Electrocardiogram-Based Automatic Sleep Staging in Sleep Disordered Breathing

S Redmond, C Heneghan

University College Dublin, Dublin, Ireland

Abstract

A system for electrocardiogram (ECG) based sleep staging in subjects with sleep-disordered breathing is described. Three sleep states are defined: wakefulness(W), REM sleep(R) and non-REM sleep. Features investigated include RR interval, RR standard deviation, RR spectra, respiratory frequency, RR interval differences, and an ECG-derived respiratory signal. A subject specific quadratic discriminant classifier was trained and tested, and yielded an estimated classification accuracy of 71% (Cohen's κ value of 0.37). When a similar subject-dependent classifier was trained and tested, the estimated classification accuracy dropped to 61% (κ=0.12). For comparison, an electroencephalogram (EEG) based classifier yielded a subject-specific accuracy of 76% (κ=0.51), and subject-independent accuracy of 75% (κ=0.43), indicating that EEG features are robust across subjects. We conclude that the ECG signal provides moderate sleep-staging accuracy, but features exhibit significant subject dependence

1. Introduction

In recent work, a system has been presented for using the electrocardiogram (ECG) in assessment for the presence of sleep-disordered breathing (SDB) [1]. However, a limitation of this system is that it provides no knowledge about sleep state to the clinician. Accordingly, we have conducted a study to see if the ECG alone can provide some degree of sleep staging.

At present, sleep staging is carried out as part of the polysomnogram scoring process. Polysomnography routinely records and analyzes electroencephalograms (EEG), electromyograms (EMG), electrooculograms (EOG), electrocardiogram (ECG), pulse oximetry and several measures of breathing. Following acquisition of the physiological signals, the subject's sleep is scored in blocks of 30 s into one of six stages: Wake, REM, and Sleep Stages 1, 2, 3 and 4, using the Rechtschaffen and Kales (R&K) standard [2]. Scoring is typically carried out in two stages; an automated system performs an initial classification, which is followed by manual scoring to correct errors. However, since the R&K rules are arbitrary, and subject to operator interpretation, even highly experienced scorers have some degree of variability (estimated κ coefficient is 0.80 [3]).

Correlates of EEG-defined sleep stages can be expected to be present in the ECG also, through autonomic modulation of the heart. Indeed, previous studies have shown that the ECG contains relevant information about sleep stages [4-8]. Several ECG derived features (powers in the VLF, LF and HF spectral bands, and the LF/HF ratio) have been described which allow discrimination between these stages. The aim of this study was to use the ECG alone to classify sleep into Wake (W), REM (R), or Non-REM Sleep (S) stages, and hence augment the system described in [1] for detection of SDB. Our methodology was as follows. Firstly, a subject specific system is trained by randomly selecting epochs from 20% of the night's sleep with suitable representation of the three defined classes. Features were extracted from each epoch, and a classifier model was trained to distinguish the three classes. The remaining 80% of the night's sleep was used to test the system. The purpose of this exercise was to illustrate the ability of the ECG features to discriminate between stages W, R and S on a subject dependent basis.

Secondly, a subject-independent system was constructed using training epochs drawn from all subjects. Its performance on a single subject's records was evaluated using a 'jackknife' paradigm (leave one subject out of the training data). In both the subject-specific and the subject-independent system the training data was used to train a quadratic discriminant classifier.

Finally, to provide a benchmark by which to assess our ECG-based automated system, we compared its performance with an implementation of a standard EEG-based automated sleep stager.

2. Database

All data was obtained from the Physionet MIT-BIH

Polysomnograph Database [8]. The database contains multiple channel recordings of subjects being evaluated for OSA in Boston's Beth Israel Hospital Sleep Laboratory. Sleep stage annotations and apnea events are included. For the purposes of this study, these sleep stage annotations were considered as "ground truth". All signals are sampled at 250Hz. The database contains 16 subjects, of which 15 were chosen for the study. All 16 subjects were male, aged 32 to 56 (mean age 43), with weights ranging from 89 to 152 kg (mean weight 119 kg).

3. Feature extraction

In designing our ECG-based sleep stager, we extracted features consistent with those suggested in the literature. In an attempt to remove subject-dependence from the features, we carried out a normalization step on the RR interval series. For each subject, a normalized RR series was calculated by dividing by the mean RR interval (producing an RR sequence with a unity mean). However, since we may want to calculate spectral features in cycles/second as well as cycles/interval, we retain both normalized and raw RR series.

RR-Interval Features: Spectral representations of the RR interval series have been widely used previously [4]. Sleep is traditionally staged in 30-second epochs. Due to the poor frequency resolution of a 30 s PSD estimate, we calculated an RR-interval spectrum based on 5 epochs centered on the epoch of interest. While this reduces the time-localization of the sleep stage information, we believe that this is offset by the increased spectral resolution. To calculate the power spectral density estimate, the data from the five epochs is zero-meaned, windowed (using a Hanning window), and the square of its Discrete Fourier Transform (DFT) is taken as a single periodogram estimator. The *x*-ordinate of this estimate is in cycles/interval, which can be converted to cycles/second by dividing by the mean RR. From this spectral estimate five features are calculated: normalized VLF (power in the 0.01–0.05 Hz band), normalized LF (power in the 0.05–0.15 Hz band), and normalized HF (power in the 0.15–0.5 Hz band). Normalization is achieved by dividing by the total power in the three mentioned bands. From the spectrum, we also estimated the mean respiratory frequency by finding the frequency of maximum power in the HF band, and the power at this frequency.

In addition to the RR spectral features, we also used a range of temporal RR features for each 30s epoch. These features were: the mean RR of the normalized RR, its standard deviation, and the difference between the longest and shortest RR interval in the epoch.

ECG Derived Respiratory Features: As an alternative source of information about respiration, we also derived an ECG-derived respiratory (EDR) signal. It was extracted by estimating the "envelope" of the ECG signal, as it is modulated by the change in resistance caused by the expansion and contraction of the chest during breathing [1]. The resulting time-domain signal was then normalized over the entire recording to have a zero mean, and unit variance (since the amplitude of this EDR modulation is highly subject and electrode-position dependent).

As with the RR-intervals the VLF, LF, HF, respiratory frequency, and power at the respiratory frequency are estimated, using a five-epoch window, centered on the epoch of interest. The mean and standard deviation of each epoch's EDR was also calculated.

4. Quadratic discriminant classifier

Following the feature extraction stage described above, each epoch now has an associated set of 8 RR-based and 7 EDR-based features. The tool used for classification is a quadratic discriminant classifier, based on Bayes' rule. Gaussianity of the feature vector distributions, and independence between successive epochs is assumed. A quadratic discriminant classifier is derived as follows. Let ω_i signify the ith class. In this application there are three classes, S, W, and R. Let $\mathbf{x}$ denote the feature vector corresponding to a certain epoch. Using Bayes' rule we wish to find the class i which will maximize the posterior probability:

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\omega_i)p(\mathbf{x} \mid \omega_i)}{p(\mathbf{x})} \tag{1}$$

Maximizing the LHS of Eq. (1) is equivalent to maximizing its log. Therefore, assuming a normal distribution for the feature vector, $p(\mathbf{x} \mid \omega_i)$ becomes:

$$p(\mathbf{x} \mid \omega_i) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right] \tag{2}$$

where $\boldsymbol{\Sigma}_i$ is the covariance matrix of the ith class, and $\boldsymbol{\mu}_i$ is the mean vector of the ith class. Substituting Eq. (2) into the natural log of Eq. (1), our problem is transformed into finding the class i which maximizes the discriminant value $g_i(\mathbf{x})$ for a given test feature vector $\mathbf{x}$:

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i \mathbf{x} + k_i \tag{3}$$

where

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}, \qquad \mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$
$$k_i = -\frac{1}{2}\boldsymbol{\mu}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i) \tag{4}$$

The class with the highest discriminant value is chosen as the assigned class for that feature vector. To construct the quadratic discriminant classifier, therefore, we must estimate the covariance matrix and mean for the features corresponding to each class, and also the prior probability

of the class occurring.

5. Subject specific classification

Firstly, we used the selected features and the quadratic discriminant classifier model to discriminate between the three classes W, R, and S for a single subject's recording. To train the classifier (i.e., estimate class prior probabilities, covariance matrices, and means) 20% of the epochs for that night are randomly selected. Before the training data is chosen the prior probabilities for each of the three stages occurring are estimated using all 15 subjects. These probabilities are calculated as: P(W)= 0.29, P(R)= 0.07, P(S)= 0.64. The training data is chosen in such a way that the ratios of each class are in the proportion of the prior probabilities where possible. However, if the covariance matrix is estimated using as many (or less) observations than there are features, the matrix will be singular, prohibiting the use of discriminant analysis. In such cases the class is simply eliminated from the training data. To test the system the remaining 80% of the subject's data is presented to the classifier.

Table 1 gives the results of such a classification scheme. Results are presented in two formats. The overall accuracy is the percentage of correctly classified epochs from the test set. Cohen's kappa statistic κ is also presented. It represents a better measure of performance than raw accuracy, since it takes account of agreements that would have occurred by chance [15]. κ takes on values between 0 and 1, with 1 indicating perfect inter-system reliability, and 0 indicating no agreement above that predicted by chance alone. A κ value above 0.7 is typically taken to indicate a high-degree of inter-system reliability. The results shown below are a double average. The accuracies and κ obtained for each of the 15 subjects are averaged to give mean accuracy and κ. Each accuracy and κ is itself derived from an ensemble of ten classifier runs, with differing selections of training data each time. Using all 15 features described in Section 3, the 3-class classifier obtains an average accuracy of 71% and a κ of 0.36. Using the RR derived features gives an accuracy of 72% and a κ of 0.37, while the set of EDR features give an accuracy of 71% and a κ of 0.36. The standard deviations of the accuracy are also included in Table 1 to show that the overall accuracy will vary significantly from run to run, and subject to subject.

Table 1: Classification results for subject specific system

Features	Mean Accuracy	Standard Dev.	Mean Kappa statistics κ
All features	71%	11%	0.36
RR features	72%	11%	0.37
EDR features	71%	12%	0.36

6. Subject independent classification

To construct a subject independent classifier, features from the 14 other subjects are pooled together to form the training data for the classifier. This is repeated 15 times, leaving one subject out of the training data each time. The remaining subject is used to test the system. Obtained accuracies and κ are averaged for an overall estimate of performance, with results shown in Table 2.

Table 2: Results for subject-independent system

Features	Mean Accuracy	Standard Dev.	Kappa statistics κ
All features	53%	14%	0.15
RR features	61%	14%	0.12
EDR features	57%	15%	0.13

The accuracy achieved by the subject independent system, using all features listed in Section 3, was 53%, and the κ was 0.15. The RR derived features gave an accuracy of 61% and a κ of 0.12. Finally, the EDR features attained an accuracy of 57% and a κ of 0.13. The performance of the classifier on the subject independent task was poor, and differed only slightly from chance.

7. EEG comparative results

To gain a perspective on the results listed in Sections 5 and 6, two identical systems were designed using spectral and time domain features from the EEG in place of the ECG features described in Section 3. These systems were designed in accordance with standard approaches outlined in the literature [9,10,11], which recommend using EEG spectral features for sleep staging. The EEG spectral features used are: average power in the delta (0.5 – 3 Hz), theta (3 – 7 Hz), alpha (7 – 13 Hz), beta (13 – 30 Hz), and spindle (12 – 14 Hz) frequency bands. The time domain features are activity, complexity and mobility [9].

Table 3: Classification results for EEG features

System	Mean Accuracy	Standard Dev.	Kappa statistics κ
Subject Specific	76%	10%	0.51
Subject Independent	75%	11%	0.43

Using the same training and classifier paradigm as outlined above, subject-specific and subject-independent classifiers were designed and tested. In the subject-specific classification task the eight EEG features achieved an accuracy of 76% and a κ of 0.51 using all features. However, the subject independent system performed almost as well, attaining an accuracy of 75% and a mean κ of 0.43, as summarised in Table 3.

8. Discussion and conclusions

Subject-specific and subject-independent simplified ECG-based sleep staging systems have been designed and compared to a standard EEG-based sleep staging system. The ECG subject-specific system performs slightly worse (72%, κ=0.37) than its EEG counterpart (76%, κ=0.51). However, its performance suggests that the ECG represents a valid physiological signal for estimating sleep stage with a reasonable degree of accuracy. To place the classification performance in context, consider that optimum EEG-based systems typically have performances in the 80-85% range [11-13] (averaged over both normal and pathological populations). Such systems are operating at accuracy levels comparable to human experts.

However, in the transition to a subject-independent system, the ECG based system is not successful. Unlike the EEG-based system, where performance figures are virtually unchanged, the ECG system performs little better than chance. Heuristically, this appears to be primarily due to the fact that the distribution of our chosen ECG features exhibits a large intersubject variability. At this point, it is unclear to us whether the variations are caused by inadequate choice of normalization strategy, or by real inter-subject physiological variations. For example, it is plausible that for different subjects who have epochs of sleep in which obstructive events occur, different cardiodynamic behavior will be observed, even if the epochs appear similar in the EEG.

A small number of non-EEG based sleep staging systems have been described in the literature. In the small study described in [16], the authors achieved excellent accuracy using measurements of RR intervals and respiration in normal infants. However, infants have quite different sleep patterns and cardio-respiratory variability than adults, so it is hard to know how their approach will generalize. In [17], body movement was used to achieve accuracies of between 78% and 89% in the same discrimination task as ours; however, they report their results using a subject-specific classifier. We conclude that the ECG signal does contain information related to sleep stages, but that a robust subject-independent ECG-based classifier has not yet been developed.

Acknowledgements

The authors are grateful to the Irish Research Council on Science, Engineering and Technology for their support.

References

[1] de Chazal P, Heneghan C, Sheridan E, Reilly R, Nolan P, O'Malley M. Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea. IEEE Trans. Bio. Eng. 2003;50:686-696.

[2] Rechtschaffen A, Kales A. Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Los Angeles: UCLA Brain Information Services/Brain Research Institute 1968.

[3] Whitney CW, Gottlieb DJ, Redline S, Norman RG, Dodge RR, Shahar E, Surovec S, Nieto FJ. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep. 1998 Nov 1;21(7):749-57

[4] Shinar Z, Baharav A, Dagan Y, Akselrod S. Automatic detection of slow-wave-sleep using heart rate variability. Computers in Cardiology 2001:593-596.

[5] Versace F, Mozzato M, De Min Tona G, Cavellero C, Stegagno L. Heart rate variability during sleep as a function of the sleep cycle. Biological Psychology 2003;63:149-162.

[6] Ichimaru Y, Clark KP, Ringler J,Weiss WJ. Effect of sleep stage on the relationship between respiration and heart rate variability. Computers in Cardiology 1990: 657-660.

[7] Penzel T,Bunde A, Heitmann J,Kantelhardt JW, Peter JH, Voigt K. Sleep stage-dependent heart rate variability in patients with obstructive sleep apnea. Computers in Cardiology 1999;26:249-252.

[8] Calcagnini G, Biancalana G, Giubilei F, Strano S, Cerutti S. Spectral Analysis of Heart Rate Variability Signal during Sleep Stages. Proceedings of the 16th Annual International Conference of the IEEE 1994;2:1252-1253.

[9] Ichimaru Y, Moody GB. Development of the polysomnographic database on CD-ROM. Psychiatry and Clinical Neurosciences 1999;53:175-177.

[10] Hjorth B. The physical significance of time domain descriptors in EEG analysis. Electroencephalogr Clin Neurophysiol. 1973;34(3):321-325.

[11] Flexer A, Gruber G, Dorffner G. Continuous unsupervised sleep staging based on a single EEG signal. Artificial Neural Networks - ICANN 2002 Lecture Notes in Computer Science. 2002;2415:1013-1018.

[12] Albertario CL, Zendell SM, Hertz G, Maberino MM, Feinsilver SH. Comparison of a frequency-based analysis of electroencephalograms (Z-ratio) and visual scoring on the multiple sleep latency test. Sleep. 1995;18(10):836-843.

[13] Schaltenbrand N, Lengelle R, Toussaint M, Luthringer R, Carelli G, Jacqmin A, Lainey E, Muzet A, Macher JP. Sleep stage scoring using the neural network model: Comparison between visual and automatic analysis in normal subjects and patients. Sleep. 1996;19(1):26-35.

[14] Katayama T, Susuki E, Saito M. Staging of Awake and Sleep based on Feature Map. Systems and Computers in Japan. 1995;26(7):98-107.

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

[16] Haddad GG, Jeng HJ, Lai TL, Mellins RB. Determination of sleep state in infants using respiratory variability. Pediatr Res. 1987 Jun;21(6):556-62.

[17] Jansen BH, Shankar K, Sleep Staging with Movement Related Signals, International Journal of Bio-Medical Computing. 1993;32 (3-4):289-297

Stephen Redmond.
Rm. 244, Electronic Engineering,
University College Dublin,
Belfield, D4, Ireland.
Stephen.Redmond@ee.ucd.ie