Automatic Detection of Overnight Deep Sleep Based on Heart Rate Variability: A Preliminary Study

Xi Long[1], *Member*, *IEEE*, Pedro Fonseca[1], Reinder Haakma[2], Jérôme Foussier[3], *Member*, *IEEE*, and Ronald M. Aarts[1], *Fellow*, *IEEE*

***Abstract*— This preliminary study investigated the use of cardiac information or more specifically, heart rate variability (HRV), for automatic deep sleep detection throughout the night. The HRV data can be derived from cardiac signals, which were obtained from polysomnography (PSG) recordings. In total 42 features were extracted from the HRV data of 15 single-night PSG recordings (from 15 healthy subjects) for each 30-s epoch, used to perform epoch-by-epoch classification of deep sleep and non-deep sleep (including wake state and all the other sleep stages except deep sleep). To reduce variation of cardiac physiology between subjects, we normalized each feature per subject using a simple Z-score normalization method by subtracting the mean and dividing by the standard deviation of the feature values. A correlation-based feature selection (CFS) method was employed to select informative features as well as removing feature redundancy and a linear discriminant (LD) classifier was applied for deep and non-deep sleep classification. Results show that the use of Z-score normalization can significantly improve the classification performance. A Cohen's Kappa coefficient of 0.42 and an overall accuracy of 81.3% based on a leave-one-subject-out cross-validation were achieved.**

I. Introduction

With polysomnography (PSG), the "gold standard" for objective sleep assessment, overnight sleep can be classified as wake, rapid-eye-movement (REM) sleep, and one of non-REM (NREM) sleep stages N1, N2 and N3 according to the guidelines of American Academy of Sleep Medicine (AASM) [1]. N3 usually corresponds to slow wave sleep (SWS) or "deep sleep". Deep sleep is the most restorative period of sleep for metabolic function, where body energy can be efficiently conserved and recovered. It is therefore important to detect deep sleep throughout the night from a healthcare point of view.

Cardiac information or more specifically, heart rate variability (HRV), has been proved to correlate with autonomic nervous system where autonomic activity differs across sleep stages [2]. For example, deep sleep is in association with decreased sympathetic activity which is reflected by the HRV low-frequency power [3]. In this matter, HRV information can thus be in turn used to detect deep sleep.

HRV data have been more and more considered for sleep staging or sleep stage detection [4], [5]. This is because, compared with traditional PSG, HRV data can be acquired easier or more unobtrusively with, e.g., photoplethysmography [6] and balistocardiograhy [7]. Many studies have investigated classifications between sleep and wake [8], between REM and NREM sleep [9], and between wake, REM and NREM sleep [4]. However, detecting deep sleep has not been well studied. Shinar *et al.* [10] developed an HRV-based deep sleep detector and achieved an overall accuracy of ∼80%, but they only chose a very small portion (a total deep and non-deep sleep duration of 50 minutes each) of the whole-night recordings from all subjects for classification. In this study, we addressed the problem of overnight deep and non-deep sleep classification using solely HRV data.

We extracted a total of 42 features from the HRV data. These features were computed on each 30-s interval (or epoch) based on the AASM guidelines [1]. For each subject, the values of each feature were normalized to have a zero mean and unit variance (i.e., Z-score normalization), aiming at reducing the between-subject variation reflected by the features (caused by the difference in cardiac physiology). This was expected to help improve the deep and non-deep sleep classification.

A linear discriminant (LD) classifier was simply adopted in this work since it has been previously shown to be an appropriate method in sleep staging or sleep stage detection using HRV data [4], [5], [8].

II. Materials and Methods

A. Data Set

Single-night PSG data of 15 healthy subjects were included in our data set. They had a Pittsburgh Sleep Quality Index (PSQI) of less than 6 [11]. Nine subjects were monitored (Alice 5 PSG, Philips Respironics) in Boston, USA, during 2009 at the Sleep Health Center and six were measured (Vitaport 3 PSG, TEMEC) in Eindhoven, the Netherlands, during 2010 at the High Tech Campus. Each subject provided an informed consent and the study protocol was approved by the Ethics Committee of the two sleep laboratories. The PSG recordings are comprised of multi-channel signal modalities such as electroencephalogram (EEG), electromyogram (EMG), electroocculogram (EOG), electrocardiogram (ECG), oxygen saturation, and respiratory effort. From the PSG recordings, only the ECG data (modified lead II, sampled at 500 Hz) were used for deep sleep detection. We clipped each PSG recordings to the time interval from the moment when the subject turned off the lights with the intention of sleep until the moment

[1]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, 5656 AE Eindhoven, The Netherlands x.long@tue.nl.

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

[3]J. Foussier is with the Chair for Medical Information Technology (MedIT), RWTH Aachen University, Pauwelsstrasse 20, 52074 Aachen, Germany foussier@hia.rwth-aachen.de.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

the lights were turned on before this subject got out of bed in the morning.

Sleep stages were manually scored as wake, REM sleep, and N1-N3 of NREM sleep on each 30-s epoch by sleep experts based on the multi-channel bio-signals of PSG according to the AASM guidelines [1]. To perform deep and non-deep sleep classification, N3 was considered the deep sleep class (DS); and wake, REM, N1, and N2 sleep were merged into a single non-deep sleep class (NDS). Table I summarizes the subject demographics and sleep statistics.

TABLE I
SUBJECT DEMOGRAPHICS AND SLEEP STATISTICS

Parameter	Mean ± Std
Gender	5 males and 10 females
Age (years)	31.0 ± 10.4
Body mass index (kg/m^2)	24.4 ± 3.3
Total recording time (hours)	7.2 ± 1.1
Sleep efficiency (%)	92.3 ± 3.8
Deep sleep (%)	20.4 ± 9.2
Non-deep sleep (%)	79.6 ± 9.2

B. HRV Features

The heart beats of an ECG signal (high-pass filtered with a cut-off frequency of 0.8 Hz and normalized in regard to mean and amplitude) were identified with an R-peak detector based on the algorithm proposed by Hamilton and Tompkins [12], resulting in inter-beat intervals or an HRV series. It was then re-sampled at a sampling rate of 4 Hz using linear interpolation. The ectopic RR intervals that are longer than 2 s or shorter than 0.3 s (possibly caused by, e.g., motion artifacts) were excluded. Here 42 epoch-based HRV features (known from literature) were extracted. They are time-domain and frequency-domain features [4] and non-linear features measured by multi-scale sample entropy [13] and detrended fluctuation analysis [14].

C. Feature Selection

We applied a correlation-based feature selection (CFS) algorithm [15] to reduce feature dimensionality and meanwhile remove correlated features. CFS is a filter-based algorithm taking the correlation between features and between features and classes into account. It towards finding an 'optimal' feature subset containing features that are as much as possible uncorrelated with each other and highly correlated with class. The heuristic evaluation criterion of a feature subset F containing k features based on CFS is given by

$$M_{F,k} = \frac{k \cdot \rho_{cf}}{\sqrt{k + k \cdot (k-1) \cdot \rho_{ff}}}, \tag{1}$$

where $M_{F,k}$ represents the "merit" of the feature subset F, ρ_{cf} is the mean feature-to-class correlation, and ρ_{ff} the mean correlation among features. Starting with no features in the subset, a forward search can be used to combine additional features one-by-one until no increase of merit was observed when in combination with them. More details of the CFS algorithm can be found elsewhere [15].

D. Feature Normalization

All the features were normalized via a Z-score method for each subject. Let us consider a feature x_s from subject s containing feature values of n epochs throughout the night, the normalized feature values can be computed such that

$$\hat{x}_s = \frac{x_s - \mu_{x_s}}{\sigma_{x_s}}, \tag{2}$$

where μ_{x_s} and σ_{x_s} are the mean and the standard deviation of x_s, respectively. As mentioned before, the use of subject-specific (Z-score) normalization should enable reduction of between-subject variation evoked by their physiological difference to a certain extent.

E. Feature Separability

We used a Mahalanobis distance metric MD to assess the separability of each feature between classes. For a single feature x, its separability is given by

$$\mathrm{MD}_x = \frac{|\mu_x^{\mathrm{DS}} - \mu_x^{\mathrm{NDS}}|}{\sigma_x}, \tag{3}$$

where μ_x^{DS} and μ_x^{NDS} represent mean values of DS and NDS, respectively, and σ_x is the population standard deviation of this feature. A higher feature separability of in discriminating between the two classes is indicated by a larger MD value.

F. Deep Sleep Detection

A well-known LD classifier was adopted to classify deep and non-deep sleep on an epoch-by-epoch basis. Conventional metrics sensitivity, specificity, precision, and overall accuracy were first used in a binary classification to assess the classification performance. In addition to these, we also utilized the Cohen's Kappa coefficient of agreement. It is considered a better metric when class distribution is imbalanced (here DS epochs account for an average of approximately 20% of the night which is much less than NDS epochs). Note that in this study DS and NDS were considered the positive and the negative class, respectively. To compare the classifiers across the entire solution, we used the Receiver Operating Characteristic (ROC) curve which plots sensitivity (true positive rate) versus 1-specificity (false positive rate) on a graph. The 'area under the ROC curve' (AUROC) was then computed as a single metric that quantifies the classification performance in the solution space. A better classification performance corresponds to a larger AUROC value.

It is known that LD classifier is sensitive to prior probability of each class. A time-varying prior probability (TVPP) has been successfully used for classifying wake, REM sleep and NREM sleep [4], [5]. This was based on the observation that the probabilities of different classes change over time throughout the night. Similarly, the TVPP of DS and NDS for each epoch was obtained by computing the percentage that specific epoch was labeled as each class with respect to time (or epoch index) based on training set [4].

In order not to bias the classification results, experiments were conducted using a leave-one-subject-out cross-validation (LOSOCV) to evaluate the classifier. During each iteration of the cross-validation, data from 14 subjects were used to train the classifier and the data from the remaining subject were used for testing. Afterwards, results of all testing sets were then averaged, yielding the overall classification performance. Note that feature selection was performed on each training set of the cross-validation, resulting in 15 feature subsets. To assemble a single feature set for evaluation and avoid biasing the results, only the features included in all those feature subsets were eventually selected and then used to test the classifier using LOSOCV.

III. Results and Discussion

After using CFS during the cross-validation procedure, three HRV features were selected for deep and non-deep sleep classification. They are: 1) SDNN, the standard deviation of inter-beat intervals, 2) MRF, the mean respiratory frequency estimated from HRV which corresponds to the frequency of spectral peak in the high-frequency band between 0.15 Hz and 0.4 Hz, and 3) PMRF, the power of MRF. The separabilities of these three features without and with applying subject-specific Z-score normalization are compared in Table II. It can be seen that normalizing the features per subject clearly increases their separability (as measured by the Mahalanobis distance MD).

TABLE II

Feature Separability (MD) of Selected Features with and without Subject-specific (Z-score) Feature Normalization

Selected feature →	SDNN	MRF	PMRF
Without normalization	0.60	0.55	0.64
With normalization	0.93	0.61	0.81

Note: Results are pooled over all 15 subjects.

Table III presents the overnight deep and non-deep classification results (obtained through the LOSOCV) using the selected HRV features. In the table, the precision, sensitivity, specificity, accuracy, Kappa, and AUROC are shown and compared with and without using subject-specific (Z-score) normalization. After applying the normalization, an average Kappa of 0.42 $\pm$ 0.16 (versus 0.35 $\pm$ 0.22) and an average accuracy of 81.3 $\pm$ 3.5% (versus 79.2 $\pm$ 7.6%) were achieved. To examine the significance of their differences, a paired Wilcoxon signed-rank test (one-sided) was used accordingly. We notice that the normalization can significantly improve the performance of deep sleep detection. Moreover, the variances of the results decrease after using the proposed Z-score feature normalization method for each subject, which indicates that this method can help reducing the between-subject variations to some extent. As shown in Figure 1, the ROC curves of deep sleep detection obtained with and without the normalization are compared in a two-dimensional solution space. In the figure, we also observe a performance enhancement after normalizing the features for each subject.

TABLE III

LOSOCV Results of Overnight Deep Sleep Detection with and without Subject-specific (Z-score) Feature Normalization

Metric	Without normalization	With normalization
Precision (%)	50.3 $\pm$ 26.1	53.1 $\pm$ 22.3
Sensitivity (%)	52.6 $\pm$ 24.8	59.6 $\pm$ 15.2
Specificity (%)	87.2 $\pm$ 11.0	87.6 $\pm$ 5.7
Accuracy (%)	79.2 $\pm$ 7.6	81.3 $\pm$ 3.5$^{\text{ns}}$
Kappa (-)	0.35 $\pm$ 0.22	0.42 $\pm$ 0.16*
AUROC (-)	0.80 $\pm$ 0.07	0.84 $\pm$ 0.07*

Note: Three features SDNN, MRF and PMRF selected by CFS were used. Results are averaged over all subjects.
*Significance of difference was examined with a Wilcoxon signed-rank test on accuracy, Kappa, and AUROC, at $p<0.05$. ns: not significant.

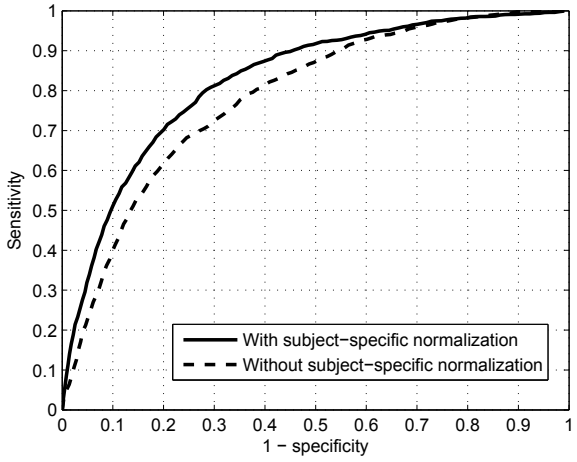

Fig. 1. ROC curves of deep and non-deep sleep classification obtained with (solid) and without (dash) subject-specific Z-score normalization.

The confusion matrix obtained with the Z-score normalization is shown in Table IV, where the misclassifications of different sleep stages and wakefulness are indicated. It can be clearly observed that around 17.5% of N2 epochs were misclassified as DS, which implies a presence of difficulty in discriminating between N2 and deep sleep based on HRV data. As a matter of fact, it has been shown that N2 and N3 sleep are very similar in regard to autonomic nervous activity [2]. This would result in performance limitation in classifying deep and non-deep sleep. Nevertheless, further explorations are encouraged in better separating these two sleep stages through the use of cardiac activity.

Although the Z-score normalization is able to improve the classification performance by reducing between-subject variations manifested by cardiac activity, it might not be the most appropriate method. For example, it is sensitive to outliers of feature values. Some other normalization methods which are robust to feature outliers (e.g., quantile normalization [16] and winsorization [17]) merit further investigation. In addition to this, the Z-score normalization assumes that all the subjects have similar proportions of sleep stages over night. However, this might not always hold, particularly for subjects in different age groups [18]. Therefore, a distribution

TABLE IV

CONFUSION MATRIX OF DEEP SLEEP DETECTION USING LOSOCV WITH SUBJECT-SPECIFIC (Z-SCORE) FEATURE NORMALIZATION

Confusion matrix		Classification	
		DS	NDS
PSG	DS	1,456	1,079
	NDS	1,274	8,787
	(Wake)	(53)	(881)
	(REM)	(100)	(2,021)
	(N1)	(63)	(906)
	(N2)	(1,058)	(4,979)

normalization method (e.g., histogram equalization [19]) might be an option to deal with this issue.

As shown in Table V, compared with the deep sleep detector developed by Shinar *et al.* [10] using HRV data, we achieved a slightly better performance (with an overall accuracy of 81.3% versus 79.5%). Hedner *et al.* [20] evaluated a sleep stager and obtained a Kappa of 0.48 for classifying deep and non-deep sleep, which is better than our result. However, they used signal modalities including peripheral arterial tone (PAT), oxyhemoglobin saturation (OS), and actigraphy (AC). Using additional signal modalities which are able to be unobtrusively acquired (e.g., respiratory effort), is therefore expected to help obtain a better performance in detecting deep sleep. This will be studied in future work.

TABLE V

COMPARISON OF DEEP SLEEP DETECTION PERFORMANCE

Study →	Hedner *et al.* [20]	Shinar *et al.* [10]	This paper
Modality	PAT†, OS‡, AC§	HRV	HRV
# Subjects	227	34	15
# Epochs	198,815	200♯	12,596
Accuracy*	88.5%	79.5%	81.3%
Kappa*	0.48	–	0.42

†Peripheral arterial tone; ‡Oxyhemoglobin saturation; §Actigraphy.
♯Results of only 200 epochs (100 deep sleep epochs) were presented.
*Results were re-computed based on the reported confusion matrix.

IV. CONCLUSION

An overnight deep sleep detector based on cardiac activity was developed. A total of 42 features were extracted from the HRV series for each 30-s epoch and three features were selected using the CFS feature selection method. By normalizing (Z-score) the feature values over the entire night for each subject, the difference between subjects in physiology manifested by the features can be reduced to some extent. This can yield deep sleep detection results that are superior to those obtained without performing the subject-specific normalization on the features. With the normalization, we achieved a Cohen's Kappa coefficient of 0.42 and an overall accuracy of 81.3% in classifying deep and non-deep sleep, tested with an LOSOCV on an LD classifier. In addition, we found that most of the misclassified deep sleep epochs are in N2 sleep.

ACKNOWLEDGMENT

The authors would like to thank Le An from the University of California, Riverside for his insightful comments.

REFERENCES

[1] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, American Academy of Sleep Medicine (www.aasmnet.org), 2007.

[2] J. Trinder, J. Kleiman, M. Carrington, S. Smith, S. Breen, N. Tan, and Y. Kim, "Autonomic activity during human sleep as a function of time and sleep stage," *J. Sleep Res.*, vol. 10, no. 4, pp. 253–264, 2001.

[3] V. K. Somers, M. E. Dyken, A. L. Mark, and F. M. Abboud, "Sympathetic-nerve activity during sleep in normal subjects," *New Engl. J. Med.*, vol. 328, no. 5, pp. 303–307, 1993.

[4] S. J. Redmond, P. d. Chazal, C. O'Brien, S. Ryan, W. T. McNicholas, and C. Heneghan, "Sleep staging using cardiorespiratory signals," *Somnologie*, vol. 11, pp. 245–256, 2007.

[5] X. Long, P. Fonseca, R. Haakma, R. M. Aarts, and J. Foussier, "Spectral boundary adaptation on heart rate variability for sleep and wake classification," *Int. J. Artif. Intell. T.*, vol. 23, no. 3, pp. 1 460 002:1–20, 2014.

[6] S. Lu, H. Zhao, K. Ju, K. Shin, M. Lee, K. Shelley, and K. H. Chon, "Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information?" *J. Clin. Monit. Comput.*, vol. 22, no. 1, pp. 23–29, 2008.

[7] C. Bruser, K. Stadlthanner, S. de Waele, and S. Leonhardt, "Adaptive beat-to-beat heart rate estimation in ballistocardiograms," *IEEE Trans. Inf. Tech. Biomed.*, vol. 15, no. 5, pp. 778–786, 2011.

[8] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, "Sleep and wake classification with actigraphy and respiratory effort using dynamic warping," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, 2014.

[9] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R. M. Aarts, "Respiration amplitude analysis for REM and NREM sleep classification," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, Osaka, Japan, 2013, pp. 5017–5020.

[10] Z. Shinar, A. Baharav, Y. Dagan, and S. Akselrod, "Automatic detection of slow-wave-sleep using heart rate variability," in *Computers in Cardiology 2001*. Rotterdam, Netherlands: IEEE Computer Society Press, 2001, pp. 403–406.

[11] D. J. Buysse, C. F. Reynolds-III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The pittsburgh sleep quality index: a new instrument for psychiatric practice and research." *Psych. Res.*, vol. 28, no. 2, pp. 193–213, 1989.

[12] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," *IEEE Trans. Biomed. Eng.*, vol. 33, no. 12, pp. 1157–1165, 1986.

[13] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of biological signals," *Phys. Rev. E*, vol. 71, no. 2, pp. 021 906:1–18, 2005.

[14] T. Penzel, J. W. Kantelhardt, L. Grote, J. H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 10, pp. 1143–1151, 2003.

[15] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, New Zealand, 1999.

[16] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[17] M. G. Hahn, J. Kuelbs, and D. C. Weiner, "The asymptotic distribution of magnitude-winsorized sums via self-normalization," *J. Theor. Prob.*, vol. 3, no. 1, pp. 137–168, 1990.

[18] M. M. Ohayon, M. A. Carskadon, C. Guilleminault, and M. V. Vitiello, "Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan," *Sleep*, vol. 27, no. 7, pp. 1255–1273, 2004.

[19] V. T. Tom and G. J. Wolfe, "Adaptive histogram equalization and its applications," in *Proc. SPIE 1982*, San Diego, 1982, pp. 204–209.

[20] J. Hedner, D. P. White, A. Malhotra, S. Herscovici, S. D. Pittman, D. Zou, L. Grote, and G. Pillar, "Sleep staging based on autonomic signals: a multi-center validation study," *J. Clin. Sleep Med.*, vol. 7, no. 3, pp. 301–306, 2011.