University Press, 1997.

[18] S. Herscovici, A. Peer, S. Papyan, and P. Lavie, "Detecting REM sleep from the finger: an automatic REM sleep algorithm based on peripheral arterial tone (PAT) and actigraphy," *Physiol. Meas.*, vol. 28, no. 2, pp. 129-0140, Feb. 2007.

[19] K. F. Whyte, *et al.*, "Accuracy of respiratory inductive plethysmograph in measuring tidal volume during sleep," *J. Appl. Physiol.*, vol. 71, no. 5, pp. 1866–1871, Nov. 1991.