PAPER • **OPEN ACCESS**

Sleep stage classification with ECG and respiratory effort

OPEN ACCESS
IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement
Physiol. Meas. **36** (2015) 2027–2040 doi:10.1088/0967-3334/36/10/2027

Sleep stage classification with ECG and respiratory effort

Pedro Fonseca[1,2,4], Xi Long[1,2,4], Mustafa Radha[1], Reinder Haakma[1], Ronald M Aarts[1,2] and Jérôme Rolink[3]

[1] Philips Research, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
[2] Department of Electrical Engineering, Eindhoven University of Technology, Postbus 513, 5600MB Eindhoven, The Netherlands
[3] Philips Chair for Medical Information Technology, RWTH Aachen University, Pauwelsstrasse 20, D-52074 Aachen, Germany

E-mail: pedro.fonseca@philips.com

Received 8 April 2015, revised 3 June 2015
Accepted for publication 12 June 2015
Published 19 August 2015

Abstract

Automatic sleep stage classification with cardiorespiratory signals has attracted increasing attention. In contrast to the traditional manual scoring based on polysomnography, these signals can be measured using advanced unobtrusive techniques that are currently available, promising the application for personal and continuous home sleep monitoring. This paper describes a methodology for classifying wake, rapid-eye-movement (REM) sleep, and non-REM (NREM) light and deep sleep on a 30 s epoch basis. A total of 142 features were extracted from electrocardiogram and thoracic respiratory effort measured with respiratory inductance plethysmography. To improve the quality of these features, subject-specific Z-score normalization and spline smoothing were used to reduce between-subject and within-subject variability. A modified sequential forward selection feature selector procedure was applied, yielding 80 features while preventing the introduction of bias in the estimation of cross-validation performance. PSG data from 48 healthy adults were used to validate our methods. Using a linear discriminant classifier and a ten-fold cross-validation, we achieved a Cohen's kappa coefficient of 0.49 and an accuracy of 69% in the classification of wake, REM, light, and deep sleep. These values increased to kappa = 0.56 and accuracy = 80% when the classification problem was reduced to three classes, wake, REM sleep, and NREM sleep.

[4] Joint first authors.

Keywords: sleep staging, ECG, respiratory effort, feature selection

(Some figures may appear in colour only in the online journal)

1. Introduction

A problem with traditional sleep monitoring, known as polysomnography (PSG), is that a wide array of potentially sleep-disturbing sensors must be applied to the body and their measurements can only be interpreted by highly trained sleep technicians or scientists. Albeit invaluable in the diagnostic of sleep disorders, traditional PSG is rather ill-suited for regular, non-diagnostic monitoring of sleep and will only introduce more sleep disturbances when applied on a daily basis by untrained individuals. This scenario makes apparent a need for unobtrusive methods of sleep monitoring, preferably inexpensive and with no training required to operate them. Cardiorespiratory monitoring can be unobtrusive and the data can be analyzed by a computer, which makes this technology a promising candidate for personal, continuous and unobtrusive sleep monitoring.

Cardiorespiratory sleep staging or sleep stage classification is often based on heart rate variability (HRV) calculated from electrocardiogram (ECG) and respiratory effort, often from respiratory inductance plethysmography (RIP). Usually cardiorespiratory information is combined with body movements from an accelerometer to more accurately distinguish wake from sleep. One of the earliest studies that presented a successful machine learning approach to cardiorespiratory sleep stage classification with these modalities was done by Redmond and Heneghan (2006). Using a set of HRV features to model the autonomic nervous activity and a set of respiratory features to model the parasympathetic tone, Redmond and colleagues showed the viability of a sleep stage classifier that can generate a simplified hypnogram for an entire night indicating, for each 30 s segment, a sleep stage, classified as either wake, rapid-eye-movement (REM) sleep, or non-REM (NREM) (wake-REM-NREM or WRN classification for short). More recent research has shown that it is possible to obtain the same cardiorespiratory information from other sensors for sleep stage classification, such as from bed-mounted ballistocardiogram (Watanabe and Watanabe 2004, Kortelainen *et al* 2010) or contactless radio frequency (de Chazal *et al* 2011). Although these studies focused on distinction between wake, REM sleep, and NREM sleep (without separating NREM sleep in other sleep stages) or between wake and sleep (merging REM and NREM sleep), these attempts promised that cardiorespiratory methods could one day be completely unobtrusive.

In previous work (Long *et al* 2014d, 2014) we proposed methods to simultaneously classify wake, REM sleep, light sleep (NREM stage S1 and S2), and deep sleep or slow wave sleep (stage S3 and S4) using respiratory activity in order to estimate an overnight wake-REM-light-deep sleep (WRLD) hypnogram. In comparison with WRN classification, achieving WRLD classification would allow a more adequate assessment of sleep since deep sleep is thought by some researchers to be important in several cortical and physiological processes, such as energy conservation (Berger and Phillips 1995), cerebral restoration (Benington and Heller 1995), and memory processing and consolidation (together with sleep spindles which occur during NREM stage 2) (Stickgold 2005, Walker 2009). In that work, we also reviewed the state-of-the-art in sleep stage classification with cardiac and/or respiratory activity. The methods presented there will be used to benchmark the method proposed in this paper. In addition, two studies with comparable results have been proposed by Domingues *et al* (2014) and Willemen *et al* (2014). However, these works only report results on a three-class task (WRN classification) or use non-standard one-minute epochs for classification, respectively.

This paper presents a methodology for automatic sleep stage classification based on machine learned models of the autonomic nervous system during sleep from ECG and RIP signals. Compared to previous studies, our methodology includes novel features, new feature post-processing methods, and a refined feature selection method which guarantees that no bias is introduced in the validation of the algorithm while avoiding the use of a hold-out validation set. These methods are applied for three-class (WNR) and four-class (WRLD) sleep stage classification of healthy subjects.

2. Materials and methods

2.1. Data sets

The data set was the same as used in earlier work (Long *et al* 2014d, 2014) and comprised full single-night polysomnographic (PSG) recordings of 48 subjects (27 females) acquired in the SIESTA project (Klosch *et al* 2001). All subjects were healthy sleepers with a Pittsburgh Sleep Quality Index (Buysse *et al* 1989) of less than 6 and had no regular sleep complaints nor earlier diagnosis of sleep disorders. The subjects had an average age of 41.3($\pm$16.1) years at the time of the recording. Full subject demographics can be found in our earlier work (Long *et al* 2014d). Sleep stages were scored by trained sleep technicians in six classes according to the R and K rules (Rechtschaffen and Kales 1968). In the scope of this study, S1 and S2 were merged in a single L (light sleep) class and S3 and S4 were merged in a single D (deep sleep) class. Each PSG recording comprised, besides the standard signals required for sleep scoring, modified lead II ECG, and (thoracic) respiratory effort recorded with respiratory inductance plethysmography (RIP). QRS complexes were detected and localized from ECG signals using a combination of a Hamilton–Tompkins detector (Hamilton and Tompkins 1986, Hamilton 2002) and a post-processing localization algorithm (Fonseca *et al* 2014). Prior to feature extraction, RIP signals were filtered with a 10th order Butterworth low-pass filter with a cut-off frequency of 0.6 Hz, after which baseline was removed by subtracting the median peak-to-through amplitude (Long *et al* 2014d).

2.2. Feature extraction

We extracted a set of 142 features from cardiac and respiratory activity, and from cardiorespiratory interaction (CRI) using a sliding window centered on each 30 s epoch, guaranteeing sufficient data to capture the changes in autonomic activity (Malik *et al* 1996). Since some features are computed based on windows which exceed the epoch length, epochs at the start and end of each recording required a special handling: for each such feature, all epochs for which the window crosses the boundaries of the recording were marked as invalid; the feature values for these epochs were interpolated during post-processing using spline fitting (section 2.2.4).

2.2.1. Cardiac features. Considering cardiac activity, 86 cardiac features were computed from the QRS complexes detected in the ECG signal. Time domain features, computed over nine epochs, include mean heart rate, mean heartbeat interval (detrended and non-detrended), standard deviation (SD) of heartbeat intervals, difference between maximal and minimal heartbeat intervals, root mean square and SD of successive heartbeat interval differences, and percentage of successive heartbeat intervals differing by >50 ms (Malik *et al* 1996, Redmond *et al* 2007). We also computed the mean absolute difference and different percentiles (at 10%, 25%, 50%, 75%, and 90%) of detrended and non-detrended heart rates and heartbeat intervals (Ylmaz *et al* 2010, Willemen *et al* 2014) as well as the mean, median, minimal, and maximal

likelihood ratios of heart rates (Basner *et al* 2007). In the frequency domain, the features included the logarithmic spectral powers in the very low frequency band (VLF) from 0.003 to 0.04 Hz, in the low frequency band (LF) from 0.04 to 0.15 Hz, in the high frequency band (HF) between 0.15 to 0.4 Hz, and the LF-to-HF ratio (Busek *et al* 2005), where the power spectral densities were estimated over nine epochs. The spectral boundaries were adapted to the corresponding peak frequency, yielding their boundary-adapted versions (Long *et al* 2014c). We also computed the maximum module and phase of HF pole (Mendez *et al* 2010) and the maximal power in the HF band and its associated frequency representing respiratory rate (Redmond *et al* 2007). Features describing non-linear properties of heartbeat intervals were quantified with detrended fluctuation analysis (DFA) over 11 epochs (Kantelhardt *et al* 2001) and its short-term (α_1), long-term (α_2), and all time scaling exponents (Iyengar *et al* 1996, Penzel *et al* 2003), progressive DFA with non-overlapping segments of 64 heartbeats (Telser *et al* 2004), windowed DFA over 11 epochs (Adnane *et al* 2012), and multi-scale sample entropy (MSE) over 17 epochs (length of 1 and 2 samples with scales of 1–10) (Costa *et al* 2005). Approximate entropy of the symbolic binary sequence that encodes the increase or decrease in successive heartbeat intervals over nine epochs was also calculated (Cysarz *et al* 2000). In addition, we propose new features based on a visibility graph (VG) and a difference VG (DVG) method to characterize HRV time series in a two-dimensional complex network where samples are connected as nodes in terms of certain criteria (Lacasa *et al* 2008, Long *et al* 2014a). The network-based features, computed over seven epochs, comprised the mean, SD, and slope of node degrees and number of nodes in VG- and DVG-based networks with a small degree ($\leqslant$3 for VG and $\leqslant$2 for DVG) and a large degree ($\geqslant$10 for VG and $\geqslant$8 for DVG), and assortativity coefficient in the VG-based network (Shao 2010, Long *et al* 2014a, Zhu *et al* 2014).

2.2.2. Respiratory features. Concerning respiratory activity, 44 features were derived from RIP signals. In the time domain, we estimated the variance of the respiratory effort signal, the respiratory frequency and its SD over 150, 210, and 270 s, the mean and SD of breath-by-breath correlation, and the SD in breath length (Redmond *et al* 2007). One of our previous studies (Long *et al* 2014d) introduced respiratory amplitude features for sleep stage classification, including the standardized mean, standardized median, and sample entropy of respiratory peaks and troughs (indicating inhalation and exhalation breathing depth, respectively), median peak-to-trough difference, median volume and flow rate for complete breath cycle, inhalation, and exhalation, and inhalation-to-exhalation flow rate ratio. These features were adopted in this work. Besides, we also computed the similarity between the peaks and troughs by means of the envelope morphology using a dynamic time warping (DTW) metric (Berndt and Clifford 1994). From the respiratory spectrum, the respiratory frequency and its power, the logarithm of the spectral power in VLF (0.01–0.05 Hz), LF (0.05–0.15 Hz), and HF (0.15–0.5 Hz) bands, and the LF-to-HF ratio were estimated (Redmond and Heneghan 2006). Respiratory regularity was measured by means of sample entropy over seven epochs (Richman and Moorman 2000, Long *et al* 2014d) and self-(dis)similarity based on DTW and dynamic frequency warping (DFW) (Long *et al* 2014b) and uniform scaling (Long *et al* 2014) were derived. The same network analysis features as for HRV were also computed for breath-to-breath intervals.

2.2.3. Cardiorespiratory interaction features. Numerous studies have shown that the interaction between cardiac and respiratory activity varies across sleep stages (Ichimaru *et al* 1990, Cysarz *et al* 2004, Long *et al* 2014a). The power associated with respiratory-modulated heartbeat intervals was quantified over windows of nine epochs (Ichimaru *et al* 1990). In addition,

we also extracted the VG- and DVG-based features for CRI (Long *et al* 2014a). These resulted in a total of 12 CRI features in our feature set.

2.2.4. Feature post-processing. In order to reduce the impact of physiological differences and equipment-related variations from subject to subject, the features of each subject were first Z-score normalized by subtracting their mean and dividing by their SD. Further, it is known that the sleep pattern of healthy adults progresses with several cycles throughout the night (Carskadon and Dement 2011). For example, REM and NREM sleep alternate with 4–6 cycles of about 90–110 min with deep sleep usually dominating the NREM periods during the first half of the night. This suggests that the autonomic physiological response with its associated sleep stage is time-variant across the night for each subject. For this reason, we were motivated to smoothen each feature for each subject by means of a cubic spline fitting method (De Boor 2001). This is also expected to help reduce signal measurement noise and variability within subjects for each sleep stage conveyed by the feature values. The latter can be caused by body movements, conscious breathing control, internal physiological variations, or other external factors such as changes in environmental noise and temperature during bedtime sleep. Instead of other simpler low-pass filters, spline fitting was chosen since it can interpolate feature values which could not be computed, for example due to motion artifacts (about 10% observed in our data set) or at the start and end of each recording for features computed with windows exceeding the epoch duration. This procedure allows all epochs in each recording to be classified.

Let $\mathbf{t}$ represent a sequence of feature values $\mathbf{v} = \{v_1, v_2, ..., v_n\}$ at their corresponding time (or epoch) indices $\mathbf{t} = \{t_1, t_2, ..., t_n\}$ (in 30 s), then a relation between them can be modeled by

$$v_i = h(t_i) + \varepsilon_i \quad (i = 1, 2, ..., n), \tag{1}$$

where h is a smoothing (spline) function, ε_i are independent and identically distributed residuals. The smoothing function can be estimated by minimizing the objective function (i.e. penalized sum of square) such that

$$\hat{h} = \arg\min_{h} \left[\sum_{i=1}^{n} [v_i - h(t_i)]^2 + \lambda \int_{t_1}^{t_n} h''(t)^2 \mathrm{d}t \right], \tag{2}$$

where λ is a smoothing parameter that controls the trade-off between residual and local variation. The smoothing function can be expressed by cubic B-splines as basis functions and determined via least squares approximation (Unser 1999, De Boor 2001).

For a specific overnight recording with a total of m epochs, it is divided in s continuous segments ($s = \lceil m/n \rceil$), designated as smoothing splines. Each segment can then be modeled by the spline function, yielding a general spline fitting for the epochs over the entire recording. n represents the smoothing window size where a larger n translates to a smoother fitting curve. In this work, a window size of nine epochs for modeling splines was experimentally found to be appropriate for the task of sleep stage classification.

2.3. Classifier

This work used a multi-class Bayesian linear discriminant with time-varying prior probabilities (Redmond *et al* 2007), similar to that used in previous work (Long *et al* 2014d). For each epoch, the selected class (D, L, R, or W) is the class ω_i that maximizes the posterior probability given an feature vector $\mathbf{x}$ (Duda *et al* 2000),

$$\omega_i(\mathbf{x}) = \arg\max_{i} [\mathrm{g}_i(\mathbf{x})] \tag{3}$$

with the the discriminant function g_i for each class given by

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^t \Sigma^{-1}(\mathbf{x} - \mu_i) + \ln P(\omega_i, t) \quad (4)$$

where μ_i is the average feature vector for class i, Σ is the pooled covariance matrix for all classes, and $P(\omega_i, t)$ is the prior probability for class i at time (since lights off) t. All parameters were estimated during training.

2.4. Feature selection

To select the final list of features we used a wrapper feature selection method based on sequential forward selection (SFS) (Whitney 1971) using as criterion the Cohen's kappa coefficient of agreement κ (Cohen 1960) on the training set. This measure of agreement between the classification predictions and the ground-truth annotations is more adequate than traditional measures of accuracy for this problem since there is a strong imbalance between classes (L epochs, for example, account for more than 50% of all epochs in the data set) and this coefficient factors out chance agreement, compensating for class imbalance.

In many machine learning studies supervised feature selection is often applied on the entire data set, even if the training and validation are kept separate (for example using cross-validation). This common pitfall is known to introduce a bias in the evaluation of a classifier's performance, which will often be overestimated (Smialowski *et al* 2010). Although keeping a hold-out set for validation would solve this problem, the limited size of the data set would either mean that the model learning would be based on potentially insufficient examples, or that the classifier would be evaluated on a very small sample, potentially unrepresentative of the problem at hand. Instead, the feature selection procedure was executed by strictly separating, on an iterative procedure akin to cross-validation, the training and validation sets. For each iteration, unbound SFS was applied using as criteria the classification performance obtained in the training set of each iteration. The final number of features was chosen as the smallest number S that yield a certain percentage of the maximum training kappa obtained across all iterations. The final list of selected features was chosen as the S features most often selected during the process.

The discriminative power of selected features was evaluated with the absolute standardized mean distance (ASMD) between the feature values of two classes, computed as

$$\mathrm{ASMD} = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sigma} \right| \quad (5)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means for class 1 and 2 and σ is the pooled sample SD.

2.5. Validation and evaluation

After feature selection is performed and the set of features is chosen, the classification results per subject were evaluated using a ten-fold cross-validation procedure. The kappa coefficient for all subjects in the data set as well as the average and pooled performance were then calculated. In addition to the kappa coefficient, we also computed the traditional metric accuracy, i.e. the percentage of correctly identified epochs. For kappa and accuracy, the results were computed both after pooling the predictions over all epochs of all subjects and after averaging the performance for each subject.

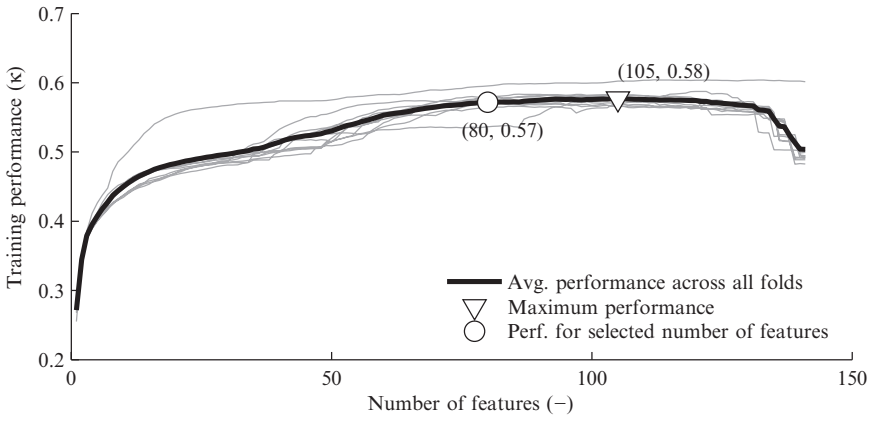

Figure 1. Training performance per iteration and average training performance during feature selection. Maximum performance and performance for the selected number of features are indicated with markers (in parenthesis, the number of features and the corresponding performance).

3. Results and discussion

3.1. Feature selection

Figure 1 indicates the kappa coefficient obtained in the training set of each iteration of the feature selection procedure, for a varying number of features. As illustrated, the maximum average training performance is obtained for 105 features, with an average kappa of 0.58. Also clear in the figure, is a plateau in performance between 70 and 100 features. This suggests that the number of features can be greatly decreased without affecting the training performance. A small feature set is often desirable to prevent over-fitting to the training data, as long as it is not so small that the model cannot learn the characteristics of the problem.

By choosing different operating points in figure 1, we can choose a smaller feature set at the expense of a reduction in training performance. By allowing a reduction of 1% in training performance (from the maximum kappa of 0.58–0.57), we can reduce the number of features from 150 to 80 features (a reduction of 23.8%). Below this number there is a statistically significant (after a Wilcoxon signed-rank test, with $p < 0.05$) decrease in training performance and further decreasing the number of features will likely lead to a decrease in classification performance after cross-validation. Using as criteria the smallest number of features which does not cause a statistically significant decrease below the maximum training performance, we chose a total of $S = 80$ features.

After ranking all features by the number of times they were selected during the iterative feature selection procedure and selecting the 80 features with the highest count, we found that all features in the final feature set were selected in at least 5 of the 10 iterations (with a mean count of 7.67) with 14 features having been selected in all 10 iterations. This illustrates the robustness of the modified SFS method: despite their simplicity and computational efficiency, sequential selection algorithms are known to suffer from a so-called 'nesting effect', potentially leading to sub-optimal feature sets (Pudil *et al* 1994). By iteratively performing several unbound SFS searches on different training sets and keeping only the features that are selected most often, this effect was reduced, as attested by the large number of iterations each feature in the final set was selected.

For brevity only the 14 features selected in all iterations will be discussed further. Table 1 indicates the discriminative power of each feature using the pooled ASMD. It was computed for each pair of classes after aggregating the feature values for all subjects and also the 90th percentile of the ASMD (in parenthesis) obtained for each feature, for all individual subjects. Pooled ASMD values below 0.5 were omitted and 90th percentile ASMD values below 1 were omitted.

The top features are clearly discriminative for different pairs of classes which helps explain the relatively large number of features selected. Additionally, it is interesting to observe that there is one feature (median likelihood ratio) which does not have a pooled ASMD above 0.5 for any class pair. However, its 90th percentile ASMD value is larger than 1 for the pairs D/W and L/W. This is a good example of a feature which is discriminative for only a subset of the subjects (at least 10%) but not for all subjects. The fact that it was selected in every single iteration using the wrapper method described in section 2.4 suggests that it is complementary to other features for some subjects, helping raise the overall training performance.

A note should be made regarding long-term cardiac features such as α_2 or larger-scale MSE features. None of these features were part of the final set of features selected with our method. One possible explanation for this is related to the length of the time series used to compute them. The choice of window sizes (for MSE, 17 epochs, i.e. 8.5 min) represents a compromise between having as much data as possible to accurately calculate the features, while at the same time not exceeding the average length of a given sleep stage (in our data set, the average length of deep and REM sleep periods was found to be 5.1 and 8.7 min, respectively). Although theoretically the window sizes we used are sufficient to calculate these features (a window of 17 epochs corresponds, at an average heart rate of 60 bpm, to 510 samples), it has been suggested that the estimation of sample entropy is low in series shorter than 10^m (where m is the pattern length, in samples) (Richman and Moorman 2000, Yentes *et al* 2013). This means that for a pattern length of $m = 2$ and scales higher than 5, the coarse-grained time series used to calculate the sample entropy will have less data points than the suggested limit and the features might not be accurate, and therefore, not representative of the autonomic characteristics of different sleep stages.

3.2. Cross-validation

Table 2 indicates the overall classification performance obtained after 10-fold cross-validation using the selected set of 80 features. In addition, it indicates the performance per class, obtained by considering each class as the positive class and merging the remaining in a single negative class. The highest performance is obtained for R detection, followed by W. The lowest performance is obtained for L. This is further confirmed by the confusion matrix of table 3 which shows that the largest proportion of errors occurs when trying to distinguish L from the other classes. For all other classes, the percentage of misclassified epochs (relative to the total number of epochs) is below 1% except for L.

In order to evaluate the performance of the classifier in a three-class task (WRN), classes D and L were merged in a single N (non-REM) class. Table 2 indicates the resulting performance. Analyzing the performance of the classifier we see that the classification performance rises substantially, to a kappa of 0.56 and an accuracy of 80%. This was expected since a large number of classification errors occurred between D and L, and in a WNR task these two classes no longer need to be distinguished.

Figure 2 illustrates examples of predicted hypnograms, as compared with the reference, for three subjects in the data set: the subject with the worst performance (with $\kappa = 0.17$), with the median performance (with $\kappa = 0.50$) and with the best performance (with $\kappa = 0.69$). A possible

Table 1. Pooled and 90th percentile ASMD for features selected in all iterations.

Feature	D/L	D/R	D/W	L/R	L/W	R/W
Respiratory features:						
VLF spectral power		0.56	1.02		0.86	0.68
		(1.34)	(1.69)		(1.52)	(1.26)
LF/HF spectral power ratio		0.56	0.85		0.95	0.70
		(1.36)	(1.62)	(1.10)	(1.62)	(1.30)
Frequency SD over 270 s	0.79	1.46	1.41	0.84	0.97	
	(1.20)	(1.82)	(1.87)	(1.38)	(1.67)	(1.09)
Mean breath-by-breath correlation	0.59	1.03	0.82			
	(1.27)	(1.78)	(1.76)	(1.61)	(1.51)	(1.46)
Sample entropy regularity		0.71		0.61	0.55	0.86
		(1.67)	(1.53)	(1.41)	(1.49)	(1.65)
DTW self-dissimilarity		0.59	0.86		0.86	0.56
		(1.62)	(1.58)		(1.68)	(1.39)
Standardized mean of troughs	0.82	1.21	0.97	0.56		
	(1.41)	(1.83)	(1.85)	(1.19)	(1.18)	(1.34)
DTW peak-to-trough similarity						0.55
		(1.06)	(1.34)		(1.04)	(1.38)
Uniform scaling self-dissimilarity	0.92	1.46	1.16	0.85	0.56	
	(1.47)	(1.87)	(1.85)	(1.50)	(1.47)	(1.22)
Cardiac (HRV) features:						
Mean likelihood ratio			0.86			
		(1.50)	(1.60)	(1.09)	(1.46)	(1.23)
Median likelihood ratio						
			(1.19)		(1.23)	
Adapted LF spectral power	0.65	0.88	0.70			
	(1.47)	(1.70)	(1.59)	(1.09)	(1.15)	(1.14)
Assortativity coefficient in VG		0.53				
		(1.34)	(1.11)	(1.22)		(1.44)
Number small-degree nodes in VG			0.59			
	(1.05)	(1.39)	(1.32)	(1.14)	(1.13)	(1.24)

Note: the features are described in section 2.2. The pooled ASMD was computed for each pair of classes after aggregating the feature values for all subjects (values below 0.5 were omitted); The 90th ASMD percentiles (in parentheses) were obtained after computing the ASMD of each feature, for each subject (values below 1 were omitted)

Table 2. Cross-validation performance for 3 and 4 classes.

	Pooled kappa	Pooled acc.	Mean kappa	Mean acc.
WRLD	0.49	0.69	0.49 ± 0.13	0.69 ± 0.08
D	0.51	0.89	0.50 ± 0.17	0.89 ± 0.04
L	0.40	0.71	0.41 ± 0.14	0.71 ± 0.07
R	0.57	0.87	0.58 ± 0.19	0.87 ± 0.08
W	0.54	0.91	0.51 ± 0.18	0.91 ± 0.04
WRN	0.56	0.80	0.56 ± 0.15	0.80 ± 0.08

Note: the pooled performance was computed after aggregating all epochs of all subjects. The mean and SD were calculated based on the performance for each individual subject

Table 3. Confusion matrix after cross-validation.

Pred.↓ ref.→	D		L		R		W	
D	**3431**	(7.6%)	1949	(4.3%)	5	(0.0%)	97	(0.2%)
L	2969	(6.6%)	**19165**	(42.6%)	2947	(6.5%)	2302	(5.1%)
R	86	(0.2%)	2071	(4.6%)	**5383**	(12.0%)	404	(0.9%)
W	31	(0.1%)	952	(2.1%)	243	(0.5%)	**2996**	(6.7%)

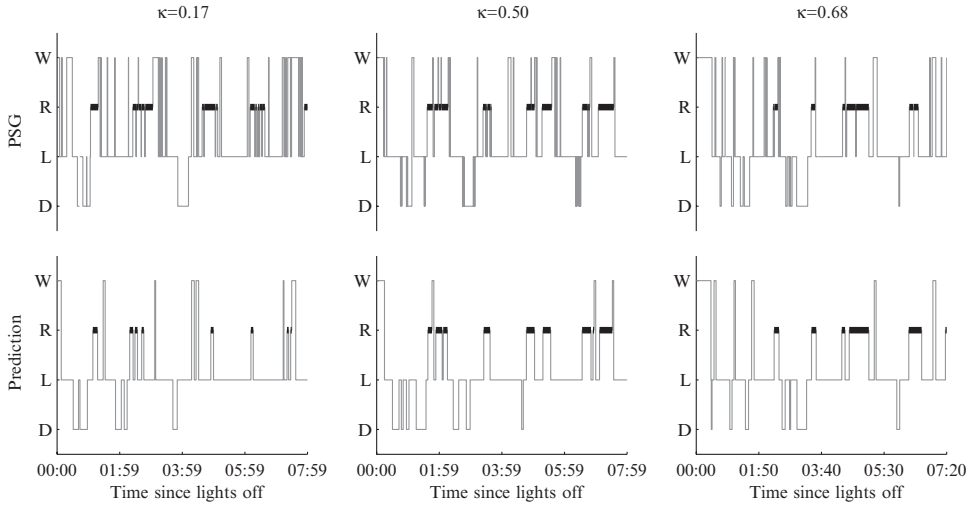

Figure 2. Example of sleep stage reference (top) and predictions (bottom) for the subject with the worst performance (left), with the median performance (middle) and with the best performance (right).

explanation for the poor performance obtained for the worst subject is that the model trained with the characteristics of the general sample population does not fully capture this subject's cardiac and respiratory expression of different sleep stages. However, despite the low kappa coefficient, the predicted hypnogram still exhibits some correct features, namely, most REM intervals were detected, albeit with the incorrect length, and the two deep sleep periods were also detected. As the performance improves, we see that the predicted hypnograms match better the characteristics of the reference hypnogram, and in the best case the most obvious mistakes are in the missed detection of brief periods of wake during the night while the rest of the sleep stages are correctly predicted. This is likely caused by the use of spline smoothing during feature post-processing, which is adequate to capture the slow-changing characteristics of most sleep stages, but penalizes short, abrupt changes such as brief periods of awakening.

3.3. Comparison with state-of-the-art

Table 4 compares the results of our work with other studies reported in literature. As indicated, only a few studies focused on WRLD classification based on cardiac and/or respiratory signals and our results are amongst the best performing. The first observation is that the results of our previous work (Long *et al* 2014), which used only respiratory features, are worse than those

Table 4. Performance comparison with state-of-the-art.

Reference	Modalities[a]	N	Age (year)	Average κ	Average accuracy
WRLD					
Our work	**RIP, ECG**	**48**	**41.3 ± 16.1**	**0.49**	**0.69**
Our previous work[b]	RIP	48	41.3 ± 16.1	0.41	0.65
Isa *et al* (2011)	ECG	16	[32 – 56]	0.26	0.60
Hedner *et al* (2011)	PAT, PO, ACT	227 OSA	49 ± 14	0.48	0.66
Willemen *et al* (2014)[c]	RIP, ECG, ACT	85 (36 subj.)	22.1 ± 3.2	0.56	0.69
WNR					
Our work	**RIP, ECG**	**48**	**41.3 ± 16.1**	**0.56**	**0.80**
Our previous work[b]	RIP	48	41.3 ± 16.1	0.48	0.77
Redmond and Heneghan (2006)	RIP, ECG	37 OSA	46.7 ± 10.4	0.32	0.67
Redmond *et al* (2007)	RIP, ECG	31	42.0 ± 7.4	0.45	0.76
Mendez *et al* (2010)	BCG	22 (11 subj.)	n.a.	0.42	0.72
Kortelainen *et al* (2010)	BCG	18 (9 subj.)	[20 – 54]	0.44	0.79
Migliorini *et al* (2010)	BCG	22 (11 subj.)	n.a.	0.55	0.77
Kurihara and Watanabe (2012)	BCG	20	22.2	0.48	0.78
Xiao *et al* (2013)	ECG	45	[16 – 61]	0.46	0.73
Domingues *et al* (2014)	RIP, ECG, ACT	20	42.1 ± 9	0.58	0.78
Willemen *et al* (2014)[c]	RIP, ECG, ACT	85 (36 subj.)	22.1 ± 3.2	0.62	0.81

[a] ECG: electrocardiography, RIP: respiratory inductance plethysmography, PAT: peripheral arterial tone, PO: pulse oximetry, ACT: actigraphy, BCG: ballistocardiography.
[b] Long *et al* (2014).
[c] 60 s epochs; 12% of all epochs excluded from validation.
Note: all data sets comprise healthy subjects unless indicated. There are other studies in literature which present subject-dependent classification results. In this comparison only results obtained with subject-independent schemes were considered.

produced in the present work, indicating that combining cardiac and respiratory activity can lead to an improved classification performance. The study of Hedner *et al* (2011) achieved similar results but they used more signal modalities including peripheral arterial tone, actigraphy, and pulse oximetry. The recent study by Willemen *et al* (2014) also achieved a good performance, although it was validated with a younger sample population, excluded 12% of the epochs from validation and used a basis of 60 s epochs instead of the standard scoring basis of 30 s which makes the results incomparable.

For WRN classification we see that, to the best of our knowledge, our results also outperform those reported in almost all of the previous studies. In comparison with one of the best performing studies (Domingues *et al* 2014), we obtain a higher accuracy (albeit a slightly smaller kappa) but require one less modality (actigraphy). Regarding the work of Willemen *et al* (2014) it is again important to note that the results in that study were obtained on basis of 60 s epochs.

These results suggest that our choice for a Bayesian linear discriminant was appropriate for this task. Besides its simplicity, it offers the benefit of a probabilistic framework which allows, for instance, the direct use of time-varying prior probabilities to improve classification. In comparison with increasingly popular black-box approaches, this classifier has the additional advantage that it does not require the tuning of critical parameters such as kernels for support vector machines, number of nodes in classification trees or number of hidden layers in neural networks.

4. Conclusions

This paper presents a method to identify overnight sleep stages using cardiorespiratory features extracted from ECG and RIP signals. These features were post-processed by means of subject-specific Z-score normalization and spline smoothing, which helps reduce the influence of signal noise, between-subject, or within-subject variability in autonomic physiology. Eighty features were selected from a set of 142 features using a modified SFS-based feature selector designed to avoid biasing the validation performance. Using a linear discriminant classifier in a ten-fold cross-validation procedure, the classification results (for both the four-class WRLD and three-class WRN classification tasks) achieved in this work (table 4) outperform most of the previous studies.

As future work it would be interesting to investigate whether this methodology would achieve a comparable performance in subjects with sleep disorders such as sleep apnea or insomnia. If successful, and given its potential for unobtrusive, prolonged use at home, it could represent a significant step towards lowering the costs and complexity of home tests and thus complement the current medical practices for diagnosis of sleep disorders.

References

Adnane M, Jiang Z and Yan Z 2012 Sleep-wake stages classification and sleep efficiency estimation using single-lead electrocardiogram *Expert Syst. Appl.* **39** 1401–13

Basner M, Griefahn B, Müller U, Plath G and Samel A 2007 An ECG-based algorithm for the automatic identification of autonomic activations associated with cortical arousal *Sleep* **30** 1349–61 (PMID: 17969469)

Benington J H and Heller H C 1995 Restoration of brain energy metabolism as the function of sleep *Prog. Neurobiol.* **45** 347–60

Berger R J and Phillips N H 1995 Energy conservation and sleep *Behavioural Brain Res.* **69** 65–73

Berndt D and Clifford J 1994 Using dynamic time warping to find patterns in time series *Workshop on Knowledge Discovery in Databases* **vol 10** pp 359–70

Bušek P, Vaňková J, Opavský J, Salinger J and Nevšímalová S 2005 Spectral analysis of the heart rate variability in sleep *Physiol. Res.* **54** 369–76 (PMID: 15588154)

Buysse D J, Reynolds C F, Monk T H, Berman S R and Kupfer D J 1989 The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research *Psychiatry Res.* **28** 193–213

Carskadon M A and Dement W C 2011 *Principles and Practice of Sleep Medicine* 5th edn, ed M H Kryger *et al* (Amsterdam: Elsevier) pp 16–26 chapter 2

Cohen J 1960 A coefficient of agreement for nominal scales *Educ. Psychological Meas.* **20** 37–46

Costa M, Goldberger A L and Peng C K 2005 Multiscale entropy analysis of biological signals *Phys. Rev.* E **71** 021906

Cysarz D, Bettermann H, Lange S, Geue D and van Leeuwen P 2004 A quantitative comparison of different methods to detect cardiorespiratory coordination during night-time sleep *Biomed. Eng. Online* **3** 44

Cysarz D, Bettermann H and van Leeuwen P 2000 Entropies of short binary sequences in heart period dynamics *Am. J. Physiol. Heart Circ. Physiol.* **278** 2163–72 (PMID: 10843917)

De Boor C 2001 *A Practical Guide to Splines* (Berlin: Springer)

de Chazal P, Fox N, O'Hare E, Heneghan C, Zaffaroni A, Boyle P, Smith S, O'Connell C and McNicholas W T 2011 Sleep/wake measurement using a non-contact biomotion sensor *J. Sleep Res.* **20** 356–66

Domingues A, Paiva T and Sanches J M 2014 Hypnogram and sleep parameter computation from activity and cardiovascular data *IEEE Trans. Biomed. Eng.* **61** 1711–9

Duda R O, Hart P E and Stork D G 2000 *Pattern Classification* 2nd edn (New York: Wiley)

Fonseca P, Aarts R M, Foussier J and Long X 2014 A novel low-complexity post-processing algorithm for precise QRS localization *SpringerPlus* **3** 376

Hamilton P S and Tompkins W J 1986 Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database *IEEE Trans. Biomed. Eng.* **33** 1157–65

Hamilton P 2002 Open source ECG analysis *Proc. IEEE Computers in Cardiology (22–25 September 2002)* pp 101–4

Hedner J, White D P, Malhotra A, Herscovici S, Pittman S D, Zou D, Grote L and Pillar G 2011 Sleep staging based on autonomic signals: a multi-center validation study *J. Clin. Sleep Med.* **7** 301–6

Ichimaru Y, Clark K P, Ringler J and Weiss W J 1990 Effect of sleep stage on the relationship between respiration and heart rate variability *Proc. Computers in Cardiology (22–26 September 1990)* pp 657–60

Isa S M, Wasito I and Arymurthy A M 2011 Kernel dimensionality reduction on sleep stage classification using ECG signal *Int. J. Comput. Sci. Issues* **8** 1178–81

Iyengar N, Peng C K, Morin R, Goldberger A L and Lipsitz L A 1996 Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics *Am. J. Physiol.* **271** R1078–84 (PMID: 8898003)

Kantelhardt J W, Koscielny-bunde E, Rego H H A, Havlin S and Bunde A 2001 Detecting long-range correlations with detrended fluctuation analysis *Physica* **295** 441–54

Klosch G *et al* 2001 The SIESTA project polygraphic and clinical database *IEEE Eng. Med. Biol. Mag.* **20** 51–7

Kortelainen J M, Mendez M O, Bianchi A M, Matteucci M and Cerutti S 2010 Sleep staging based on signals acquired through bed sensor *IEEE Trans. Inf. Technol. Biomed.* **14** 776–85

Kurihara Y and Watanabe K 2012 Sleep-stage decision algorithm by using heartbeat and body-movement signals *IEEE Trans. Syst. Man and Cybern.* **42** 1450–9

Lacasa L, Luque B, Ballesteros F, Luque J and Nuño J C 2008 From time series to complex networks: the visibility graph *Proc. Natl Acad. Sci.* **105** 4972–5

Long X, Fonseca P, Aarts R M, Haakma R and Foussier J 2014a Modeling cardiorespiratory interaction during human sleep with complex networks *Appl. Phys. Lett.* **105** 203701

Long X, Fonseca P, Foussier J, Haakma R and Aarts R M 2014b Sleep and wake classification with actigraphy and respiratory effort using dynamic warping *IEEE J. Biomed. Health Inform.* **18** 1272–84

Long X, Fonseca P, Haakma R, Aarts R M and Foussier J 2014c Spectral boundary adaptation on heart rate variability for sleep and wake classification *Int. J. Artif. Intell. Tools* **23** 1460002

Long X, Foussier J, Fonseca P, Haakma R and Aarts R M 2014d Analyzing respiratory effort amplitude for automated sleep stage classification *Biomed. Signal Proc. Control* **14** 197–205

Long X, Yang J, Weysen T, Haakma R, Foussier J, Fonseca P and Aarts R M 2014 Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging *Physiol. Meas.* **35** 2529–42

Malik M, Bigger J T, Camm A J, Kleiger R E, Malliani A, Moss A J and Schwartz P J 1996 Heart rate variability: standards of measurement, physiologic interpretation, and clinical use *Eur. Heart J.* **17** 354–81

Mendez M O, Matteucci M, Castronovo V, Ferini-Strambi L, Cerutti S and Bianchi A M 2010 Sleep staging from heart rate variability: time-varying spectral features and hidden Markov models *Int. J. Biomed. Eng. Technol.* **3** 246–63

Migliorini M, Bianchi A M, Nisticò D, Kortelainen J, Arce-Santana E, Cerutti S and Mendez M O 2010 Automatic sleep staging based on ballistocardiographic signals recorded through bed sensors *Proc. of the Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 3273–6

Penzel T, Kantelhardt J W, Grote L, Peter J H H and Bunde A 2003 Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea *IEEE Trans. Biomed. Eng.* **50** 1143–51

Pudil P, Novovičová J and Kittler J 1994 Floating search methods in feature selection *Pattern Recognit. Lett.* **15** 1119–25

Rechtschaffen A and Kales A 1968 *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects* (Bethesda, Md: US Public Health Service)

Redmond S J, de Chazal P, O'Brien C, Ryan S, McNicholas W T and Heneghan C 2007 Sleep staging using cardiorespiratory signals *Somnologie-Schlafforschung Schlafmedizin* **11** 245–56

Redmond S J and Heneghan C 2006 Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea *IEEE Trans. Biomed. Eng.* **53** 485–96

Richman J S and Moorman J R 2000 Physiological time-series analysis using approximate entropy and sample entropy *Am. J. Physiol. Heart Circ. Physiol.* **278** H2039–49 (PMID: 10843903)

Shao Z G 2010 Network analysis of human heartbeat dynamics *Appl. Phys. Lett.* **96** 073703
Smialowski P, Frishman D and Kramer S 2010 Pitfalls of supervised feature selection *Bioinformatics* **26** 440–3
Stickgold R 2005 Sleep-dependent memory consolidation *Nature* **437** 1272–8
Telser S, Staudacher M, Ploner Y, Amann A, Hinterhuber H and Ritsch-Marte M 2004 Can one detect sleep stage transitions for on-line sleep scoring by monitoring the heart rate variability? *Somnologie* **8** 33–41
Unser M 1999 Splines: a perfect fit for signal and image processing *IEEE Signal Proc. Mag.* **16** 22–38
Walker M P 2009 The role of slow wave sleep in memory processing *J. Clin. Sleep Med.* **5** S20–6
Watanabe T and Watanabe K 2004 Noncontact method for sleep stage estimation *IEEE Trans. Biomed. Eng.* **51** 1735–48
Whitney A W 1971 A direct method of nonparametric measurement selection *IEEE Trans. Comput.* **100** 1100–3
Willemen T, Van Deun D, Verhaert V, Vandekerckhove M, Exadaktylos V, Verbraecken J, Huffel S V, Haex B and Vander Sloten J 2014 An evaluation of cardio-respiratory and movement features with respect to sleep stage classification *IEEE J. Biomed. Health Inf.* **18** 661–9
Xiao M, Yan H, Song J, Yang Y and Yang X 2013 Sleep stages classification based on heart rate variability and random forest *Biomed. Signal Proc. Control* **8** 624–33
Yentes J M, Hunt N, Schmid K K, Kaipust J P, McGrath D and Stergiou N 2013 The appropriate use of approximate entropy and sample entropy with short data sets *Ann. Biomed. Eng.* **41** 349–65
Yılmaz B, Asyalı M H, Arıkan E, Yetkin S and Özgen F 2010 Sleep stage and obstructive apneaic epoch classification using single-lead ECG *Biomed. Eng. Online* **9** 39
Zhu G, Li Y and Wen P P 2014 Analysis and classification of sleep stages based on difference visibility graphs from a single channel EEG signal *IEEE J. Biomed. Health Inf.* **18** 1813–21