Preprint

Citation: Karlen, W, Mattiussi, C and Floreano, D (2009) Sleep and Wake Classification With ECG and Respiratory Effort Signals *IEEE Transactions on Biomedical Circuits and Systems*, To appear

The definitive version of this article will be available at: http://ieeexplore.ieee.org

Sleep and Wake Classification With ECG and Respiratory Effort Signals

Walter Karlen*, Claudio Mattiussi and Dario Floreano

Laboratory of Intelligent Systems, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

We describe a method for the online classification of sleep/wake states based on cardiorespiratory signals produced by wearable sensors. The method was conceived in view of its applicability to a wearable sleepiness monitoring device. The method uses a Fast Fourier Transform as the main feature extraction tool and a feed-forward Artificial Neural Network as a classifier. We show that when the method is applied to data collected from a single young male adult, the system can correctly classify on average 95.4% of unseen data from the same user. When the method is applied to classify data from multiple users with the same age and gender, its accuracy is reduced to 85.3%. However, a Receiver Operating Characteristic analysis shows that, compared to actigraphy, the proposed method produces a more balanced correct classification of sleep and wake periods. Additionally, by adjusting the classification threshold of the neural classifier, 86.7% of correct classification is obtained.

biomedical signal analysis | wearable computing | sleep and wake classification | electrocardiography | respiratory effort | neural classifier

Introduction

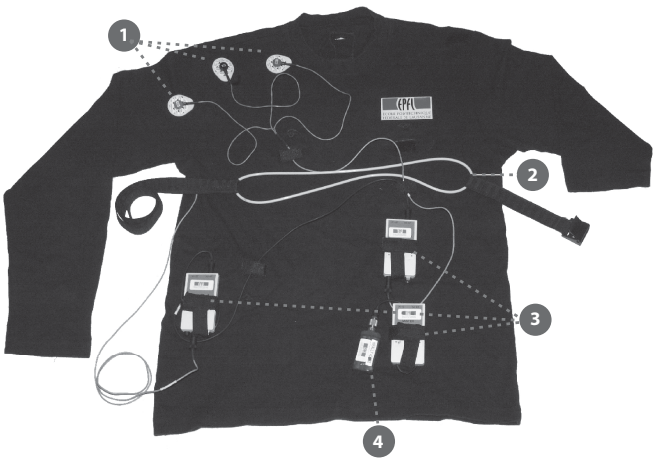
Increased sleepiness over daytime has been identified as an important cause of accidents in transportation and factory plants (1). It is therefore a major health interest to continuously monitor and report the sleepiness level of high risk persons such as pilots, truck drivers or shift workers. Continuously updated information about the persons' "need for sleep" could help these persons to better schedule their breaks and sleep times. Currently, transport industries focus mainly on emergency situation prevention by means of vehicle centered systems that alert the user either by monitoring the vehicle performances (e.g. lane deviation) or operators' behavioral responses (e.g. eye blinks). Other fatigue detection techniques include fitness-for-duty tests and mathematical alertness models (2; 3). Our approach consists in using mathematical models in combination with physiological measurements to establish a continuous sleepiness profile of the subject and give warnings even before a certain task begins or an emergency situation related to fatigue can occur.

Different mathematical models to estimate sleepiness have been suggested (4). These models are mainly based on the previous sleep and wake durations (homeostatic process), and daily alertness rhythms (circadian process). In this paper we describe a method for the estimation of the homeostatic component with a wearable device. In order to be wearable, the sleep/wake detection device should be energetically autonomous. Since the person is expected to wear it for several days, the device should also be light-weight and comfortable. This puts tight restrictions not only on the choice of signals for the classification task, but also on the signal recording, processing, and on the computational requirements of the classifier. In addition, such a device is intended for a large public. Therefore, it should be easy to use and should not depend on complicated calibration methods.

The gold standard for assessing sleep in humans is the analysis of brain wave patterns (EEG) first described by Rechtschaffen and Kales (5). The most common sleep analysis method is called polysomnography (PSG), which combines EEG recordings with different physiological signals like electromyography (EMG), electrooculography (EOG), respiratory effort, blood oxygen saturation, electrocardiograms (ECG) and video analysis. In PSG, 30-second epochs of the signals are used for decision making. The method is normally carried out in a controlled hospital environment and needs medical assistance for setting up sensors, monitoring and analysis. Although the analysis is typically computer-assisted (6), it still requires a sleep expert and is therefore expensive and time consuming. It is difficult to integrate PSG sensors into a wearable device, as they are rather bulky, power-consuming and highly susceptible to noise. Furthermore, EEG recordings require many electrodes to be glued to the scalp, which makes it very cumbersome and uncomfortable for the user.

In home environments, where PSG is typically not available, physicians rely on actigraphy for sleep monitoring (7). In this method, the acceleration of the extremities (typically wrist) are recorded over several days with a watch-like device using miniature accelerometers and a storage medium. Periods of low activity are later classified as sleep by offline computer processing. Many different classification algorithms

*To whom correspondence may be addressed. E-mail: walter.karlen@epfl.ch



Healthy recording system mounted on a shirt. 1) ECG gel electrodes; 2) inductive belt sensor; 3) electronics modules; 4) NiMH battery. The EMG and EOG electrodes are not shown.

have been suggested for actigraphy (8; 9), but often they cannot cope with the problem of misclassifying low activity tasks like reading and watching television or the case where the sensor band is not worn (7; 9). Recently, alarm clocks using accelerometers have been commercialized (10; 11). The activity is used to detect the best sleep phase for easy wake-up in a given time window (10 to 30 minutes). However, the accelerometers are only active at night and the clocks do not calculate sleep duration.

Changes in the activity of the Autonomic Nervous System (ANS) during sleep/wake transitions have been successfully identified as a reliable source of information (12). Changes in activity of the ANS are reflected in various physiological signals such as heart rate, blood pressure, skin conductance, etc. The main focus of current research is on fluctuations of heart rate variability (HRV) during sleep (13; 14; 15). However, due to differences in the methods used to calculate HRV, the results are sometimes contradictory (16). Moreover, HRV measures are very susceptible to noise. A wearable application of this technique is therefore difficult.

Recently, Redmond and Heneghan (17) have added respiratory signals to the HRV to show the feasibility of using cardiorespiratory signals for discriminating sleep stages in subjects with obstructive sleep apnea. The advantage of cardiorespiratory signals is that they are relatively easy to measure and the sensors can be applied by non-experienced users.

For this reason, we decided to use cardiorespiratory signals together with an Artificial Neural Network (ANN) in our sleep detection system. Cardiorespiratory signals recorded from wearable sensors typically contain artifacts due to movements of the subject wearing the sensors. Instead of filtering out these artifacts with sophisticated signal reconstruction and artifact-rejection algorithms, our method treats the artifacts as relevant information in the signal. Within this perspective, movement artifacts can give an indication of the activity of the user as actigraphy would do, but without the need of using an additional sensor. Contrary to all other studies, we rely on day and night recordings obtained in a non-hospital environment to obtain more realistic data. We have selected ANN classifiers because of their capabilities for nonlinear class separation and the possibility to efficiently program them into a microcontroller.

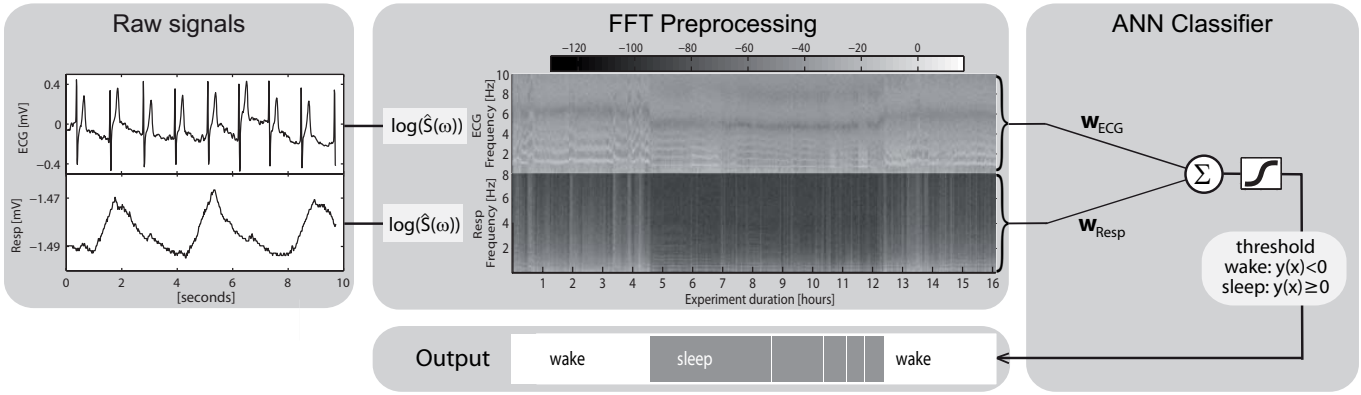
Method

The raw data used in the current study consist of ECG and respiratory effort signals of different subjects recorded over day and night periods. Additionally, video, EMG and EOG were recorded for labeling the users' state as wake or sleep by a technician. This information was used for training the neural classifier. The preprocessing consisted in calculating an estimation of the power spectral density (PSD) for the raw ECG and respiratory signals with the help of a Fast Fourier Transform (FFT). Three classifier architectures were designed, each having as input either the PSD values of ECG, respiration or their combination (Fig. 2). For each architecture, an ANN was first trained and tested using data from a single user. To investigate its capability to generalize to multiple users, each architecture was then trained and tested using data from multiple users.

Data Recordings. We conducted home recordings with 6 healthy male subjects, aged between 23 and 29 years. ECG and respiratory effort were recorded with a Healthy system (Fig. 1, Koralewski Industrie Elektronik, Celle, Germany). The Healthy system is a portable recording system that uses an inductive belt sensor for measuring ribcage respiratory effort and gel electrodes for measuring ECG. We have chosen the sampling frequencies f according to the requirements for digitalized PSG (6). The respiratory signal is sampled at $f_{Resp} = 50$ Hz and the 1-lead ECG at $f_{ECG} = 100$ Hz. Additionally, the Healthy system offers the possibility to measure the EMG (recorded from the right shoulder muscle *trapezius* at 200 Hz) and EOG (recorded at 200 Hz) as reference. EOG was only measured during the night, in order not to disturb the subjects too much during daily activities. During nighttime a video of the upper part of the body was recorded. We did not consider the possibility of recording EEG signals, because subjects wearing the monitoring device were expected to move freely and perform undisturbed daily activities.

In order to obtain an equal amount of data for both sleep and wake, subjects wore the recording system for 16 hours per session. This recording time corresponds to the double of the average sleep time for the studied age group (18). The recording started approximatively 4 hours before the regular bed time of the subject. A total of 18 recording sessions were carried out, 8 sessions for one subject (subject A) and 2 sessions for each of the other subjects (subject B, C, D, E and F). Each session contained a mean of 7.19 $\pm$ 1.65 hours of sleep and 8.45 $\pm$ 3.19 hours of wake. A total of 130.47 hours of sleep and 161.64 hours of wake were analyzed. In case of sensor failure or detachment, the corresponding data segments were discarded. The subjects reported no major discomfort during sleep because of the recording system. However, at the end of the recording, itching at the electrode sites was reported.

The videos were analyzed by a human expert to determine if the subject was asleep or not. We did not distinguish between light, deep or REM sleep, because the binary discrimination between sleep and wake was sufficient for the present study. The video was divided into segments of 10 seconds and



Overview of the sleep/wake classification system. Raw ECG and respiratory effort signals are mapped to the frequency space with the help of a Fast Fourier Transform (FFT). The resulting frequency data (represented here by a spectrogram) are fed to a feed-forward, single-layer Artificial Neural Network (ANN) with a tangent-sigmoid transfer function and a symmetric classification threshold.

each segment was labeled using the following criteria derived from (8; 12; 19):

1. The person is considered to be awake if his eyes are open or body movements occur for more than 10 seconds.

2. If the eyes are closed, the subject is considered to be asleep when muscle tonus is released or slow eye movements are present. If segments of the video analysis were uncertain, the EOG and EMG signals were examined.

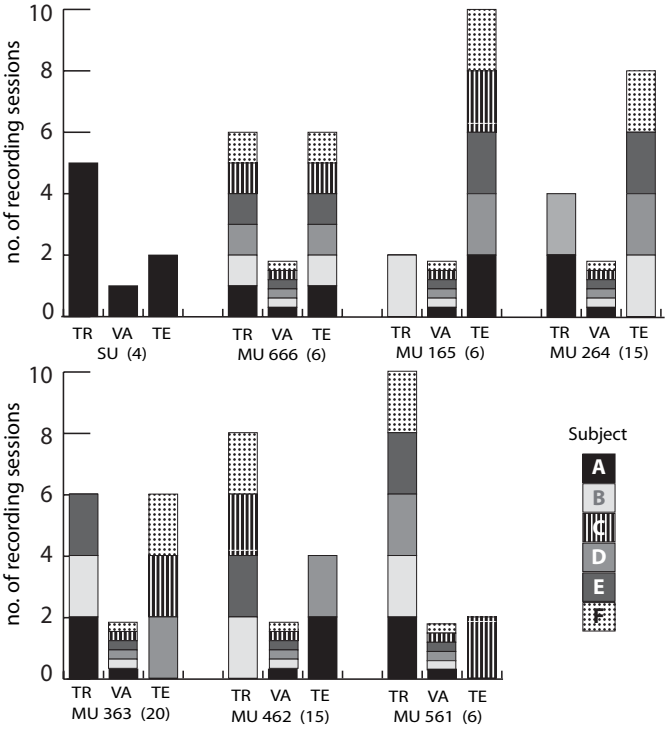
3. In doubtful cases, where neither EOG nor EMG signals could help to clearly identify sleep, the state was set to awake.

To prevent undetected wake with closed eyes, the subjects were asked to open the eyes if they woke up during the night.

Preprocessing and Feature Extraction. The feature extraction step consisted in calculating the power spectral density (PSD) of raw ECG and respiratory signals. PSD estimation methods are widely used for this purpose in biomedical signal analysis (20). The periodogram method that we chose partitioned the original sequence of samples into equally sized segments s. Each segment s was windowed with a Hamming window w to reduce the effects of spectral leakage at the first side lobe (21). For each segment a periodogram $\hat{\mathbf{S}}$ was then calculated with an FFT, as follows

$$\hat{\mathbf{S}} = |FFT\{s(m)w(m)\}|^2 \qquad m = 0, 1, \ldots, N-1 \qquad (1)$$

where N is the number of samples that form a segment s. The periodogram $\hat{\mathbf{S}}$ gives an estimation of the frequency content in the given time segment. Computationally efficient FFT calculation in a digital signal processing (DSP) microcontroller requires a number of samples $N = 2^l$ where l is a positive integer. The values of $N_{ECG} = 4096$ and $N_{Resp} = 2048$ were selected according to this constraint and correspond to a segment length of 40.96 seconds. Note that this is the next larger possible segment size compared to the traditional segment length of 30 seconds of PSG. In previous unpublished work we analyzed the effects of increasing the segment size for a better PSD resolution. Since the performance of the classifier did not improve, there appears to be no reason to incur



Experimental design for training the neural classifier. SU = trained and tested on single user. MU = trained and tested on multiple users. Numbers indicate users in the training set (TR), users in the validation set (VA), users in the test set (TE) (no. of repetitions with different combinations of users/sessions in training and testing).

the additional computational cost entailed by the increased segment size. As the input values are real numbers, the PSD output is symmetrical around the DC component and it is sufficient to use half of the output points ($N/2$). The DC component was eliminated, because it contained mainly the offset of the uncalibrated sensors. Further experiments (not

[1] The parameters were: μ: 0.001; μ increase: 10; μ decrease: 0.1; μ max: 10^{10}; min gradient: 10^{-10}; max validation failures: 15.

detailed in this paper) showed that the high frequency components of the ECG and respiratory spectrogram can be pruned without degrading the performance of the classifier. We found experimentally that the ECG spectrogram can be truncated at 10 Hz and the respiratory spectrogram can be truncated at 8 Hz. Correspondingly, we reduced the input size of the ANN from $N_{ECG}/2$ and $N_{Resp}/2$ to $n_{ECG} = 409$ and $n_{Resp} = 327$, respectively.

Neural Classifier. We used a feed-forward ANN with no hidden layers and one single output unit with a tangent-sigmoid transfer function (see Fig. 2, ANN classifier). We also experimented with an ANN with one hidden layer, but the performance was not better and the training time increased considerably. To train the ANN and update the synaptic weights we used the Levenberg-Marquardt backpropagation algorithm (22)[1]. We studied three different architectures, which differed in the type of input signal. The input vector of the first architecture *ECG+Resp* was composed of the logarithm of the periodograms $\hat{\mathbf{S}}_{ECG}$ and $\hat{\mathbf{S}}_{Resp}$ (Fig. 2). The other two architectures *ECG* and *Resp* used only the logarithm of the periodogram of one of the two signals, ECG or respiratory effort, respectively.

Initialization of the weights was done with the Nguyen-Widrow method (23). The output of the neuron was thresholded so that $y(x) \geq 0$ is mapped to sleep and $y(x) < 0$ is mapped to wake. To train the networks, the data were divided into three sets: training, validation and test. The training set (TR) contained the data used to update the synaptic weights. The performance of the network was evaluated on the validation set (VA) after each iteration and the training was stopped if the performance of VA did not increase for more than 15 training iterations or the minimal gradient was reached. The test set (TE) was used to measure the performance of the network after the training.

Experiments.

Single-User Experiments

With this set of experiments, we investigated the performance of the method when trained and tested on the same person. We used subject A, for whom we had the highest number of recording sessions. The 8 available sessions were randomly divided into TR containing 5 sessions, VA containing 1 session, and TE containing 2 sessions. 5 independent runs were performed from different initial weight values. In order to prevent performance biases due to the choice of sessions used for training and testing, we repeated the experiment 4 times with different sessions in the training and testing set.

Multi-User Experiments

Most algorithms for sleep/wake detection are based on data from a multitude of users and are expected to generalize to other users (24; 25; 9). We investigated the performance of our method when trained on a single person and tested on multiple persons, and when trained on multiple persons and tested on multiple persons. Six experiments were carried out, each with an increasing number of persons in the training set (1 to 6) and all remaining persons in the testing set (in the only case when all 6 persons were in the training set, we made sure that the two sets contained different recording sessions). The validation set was composed of 2 hours of data from each user, randomly sampled over the available 2 sessions and containing an equal amount of sleep an wake labels. This data was neither used for training nor for testing. Five independent runs of each experiment were performed from different initial weight values.

In order to prevent performance biases due to the choice of sessions, we repeated each experiment with all possible combinations of sessions in the testing and training set, making sure that the same session did not appear both in the training and in the testing set (the number of repetitions for each experiment is indicated between brackets in Fig. 2).

Results and Discussion

We determined the accuracy of our algorithm by calculating the percentage of true (correct) classifications of sleep and wake of TE according to Eq. 2)

$$accuracy = \frac{true\ sleep + true\ wake}{all\ sleep + all\ wake} \tag{2}$$

To further quantify the performance of the system we computed two additional quantities: the number of segments classified as sleep per session *total sleep time*, which is an important parameter in the sleepiness estimation model described in the introduction, and the *awakenings* (total number of sleep-to-wake transitions during the period between the first and last segment labeled as sleep in a session), which is an indicator of sleep quality of the subject.

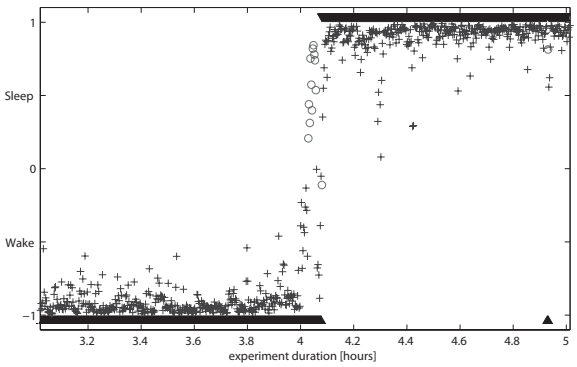
The comparison of the total sleep time labeled according to the video data and estimated by the classifier does not take into account whether the estimated sleep epochs are correctly classified. we plotted a series of Receiver Operating Characteristic (ROC) curves (Fig. 5). ROC curves allow the assessment of the results of classifier data in which the classes are not equally distributed (which is the case for our data). For this reason it is often used in medical decision making and has been introduced in sleep analysis comparisons by Tryon (26). An ROC curve shows the fraction of correctly classified sleep points called *sensitivity* (Eq. 3) vs. the fraction of wrongly classified wake points (1-*specificity*, Eq. 4), when the classification threshold of the ANN output is varied from -1 to 1.

$$sensitivity = \frac{true\ sleep}{true\ sleep + false\ sleep} \tag{3}$$

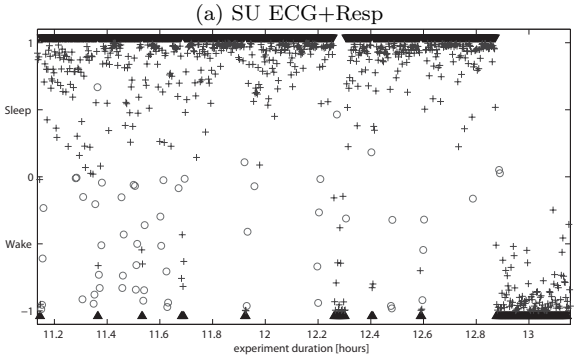
$$specificity = \frac{true\ wake}{true\ wake + false\ wake} \tag{4}$$

In such a representation the output of an ideal classifier is located at (0,1), which corresponds to perfect classification of all wake and sleep states.

Single User. In the single-user condition (SU), the classification accuracy of the networks using both ECG and respiration signals was much better than that of the networks using only ECG signals (95.4% vs. 91.7%, $p<0.001$, t-test) and slightly better than the networks using only respiration signals (95.4% vs. 93.3%, $p<0.01$, t-test) (Table 2). Fig. 2a shows a 2 hour detailed view of an output of a single-user classifier using respiration and ECG signals as input. In this example we can observe that the process of falling asleep is gradually detected by the classifier, although the labeling is binary (black line composed of triangles). From a physiological point of view



(a) SU ECG+Resp



(b) SU Resp

Unthresholded output of a single-user classifier on two different sessions with a) ECG and respiratory effort and b) respiratory effort only. The crosses represent correctly classified segments when a classifier threshold of 0 is applied, the circles represent the wrong classifications. The black triangles are the labeling values from the technician.

Single-user (SU) test classification accuracies and standard deviations in percent

Experiment	ECG	Resp	ECG + Resp
SU	91.67 ± 2.74	93.27 ± 1.48	95.42 ± 1.61

Comparison of SU sleep parameters

Sleep parameters[a]	Mean ± SD	Mean difference from Video ± SD
1. Total sleep time (hours)		
Video (label)	7.55 ± 1.7	
SU ECG	7.79 ± 2.11	-0.25 ± 1.22
SU Resp	7.46 ± 1.54	0.09 ± 0.3
SU ECG+Resp	7.77 ± 1.74	-0.22 ± 0.44
2. Awakenings (numbers)		
Video (label)	23.75 ± 8.81	
SU ECG	88.71 ± 86.96	-64.96 ± 85.24
SU Resp	92.61 ± 39.67	-68.86 ± 36.52
SU ECG+Resp	49.6 ± 24.37	-25.85 ± 20.61

[a] The sleep parameters are calculated on each session individually and not over the entire test set that contains several sessions.

this makes sense because sleep onset is a gradual, rather than a discrete process. A more detailed probabilistic classification taking into account the uncertainty of the classification would presumably reveal that the uncertainty increases in this transition phase, reducing the reliability of the classification. The same figure also shows that the first awakening of the subject

Tuned multi-user (MU) test classification accuracies and standard deviations in percent

Experiments	ECG	Resp	ECG + Resp
MU 165	67.30 ± 6.71	84.06 ± 3.47	77.75 ± 5.53
MU 264	68.28 ± 5.06	86.64 ± 2.39	82.45 ± 3.68
MU 363	69.93 ± 5.87	87.75 ± 2.44	84.32 ± 3.13
MU 462	71.32 ± 8.30	88.68 ± 2.88	85.76 ± 3.37
MU 561	75.18 ± 11.29	89.52 ± 3.79	86.68 ± 5.33
MU 666	78.98 ± 1.78	90.24 ± 1.29	89.04 ± 2.25

after 48 minutes could not be correctly classified. This difficulty to detect short awakenings inside long sleep epochs has also been reported by various actigraphy studies (24; 25; 9).

In comparison, the only other study in the literature where both ECG and respiration signals from a single subject were combined, reported an accuracy of 81% (17). However, a direct comparison of the two results is difficult, for the following

Multi-user (MU) test classification accuracies and standard deviations in percent

Experiments	ECG	Resp	ECG + Resp
MU 165	65.54 ± 7.76	83.09 ± 2.78	76.53 ± 5.39
MU 264	66.74 ± 5.88	85.69 ± 2.05	81.65 ± 3.82
MU 363	68.17 ± 5.98	86.64 ± 2.33	83.59 ± 3.21
MU 462	68.63 ± 8.99	87.34 ± 2.91	84.96 ± 3.52
MU 561	69.51 ± 14.96	87.31 ± 4.68	85.25 ± 5.25
MU 666	78.72 ± 1.66	89.77 ± 1.37	88.34 ± 2.30

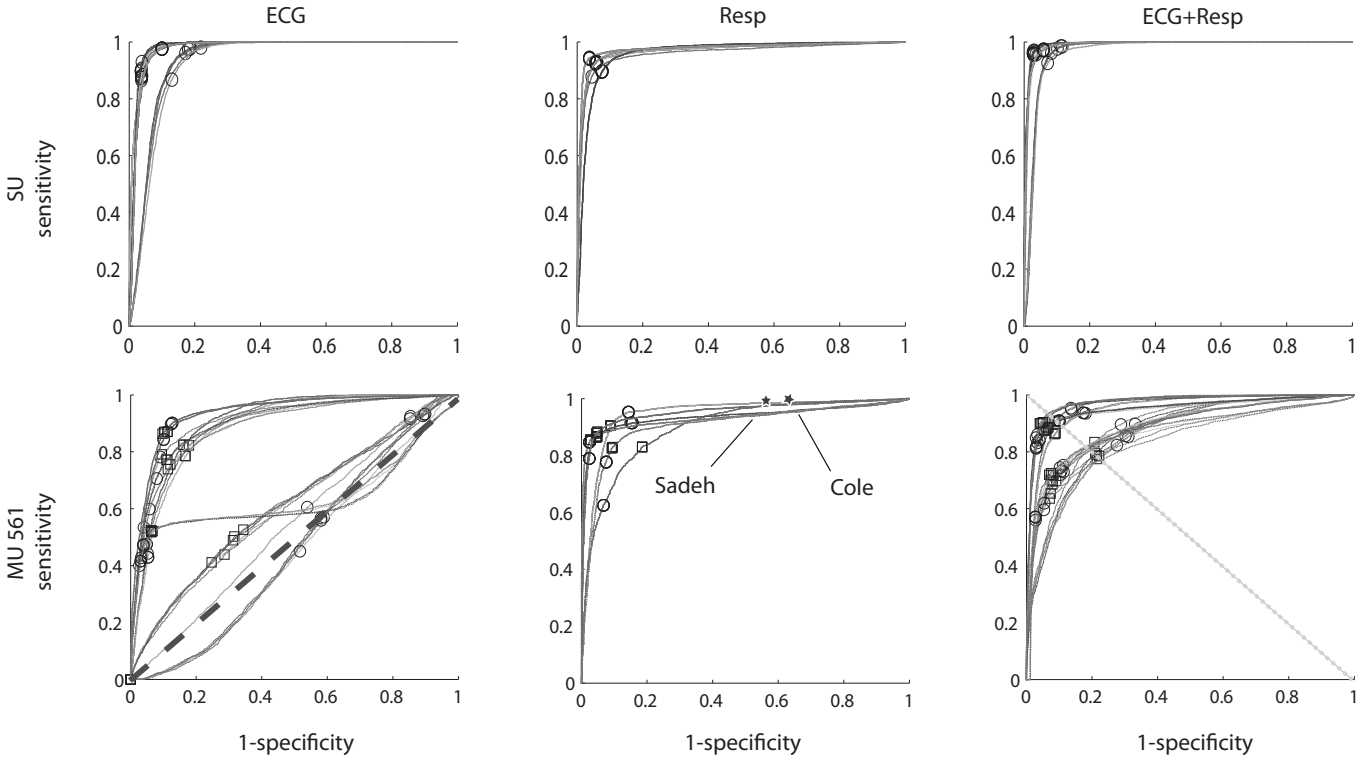
Multi-user (MU) mean training accuracies and standard deviations in percent

Experiments	ECG	Resp	ECG + Resp
MU 165	96.63 ± 1.81	95.81 ± 1.45	99.53 ± 0.34
MU 264	92.62 ± 1.99	93.83 ± 1.25	97.62 ± 1.12
MU 363	90.31 ± 2.04	93.06 ± 0.84	96.02 ± 0.89
MU 462	88.46 ± 1.98	92.64 ± 0.60	95.05 ± 0.67
MU 561	86.92 ± 1.63	92.34 ± 0.34	94.63 ± 0.56
MU 666	88.35 ± 3.31	92.72 ± 1.07	95.84 ± 1.02

Comparison of MU 561 sleep parameters

Sleep parameters[a]	Mean ± SD	Mean difference from Video ± SD
1. Total sleep time (hours)		
Video (label)	4.27 ± 1.14	
MU 561 ECG	4.44 ± 3.07	-0.17 ± 2.6
MU 561 Resp	3.88 ± 1.12	0.38 ± 0.9
MU 561 ECG+Resp	4.17 ± 1.6	0.09 ± 0.9
2. Awakenings (numbers)		
Video (label)	31.38 ± 13.27	
MU 561 ECG	118.5 ± 68.7	-86.79 ± 67.52
MU 561 Resp	106.08 ± 50.99	-74.7 ± 50.49
MU 561 ECG+Resp	98.82 ± 52.59	-67.82 ± 49.84

[a] The sleep parameters are calculated on each session individually and not over the entire test set that contains several sessions.



Receiver Operating Characteristics (ROC) curves of the three neural classifiers for the SU (top) and the MU 561 (bottom) configuration. Each curve corresponds to a different run obtained by changing the training set configuration or the initial network weights. The circles are the ROC points of the actual classification of the test set with the same classification threshold as in the training. The squares are the ROC points with the highest accuracy using the same weights as obtained in the training, but with a tuned classification threshold. Bottom left: The dashed line corresponds to the ROC line of a random classifier. Bottom middle: The stars correspond to the ROC points obtained by the two actigraphy algorithms proposed by Cole (8) and Sadeh (27) that were used in (9). Bottom right: The dotted line corresponds to ROC points with perfectly balanced sensitivity and specificity.

reasons:

a) the value reported in (17) was measured in the more difficult task of classifying wake, sleep, and REM sleep;

b) it was obtained with data recorded in a controlled hospital environment using PSG equipment;

c) it used only data from night recordings; and

d) it used a computationally expensive pre-processing algorithm calculating 27 features, which may be difficult to implement in a low-power and wearable device.

Table 2.1 shows that on average the ECG and the ECG and respiration classifiers overestimate the total sleep time, whereas the classifier using only respiratory data slightly underestimates the total sleep time. Similar findings were reported in actigraphy studies (25; 9). A possible explanation of sleep overestimation when ECG signal is used as an input is that the ECG signal might vary substantially between different sleep stages. This might be the case for REM-sleep for example, but further analysis with a more detailed labeling of the sleep data would be needed to confirm this statement.

Table 2.2 shows that the mean number of awakenings is overestimated by all classifiers. This can be explained by a relatively noisy output of the classifier around the classification threshold, where consecutive epochs are classified alternatively as sleep and wake and therefore counted as several awakenings. An example of this phenomenon can be observed in Fig. 2b. The occasional mis-labeling of the video data might also contribute to this overestimation. Note that the combination of ECG and respiration performs relatively better in terms of the estimate of the number of awakenings.

The circles in the top of the ROC graphs for the single-user experiments shown in the top of Fig. 5 correspond to the output of the test set with a classification threshold set to 0 (the value used in the training). We observe equally balanced sensitivity and specificity in most cases, which means that the classifier classifies sleep as accurately as wake.

Multiple Users. In all six multi-user conditions (MU), the accuracies in the test condition dropped compared to the single-user conditions. The accuracy drop was the largest for the networks using only ECG signals (Table 2, first column). Table 2 shows that, as expected, the training performance of the networks using both ECG and respiratory signals always improves with respect to the training performance of the networks using only one of the two signals. However, the networks using both ECG and respiration signals (Table 2, third column) displayed lower accuracy in the test condition than the networks using only respiration signals (Table 2, second column). Also, the variability of the accuracy across multiple replications was higher when respiration was used in combination with ECG as compared to the condition when respiration alone was used. These results suggest that ECG signals have unique features that are specific to each individual, so that networks trained on the data obtained from a small number

of subjects do not generalize well to other individuals for the purpose of discriminating sleep and wake states. This points to the necessity of increasing the number of subjects contributing to the formation of the training and validation datasets, when using both the ECG and respiration signals, in order to prevent the classifier from overfitting the peculiarities of the training data obtained from a restricted sample of subjects. Although one may notice a positive correlation between accuracy and number of users in the training set, our data sets are not sufficiently large to draw conclusions on the observed differences among the six multi-user conditions.

Table 2 shows that in the MU 561 case, which is the most similar to the condition used in actigraphy studies (25; 9), the ECG and the ECG and respiration classifiers produce a good mean estimate of the total sleep time, whereas all classifiers overestimate the number of awakenings, which has also been observed in the single-user case (Table 2).

The ROC data for the MU 561 case reveals, as previously indicated, that the ECG signal alone does not generalize well and for some cases does not find a solution better than random (Fig. 5, circles close to dashed line, bottom left). Although the accuracy of our multi-user classifier with respiration signal is comparable to the 91% accuracies of the Cole (8) and Sadeh (27) algorithms with only accelerometer signals used by de Souza (9), it produces a smaller fraction of wrongly classified wake points. As we mentioned in the introduction, actigraphy may often mis-classify wake periods of low activity (reading, watching television, lying in bed) as sleep periods. Indeed, the specificity of the actigraphy methods in (9) was only 44% (Sadeh) and 34% (Cole) during the wake periods (stars, Fig. 5, bottom middle), whereas in our case it was 88.12% $\pm$ 9.6 when only the respiration signal was used and 91.87% $\pm$ 5.23 when both ECG and respiratory effort were used. In general, we can say that our solution using the respiration signal as input achieves a better balance between sensitivity and specificity than those obtained using actigraphy. The findings of Kushida et al. (25), who compared actigraphy with PSG data from a large group of sleep-disordered patients, support this statement (accuracy = 78%, sensitivity = 92%, specificity= 48%).

In most multi-user conditions the solution with a fixed classification threshold at 0 (Fig. 5, circles) does not correspond to the highest possible accuracy in the test set for the classifier weights obtained from the training. Adjusting the classification threshold leads to an increase in mean accuracy (Table 2). Further, it produces a even more balanced solution (Fig. 5, squares). This tuning is easy to implement, since it requires only the tuning of a single parameter. A further performance increment can be obtained by updating the weights of the ANN, but this is difficult to implement on a low power microcontroller.

Conclusion

The method and results presented in this paper demonstrate that the combination of ECG and respiratory signals can discriminate with high accuracy between sleep and wake states for individual young male adult users. Our choice of FFT signal preprocessing and ANN classification makes it possible to implement the method in an off-the-shelf, low-power DSP microcontroller. With currently available DSP microcontrollers, the presented calculation can be done with less than 150k instructions (3.75 ms), which fits largely in a 10 ms window between two sampling instructions of ECG and respiration and makes online sleep/wake classification possible. However, the power consumption in this configuration would still be too high for long-term sleepiness estimation. A promising way to increase autonomy could be to implement the preprocessing and the ANN classification on an analog VLSI chip, as done recently by Aziz et al. for epileptic seizure detection (28). However, this might not be possible with more advanced preprocessing and classification techniques than those presented in this paper.

With the current recording system, the users reported some discomfort linked to the loose cables and the ECG electrodes glued on the skin. We plan to increase the acceptance of the system by using another recording system that makes use of dry electrodes instead of the gel electrodes available for the Heally system, and integrates the wires and sensors into the fabrics. We can envisage an reduction of the number of sensors by using electrically derived respiration from the ECG signal. However, this requires further investigation aimed at determining the optimal position of the electrodes and at assessing the quality of the estimate of the respiratory signal thus obtained.

The method proposed in this article requires a time-consuming preliminary stage of labeling recorded data into sleep and wake states, which only few categories of persons (athletes, high-endurance workers, pathological cases, etc.) may be willing to do. However, our results show that the system can be pre-trained using data from a relatively small sample of subjects, and used in generalization mode for several other users.

Although the accuracy of the method in multi-user conditions is lower than in the single-user case, it is comparable to actigraphy methods, but with the advantage of achieving a better balance in the correct classification of sleep and wake periods. Interestingly, our results indicate that respiratory signals alone are sufficient and perform even better than the combined respiratory and ECG signals. Respiratory signals are convenient to measure because they do not require electrodes on the skin, and persons may wear the sensors for periods of several days and weeks. Therefore, we think that this method represents a very promising solution for continuous monitoring of sleep and wake states. In view of its application to larger population groups we plan to enlarge the dataset used for training, validation, and testing by including female subjects and increasing the range of subject ages included in the dataset. Our current work consists of the implementation of the algorithm into a DSP microcontroller that estimates the sleepiness of the person with the help of the models mentioned in the introduction.

Acknowledgments

The authors thank Jean-Christophe Zufferey and Sara Mitri for several useful discussions. The authors also thank all subjects who accepted to participate in this study. Special thanks go to Dr. med Werner Karrer, Dr. med Thomas Rote and Isabelle Arnold of the Luzerner Höhenklinik Montana, Switzerland for their help with physiological signal recording and sleep analysis.

References

1. T. Akerstedt, "Consensus statement: fatigue and accidents in transport operations." *Journal of Sleep Research*, vol. 9, no. 4, p. 395, 2000.

2. D. Dinges, M. Mallis, G. Maislin, and J. Powell, "Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management," Tech. Rep., 1998.

3. T. Horberry, L. Hartley, G. Krueger, and N. Mabbott, "Fatigue detection technologies for drivers: a review of existing operator-centred systems," *Human Interfaces in Control Rooms, Cockpits and Command Centres, 2001. People in Control. The Second International Conference on (IEE Conf. Publ. No. 481)*, pp. 321–326, 2001.

4. P. Achermann and A. A. Borbely, "Mathematical models of sleep regulation," *Frontiers in Bioscience*, vol. 8, pp. 683–93, 2003.

5. A. Rechtschaffen, A. Kales, R. Berger, and W. Dement, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Public Health Service, US Government Printing Office*, 1968.

6. T. Penzel and R. Conradt, "Computer based sleep recording and analysis," *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 131–148, 2000.

7. A. Sadeh and C. Acebo, "The role of actigraphy in sleep medicine." *Sleep Medicine Reviews*, vol. 6, no. 2, pp. 113–124, 2002.

8. R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, no. 5, pp. 461–9, 1992.

9. L. de Souza, A. A. Benedito-Silva, M. L. Pires, D. Poyares, S. Tufik, and H. M. Calil, "Further validation of actigraphy for sleep studies," *Sleep*, vol. 26, no. 1, pp. 81–5, 2003.

10. (2007) Sleeptracker. [Online]. Available: `http://www.sleeptracker.com`

11. (2007) Axbo shop. [Online]. Available: `http://www.axbo.com`

12. R. D. Ogilvie, "The process of falling asleep," *Sleep Medicine Reviews*, vol. 5, no. 3, pp. 247–270, 2001.

13. M. H. Bonnet and D. L. Arand, "Heart rate variability: sleep stage, time of night, and arousal influences," *Electroencephalography and Clinical Neurophysiology*, vol. 102, no. 5, pp. 390–396, 1997.

14. 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, no. 2, pp. 33–41, 2004.

15. Z. Shinar, S. Akselrod, Y. Dagan, and A. Baharav, "Autonomic changes during wake-sleep transition: A heart rate variability based approach." *Autonomic Neuroscience*, vol. 130, no. 1-2, pp. 17–23, 2006.

16. T. F. of the European Society of Cardiology and the North American Society of Pacing Electrophysiology, "Heart rate variability : Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

17. S. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 485–496, 2006.

18. A. Steptoe, V. Peacey, and J. Wardle, "Sleep duration and health in young adults," *Archives of Internal Medicine*, vol. 166, no. 16, pp. 1689–1692, 2006.

19. S. J. Closs, "Assessment of sleep in hospital patients: a review of methods." *Journal of Advanced Nursing*, vol. 13, no. 4, pp. 501–510, 1988.

20. A. Cohen, *Biomedical Engineering Handbook*, 3rd ed., ser. The electrical engineering handbook series. Boca Raton : CRC Taylor & Francis, 2006, vol. 2, ch. Biomedical Signal Analysis, pp. I1–I22.

21. F. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, Jan 1978.

22. M. Hagan and M. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

23. D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," in *Proceedings of the International Joint Conference on Neural Networks IJCNN*, vol. 3, 1990, pp. 21–26.

24. C. P. Pollak, W. W. Tryon, H. Nagaraja, and R. Dzwonczyk, "How accurately does wrist actigraphy identify the states of sleep and wakefulness?" *Sleep*, vol. 24, no. 8, pp. 957–965, Dec 2001.

25. C. A. Kushida, A. Chang, C. Gadkary, C. Guilleminault, O. Carrillo, and W. C. Dement, "Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients." *Sleep Medicine*, vol. 2, no. 5, pp. 389–396, Sep 2001.

26. W. W. Tryon, *Activity Measurement in Psychology and Medicine*, ser. Applied Clinical Psychology. Plenum Press, New York, 1991, ch. 6. Activity and sleep, pp. 149–195.

27. A. Sadeh, K. M. Sharkey, and M. A. Carskadon, "Activity-based sleep-wake identification: an empirical test of methodological issues." *Sleep*, vol. 17, no. 3, pp. 201–207, Apr 1994.

28. J. N. Y. Aziz, R. Karakiewicz, R. Genov, A. W. L. Chiu, B. L. Bardakjian, M. Derchansky, and P. L. Carlen, "On-silicon neural activity monitoring and time-frequency analysis for early detection of epileptic seizures," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, to be published.