Automatic Classification of Sleep Apnea Epochs using the Electrocardiogram

P de Chazal, C Heneghan, E Sheridan, R Reilly, P Nolan, M O'Malley

University College Dublin, Dublin, Ireland

Abstract

This study investigated the automatic prediction of epochs of sleep apnea from the electrocardiogram. A large independently validated database of 70 single lead ECGs, each of approximately 8 hours in duration, was used throughout the study. Thirty five of these records were used for training and 35 retained for independent testing. After considering a wide variety of features we found that features based on the power spectral density estimates of the R-wave maxima and R-R intervals to be the most discriminating. Results show that a classification rate of approximately 89% is achievable.

1. Introduction

Sleep apnea is commonly defined as the cessation of breathing during sleep. Clinicians usually divide sleep apnea into three major categories – obstructive, central, and mixed sleep apnea. Obstructive sleep apnea (OSA) is characterized by intermittent pauses in breathing during sleep caused by the obstruction and/or collapse of the upper airway. This is typically accompanied by a reduction in blood oxygen saturation, and leads to wakening from sleep in order to breathe. Central sleep apnea (CSA) is a neurological condition which causes the loss of all respiratory effort during sleep, and is also usually marked by decreases in blood oxygen saturation. Mixed sleep apnea combines components of both CSA and OSA, though treatment of the OSA portion often spontaneously leads to improvement in the CSA condition also.

In order for sleep apnea to be considered as clinically significant, the apnea episodes should be of ten seconds or longer duration, and occur more than five times per hour (the exact definitions vary from specialist to specialist [1]). Patients suffering from sleep apnea are more prone to hypertension, heart disease, stroke, and irregular heart rhythms (arrhythmias). Continued interruption of quality sleep is also associated with depression, irritability, loss of memory, lack of energy, and a higher risk of car and workplace accidents.

Currently, a definitive diagnosis of sleep apnea is made using a polysomnogram. This is a recording in which multiple signals are recorded simultaneously from the patient while asleep (a typical polysomnogram includes measurements of blood oxygen saturation, blood pressure, EEG, ECG, EOG, EMG, nasal/oral airflow, chest effort, and abdominal effort). Typically a full night's sleep is observed before a diagnosis is reached, and in some patients a second night's recording is required. Because of the number and variety of measurements made, this test is somewhat uncomfortable for the patient and also has a relatively high cost. Techniques which can reliably provide a diagnosis of sleep apnea with less invasive measurements, and without the need for a specialized sleep laboratory will be of benefit.

This paper considers the use of the electrocardiogram (ECG) for detection of sleep apnea using a variety of classification features based on the ECG timing intervals and amplitudes. This classification technique was evaluated using a database provided by Philipps-University. This database contains ECG recordings which have been annotated on a minute-by-minute basis for the presence of sleep apnea. This provides a gold-standard measure based on the assessment of a clinical panel with access to the full set of polysomnogram measurements. The database contains signals from 70 subjects with approximately eight hours of data per subject. The ECG signal and classifications of 35 of the recordings are available for training and the other 35 classifications are withheld for independent validation of classifiers.

2. Methods

The ECG signals contained in the database consisted of 12-bit samples, recorded at a sampling rate of 100Hz. Unvalidated QRS onset times were also supplied with the ECG database.

2.1. Data pre-processing

The raw ECG signal was processed using a linear phase high pass filter with a cutoff frequency of 0.5 Hz to remove baseline wander. R-R intervals were defined as the

0276-6547/00 $10.00 © 2000 IEEE

Computers in Cardiology 2000;27:745-748.

interval between successive R wave maxima All the QRS detection times were realigned to the R wave maxima by searching for the maximum in the region 100 milliseconds beyond the QRS onset.

Plots of the R-R intervals defined in this manner showed that records 2, 6, 12 and 25 of the training data and 13, 14, 15, 17, 20, 25, and 26 of the testing data had a significant number of physiologically unreasonable R-R intervals. A first processing step was to correct these. All suspect R-R intervals were found by applying a median filter of width 5 to the sequence of R-R intervals. This provided a robust estimate for each R-R interval of its expected value. Variations from this expected value led to it being flagged as a suspect interval.

Extraneous QRS detections were found by comparing the sum of adjacent R-R intervals with the robust R-R estimate. If this sum was numerically closer to the robust estimate than either of the individual R-R intervals then an extraneous detection was present. The two R-R interval were merged to form a single interval.

Conversely, if an R-R interval was a factor of 1.8 times or more than the robust estimate then it is probable that a QRS complex was not detected. To recover the missing QRS complexes the R-R interval was divided by the sequence of integers 2,3,4,... until it best matched the robust estimate of the R-R interval. The single R-R interval was then subdivided by the appropriate integer to form a series of new detections. For each new detection, a search was made in region of 100 milliseconds either side of that detection for the maximum of the ECG signal. If this maximum was similar to the maxima of the surrounding QRS complexes, it time of occurrence was accepted as a valid QRS detection point otherwise the original new detection point was used. A visual inspection of the new R-R intervals showed a significant improvement. The R-wave amplitude was defined as the value of the ECG signal at the QRS detection points defined by the above processing.

2.2. Feature sets

The pre-processing steps outlined above result in (a) an ECG signal with baseline wander removed and (b) a robust set of valid R-R intervals. Based on these, we considered a large set of features that could potentially be used for classification. Since the database classifications were provided for one minute segments of data, features were generated for each one minute segment. Features typically reflected only data from within a single segment, though we also considered multi-segment features. Some of the features considered were:

- Mean RR interval
- Standard deviation of the RR interval
- First and second serial correlation coefficients of the RR intervals
- Other time domain measures of RR intervals such as NN50 and pNN50 [2]
- Allan variance of the RR intervals evaluated at various time scales
- Mean PR interval
- Standard deviation of the PR intervals
- Serial correlation coefficients of the PR intervals
- Count- (or rate-)-based spectrum of the RR intervals.
- Interval based spectrum of the RR intervals
- R-wave amplitude spectrum

A useful summary of RR-interval based measures is contained in [3]. PR-interval measurements were investigated since changes in autonomic function are known to modulate PR interval length [4]. The PR interval was defined by using a wavelet-based algorithm to detect the peak of the P-wave, which could then be used to derive a PR interval. All of the measures listed above, except for the R-wave amplitude spectrum, reflect timing information only. This emphasis on timing-based measures was based on two assumptions: (a) timing information is more robust to data acquisition artifacts such as noise, motion, baseline wander, *etc.*, and (b) the processes leading to apnea occur at a location external to the heart, so that amplitude and shape characteristics of the ECG only reflect local cardiac conditions, not the process leading to apnea. However, in light of our results, in which the R-wave amplitude appears to be significant, the second assumption may need further consideration.

Since our results show that the interval-based RR and the R-wave amplitude power spectral density are the most useful features for classification, it is worth carefully defining how these quantities were calculated. In particular, there is significant confusion in the literature between interval-based and count-based spectra of RR data, as noted in [5].

The RR interval based PSD was calculated in the following way. For each record (the complete set of one-minute segments for an individual subject), the sequence of RR intervals was normalized to have a zero mean and unit variance. The purpose of this was to make the spectral features independent of the overall heart rate for each patient. A sequence of RR intervals was associated with each one-minute segment. The index for this sequence was beat number, not time. The mean RR interval for that segment was removed from each value, to yield a zero-mean sequence. The sequence was zero-padded to length 256, and the fast Fourier transform was taken of the entire sequence. This yields a periodogram estimate of the power spectral density, which has a high variance. Averaging of four adjacent frequency bins was used to yield a 64-point PSD estimate (of which only bins 0–32 are relevant since bins 33-63 are complex conjugates of 1-31). The x-axis has units of cycles/beat (not Hz as for a rate-based PSD).

The R-wave amplitude spectrum was calculated as

	Number of Features	Cross-validated test set			Cross-validated training Set			Training set			Indep. test set
		Acc(%)	Sens(%)	Spec(%)	Acc(%)	Sens(%)	Spec(%)	Acc(%)	Sens(%)	Spec(%)	Acc(%)
RR	32	79.2	72.5	83.3	80.3	73.9	84.2	80.3	74.0	84.3	
R	32	77.6	64.8	85.6	78.8	67.2	86.0	78.9	67.2	86.1	
RR+R	64	84.1	80.0	86.6	85.2	81.8	87.3	85.2	81.8	87.3	
RR+R+REL	128	88.2	84.1	90.7	89.8	86.5	91.9	89.7	86.4	91.8	88.9
class size								*17045*	*6514*	*10531*	*17268*

Table 1: The overall accuracies, sensitivities and specificities for the different feature sets for the cross-validated test and training data, full training data and independent test data.

Key: **RR**: features derived from frequencies bins of the PSD of the RR intervals **R**: features derived from frequencies bins of the PSD of the R wave heights **RR+R**: both PSD features combined **RR+R+REL:** both PSD features combined plus the relative features (see text for description)

follows. For each record, the sequence of R-wave amplitudes was normalized to have zero mean and unit variance. A discrete sequence of R-wave amplitudes for each 1-minute segment was formed. The mean value for the block was removed prior to spectral estimation using the periodogram technique outlined above.

For all of the features considered above a relative feature was generated by taking the feature value in the current segment and subtracting off the mean of the same feature values of the four segments immediately prior and following the current segment. Thus, a value of a relative feature close to zero indicated that the feature was similar to its neighbours and a value far from zero indicated that it was different.

2.3. Classifier

A supervised training technique was used to derive all classifiers. In supervised training, a classifier model that maps the input features to the required output classes is chosen. The model has a set of adjustable parameters that are optimized using training data. For this study linear discriminants classifier models were used. This model provides a parametric approximation to Bayes rule [6], so in response to a set of input features the output of each classifier is a set of numbers representing the probability estimate of each class. The final classification is obtained by choosing the class with the highest probability estimate.

Linear discriminants partition the feature space into the different classes using a set of hyper-planes. Optimisation of the model is achieved through direct calculation and is extremely fast relative to other models.

Other classifier models including quadratic discriminants and neural networks were trialed during the project. Although they resulted in an increase in training set accuracy their test set performance was always poorer than linear discriminants.

2.4. Feature selection

The performance of most classifier training algorithms is degraded when one or more of the available features are redundant or irrelevant. Redundant features occur when two or more features are correlated whereas irrelevant features do not separate the classes to any useful degree. The classification performance of a given set of features may often be improved by searching for a subset of the features with higher performance. Finding this optimal subset is generally computationally intractable for anything apart from small feature sets. This is because the number of possible subsets rises exponentially with size of the feature set. In practice a sub-optimal heuristic search such as the stepwise procedure is used [6]. A stepwise procedure for feature selection was used in this study.

When comparing the subsets, the best performance measure to use is the classification performance but again computational restrictions prevent this being implemented. We have used Wilk's Lambda [6], which is a measure of class separation to measure the performance of the subsets. A low value of Wilk's Lambda indicates good separation of the classes and indicates probable high classification performance. Hence feature selection involves finding a subset with the lowest value of Wilk's Lambda.

2.5. Classification performance estimation

When developing a classifier it is important to be able to estimate the expected performance of the classifier on data not used in training. The available data must be divided into independent training and testing sets. There are a number of schemes for achieving this and the most suitable for the size of data set used in this study, is *n*-fold cross validation [7]. This scheme randomly divides the available data into *n* approximately equal size and mutually exclusive "folds". For an *n*-fold cross validation run, *n* classifiers are trained with a different fold used each time as the testing-set, while the other *n*-1 folds are used for the training data. Cross validation estimates are generally pessimistically biased, as training is performed using a subsample of the available data.

For this study we divided the available training data

into 35 folds with each fold containing the data for one record.

In this study we report the overall classification accuracy, sensitivity and specificity. The overall accuracy is the percentage of total epochs correctly classified. Sensitivity is the percentage of apnea epochs correctly classified. The specificity is the percentage of normal epochs correctly classified.

2.6. Implementation

The work for this project was performed on a 600 MHz Pentium II PC running MATLAB version 5.3. All algorithms for feature selection, classifier training and data partitioning were developed in-house.

Calculation of the PSD features took approximately 15 minutes for the training and testing sets. A run of cross-validation with 128 features took about 5 minutes.

3. Results

Table 1 shows the classification results of the different feature sets. Three sets of results are shown. The cross-validated test set performance, the cross-validated training set performance and the independent test set accuracy. The testing set results are discussed below.

The interval-based PSD features of the R-R intervals resulted in a classification accuracy of 79.2% on the cross-validated test set. The sensitivity was 72.5% while the specificity was higher at 83.3%. Slightly lower results were obtained for the feature set derived from the PSD of the R-wave amplitudes. The accuracy for this set was 77.4% with a sensitivity of 64.8% of and specificity 85.6%. All though giving similar performance, the two PSD feature sets provided complementary classification information as, when combined, the accuracy increased to 84.1% (row 3 of table 1). When relative features were introduced to the combined PSD feature set the accuracy increased to 88.2% with a sensitivity of 84.1% and specificity of 90.7%. This classifier was submitted for independent classification and the overall reported accuracy was 88.9%.

Applying feature selection to the above feature groups did not improve the test-set classification performance so all features were retained. It was noted that at the expense of a small reduction in classification performance (<0.5%) the number of features could be significantly reduced.

By considering the full set of epochs for a patient record, we were also able to screen patients for the presence of clinically significant apnea. Of the thirty non-borderline records presented in the database, our system successfully classified them into normal and pathological cases.

4. Discussion

The results of this work indicate that detection of sleep apnea epochs is possible with an accuracy of approximately 89%. Both the time of occurrence of the QRS complex, and its R-wave amplitude are of use for classification. It is possible that the R-wave amplitude changes are an artifact of the ECG recording system due to the motion of the electrodes during breathing, rather than reflecting a physiological response of the heart. Nevertheless it was a consistent characteristic of the ECG.

If apnea is accompanied by changes in activity in the autonomic nervous system, it may influence the autonomic input to the heart. This may explain the importance of the RR-interval variations in the classification process.

Classification of epochs leads to the ability to screen patients for the presence of apnea with high reliability.

Acknowledgements

This work was supported by the Health Research Board, by the Conway Institute, and by the UCD President's Research Award. The authors are grateful to M. C. Teich of Boston University for bringing this topic to our attention.

References

[1] Mendelson W. Human Sleep – Research and Clinical Care, Plenum Medical. 1987.

[2] Task Force of the European Society of cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability – standards of measurement, physiological interpretation and clinical use. Euro. Heart J., 1996;17:354–81.

[3] Teich, MC, Lowen SB, Jost BM, Vibe-Rheymer K, Heneghan C. Heart rate variability: measures and models, in Nonlinear Biomedical Signal Processing, vol. II. ed. M. Akay, IEEE Press, Piscataway, NJ, 2000.

[4] Leffler CT, Saul JP, Cohen RJ. Rate-related and autonomic effects on the atrioventricular conduction assessed through beat-to-beast PR interval and cycle length variability. J. Cardiovasc Electrophysiol, 1994;5:2–15.

[5] DeBoer RW, Karemaker JM, Strackee J. Comparing spectra of a series of point events particularly for heart rate variability data. IEEE Trans. Biomed. Eng., vol. BME-31, 1984;384–7.

[6] Ripley BD. Pattern Recognition and Neural Networks. Cambridge University Press. 1996.

[7] Kohavi R. A study of cross validation and bootstrap for accuracy estimation and model selection. In: 14th Int. Joint Conference on Artificial Intelligence 1995;1137-43.

[8] Keyl C, Lemberger P, Pfeifer M, Hochmuth K, Geisler P. Heart rate variability in patients with daytime sleepiness suspected of having sleep apnoea syndrome: a receiver-operating characteristic analysis. Clinical Science, 1997.

Address for correspondence.

Philip de Chazal
Department of Electronic and Electrical Engineering,
University College Dublin, Dublin 4, IRELAND
philip@ee.ucd.ie