Cardiorespiratory-Based Sleep Staging in Subjects With Obstructive Sleep Apnea

Stephen J. Redmond* and Conor Heneghan, *Member, IEEE*

***Abstract*—A cardiorespiratory-based automatic sleep staging system for subjects with sleep-disordered breathing is described. A simplified three-state system is used: Wakefulness (W), rapid eye movement (REM) sleep (R), and non-REM sleep (S). The system scores the sleep stages in standard 30-s epochs. A number of features associated with the epoch RR-intervals, an inductance plethysmography estimate of rib cage respiratory effort, and an electrocardiogram-derived respiration (EDR) signal were investigated.**

A subject-specific quadratic discriminant classifier was trained, randomly choosing 20% of the subject's epochs (in appropriate proportions of W, S and R) as the training data. The remaining 80% of epochs were presented to the classifier for testing. An estimated classification accuracy of 79% (Cohen's κ value of 0.56) was achieved. When a similar subject-independent classifier was trained, using epochs from all other subjects as the training data, a drop in classification accuracy to 67% ($\kappa = 0.32$) was observed. The subjects were further broken in groups of low apnoea-hypopnea index (AHI) and high AHI and the experiments repeated. The subject-specific classifier performed better on subjects with low AHI than high AHI; the performance of the subject-independent classifier is not correlated with AHI.

For comparison an electroencephalograms (EEGs)-based classifier was trained utilizing several standard EEG features. The subject-specific classifier yielded an accuracy of 87% ($\kappa = 0.75$), and an accuracy of 84% ($\kappa = 0.68$) was obtained for the subject-independent classifier, indicating that EEG features are quite robust across subjects.

We conclude that the cardiorespiratory signals provide moderate sleep-staging accuracy, however, features exhibit significant subject dependence which presents potential limits to the use of these signals in a general subject-independent sleep staging system.

***Index Terms*—Breathing, ECG, EEG, obstructive sleep apnea, respiration, sleep stage.**

I. Introduction

SLEEP apnea is a cardio-respiratory disorder characterized by brief interruptions of breathing during sleep, and is often more generically described as sleep disordered breathing (SDB). Typical sleep patterns of a sufferer can involve heavy snoring interspersed with both partial and complete obstruction of the upper airway, leading to partial or complete waking and gasping for breath. The primary health implications of sleep apnea are its impact on the cardiovascular system (increased levels of hypertension, coronary arterial disease, arrhythmias), increased accident levels due to sleepiness, and quality of life issues. Moreover, obstructive sleep apnea (OSA) is not a rare condition. It occurs in 2%–4% of middle-aged adults [1] and in 1%–3% of preschool children [2]. However, despite the fact that apnea has such health and quality of life implications, there is a surprisingly low public and medical awareness of the illness. For example, of the 10–20 million sufferers in the U.S. with moderate-to-severe sleep apnea, it is estimated that only 10%–15% have been diagnosed [3].

A contributing factor to the low level of awareness of this disease is the relatively limited access to diagnostic tests in the general population. In most countries, the gold standard for diagnosis of sleep apnea is overnight polysomnography (PSG) (sleep study), which is carried out in a specialized hospital-based sleep laboratory. Polysomnography routinely records and analyzes electroencephalograms (EEGs), electromyograms, electrooculograms, electrocardiogram (ECG), pulse oximetry, airflow, and respiratory effort. One of the outcomes of PSG is the apnea hypopnea index (AHI), which is a measure of the average number of apneas and hypopneas per hour of sleep. A complete cessation of breathing is termed an apnea, and a partial reduction in airflow (e.g., to less than 50% of its normal value) is termed a hypopnea. It is widely accepted that PSG is a thorough and reliable test. However, it is relatively expensive, due to the need for the study to take place in a hospital setting, the requirement to have a sleep technician in attendance overnight, and the need to manually 'score' the resultant measurements. Hence, many sleep centers worldwide are currently operating at full capacity and PSG usually suffers from a low availability reflected in up to 6-mo. waiting lists for testing. Therefore, there is considerable interest in the development of reliable low-cost techniques for identification of subjects with sleep apnea.

An interesting possibility to overcome this diagnostic bottleneck is the use of more limited ambulatory cardio-respiratory studies, suitable for at-home use. Recent examples of such technology include the Embletta system (Medcare, Reykjavik, Iceland) which measures respiratory effort, pulse rate, airflow, oxygen saturation, position and activity [4], and the NovasomQSG (Sleep Solutions, Palo Alto, CA) [5] which measures airflow, respiratory effort, oxygen saturation, and pulse rate. These systems can directly reveal changes in the respiratory patterns, and hence be used to recognize apnea and hypopneas. Other recent work has focused on the use of the surface ECG obtained from Holter monitoring to reliably discriminate those suffering from obstructive sleep apnea [6]–[11]. These systems work by monitoring characteristic time-domain variations in heart rate [cyclical variations in heart rate (CVHRs)], which are associated with obstructive apnea events, and through the use of ECG-derived respiration signals.

Manuscript received September 23, 2004; revised June 11, 2005. This work was supported by the Irish Research Council for Science Engineering and Technology. *Asterisk indicates corresponding author.*

*S. J. Redmond is with the Department of Electronic Engineering, University College Dublin, Belfield D4, Ireland (e-mail: Stephen.Redmond@ee.ucd.ie).

C. Heneghan is with the Department of Electronic Engineering, University College Dublin, Belfield D4, Ireland (e-mail: Conor.Heneghan@ucd.ie).

Digital Object Identifier 10.1109/TBME.2005.869773

However, a limitation of both respiration-based and ECG-based systems is that they provide no information about sleep state to the clinician, even at the level of distinguishing sleep-wake states. Accordingly, this study was designed to see if cardio-respiratory measurements alone (ECG and respiratory effort in this case) would be sufficient to provide information about the sleep state of the subject.

It is worth briefly reviewing the concept of sleep state as defined clinically. Sleep is broken into two distinct classes: rapid eye movement (REM) and non-REM sleep. Non-REM sleep is further subdivided into four levels—sleep stages 1–4, which represent progressively deeper stages of sleep. Sleep states are defined primarily with respect to the EEG, following rules established by Rechtschaffen and Kales [12]. The process of determining sleep state is called sleep staging and is typically carried out as the first part of the polysomnogram scoring process. 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 1–4. 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 somewhat arbitrary, and subject to operator interpretation, even highly experienced scorers have some degree of variability (with an estimated kappa (κ) coefficient of 0.80 [13]). In [14] five expert scorers from three sleep laboratories were asked to independently score 62 records. A mean agreement (where all 5 scorers agree on a epoch) of 73% was achieved.

Sleep staging is clinically useful in the assessment of sleep apnea for several reasons: the Apnea-Hypopnea Index counts only apneas and hypopneas which occur during sleep; and an overall level of sleep quality or sleep disruption can be judged by the relative distribution of sleep stages. Therefore, systems which attempt to derive an Apnea Hypopnea Index should ideally incorporate some mechanism for determining sleep state. Since sleep state by definition is based on EEG analysis, it is nontrivial to determine sleep state by measurement of other physiological variables.

However, it is not unreasonable to expect that correlates of the EEG-defined sleep stages can also be present in the ECG, primarily through autonomic modulation of the heart's activity. Indeed, previous studies have shown that the ECG contains relevant information about sleep stages [15]–[20]. In these studies, several ECG-derived features (powers in the very low frequency (VLF), low-frequency (LF), and high-frequency (HF) spectral bands, and the LF/HF ratio) have been described which allow discrimination with various degrees of accuracy between sleep stages. Changes in respiration have also been observed with respect to sleep state. For example, it is generally accepted that respiration tends to be more irregular during REM sleep than non-REM [21]. Kantelhardt *et al.* have proposed that long-range temporal correlation properties differ for REM and non-REM sleep [22].

Given this background, the aim of this current study is to see whether measurements of ECG and respiration can provide a classification at the level of Wake, REM Sleep and Non-REM Sleep (which we denote W, R, and S in the following), with the goal of augmenting ambulatory cardio-respiratory systems for detection of SDB.

Furthermore, we considered two possible scenarios. In the first, we attempt to design a *subject-dependent* classification system, using a supervised classification methodology. In such a scenario, each subject has at least one night of recording which has been at least partially sleep staged using full EEG recording. The resulting classifier can then be used to classify the unlabeled section of the night's sleep, or more likely, subsequent nights of recording. The practical benefit of that is to allow multi-night recordings in the home using the limited cardio-respiratory measurements. A theoretical benefit is that it shows whether or not there is sufficient information in the ECG and respiration signals to perform sleep staging.

A second scenario, which is more practical, is to design a *subject-independent* system, whereby classifier training is carried out across a number of subjects, and where the resultant system should provide robust performance on randomly chosen subjects not represented in the training set.

In order to assess the accuracy of our results, we will use several performance measures. The first is overall accuracy (which is simply the number of 30-s epochs correctly classified). The second is the kappa coefficient κ (see Appendix I), which is a measure that also accounts for classifications which agree purely due to chance alone. Finally, we will consider the error between the true sleep efficiency (total sleep time/time in bed), and the estimated sleep efficiency.

This paper is organized as follows. In Section II, we describe the datasets used, and the processing of these signals to produce robust RR, ECG-derived respiration, and respiratory effort signals. In Section III, we describe the methodology for designing both subject dependent and subject-independent classifiers. We provide details of the features used in classification, feature selection techniques, the chosen classifier model, and the validation techniques used to reduce training bias. In Section IV, we describe the results of the classifiers designed in Section III, and also make comparison with a benchmark EEG-based automated sleep staging system designed on the same data sets. Section V discusses some of the implications of our results.

II. Methods: Signal Processing

A. Database

Data from a total of 37 subjects who were being evaluated for the presence of obstructive sleep apnea were used in this study. The subjects were drawn from data supplied by two laboratories. 27 subjects were examined at the Stanford University Sleep Disorders Center in autumn of 2003. A full polysomnograph (Sandman, Puritan Bennett, Kanata, ON, Canada) was obtained for each subject, but in this study we only consider the EEG channel C4-A1, the ribcage respiratory effort as measured by inductance plethysmography, and the ECG (modified lead V2). The EEG signal was sampled at 128 Hz, the ribcage respiratory effort at 16 Hz and the ECG signal at 256 Hz. The remaining 10 subjects were examined for OSA at St. Michael's Hospital, Toronto, also using a full polysomnograph (Sandman, Puritan Bennett, Kanata, ON, Canada). Again, only the C4-A1 EEG channel, ribcage respiratory effort, and ECG signals were utilized. The sampling rates were identical to those in the Stanford data. In both cases, sleep staging, and subsequent respiratory event scoring was carried out by a single experienced polysomnogram technician. Table I summarizes the demographic and clinical data for all subjects.

TABLE I
DEMOGRAPHIC INFORMATION

Subject group	Mean age	Mean BMI	Mean AHI	Mean Sleep Efficiency
All Subjects (37 subjects)	46.7 ± 10.4	27 ± 4.5	11.9 ± 16.4	76% ± 12%
Low AHI subjects (23 subjects)	46 ± 10.3	26 ± 4.1	3.4 ± 2.2	75% ± 12%
High AHI subjects (14 subjects)	47.3 ± 11.7	28 ± 4.9	26 ± 19.8	78% ± 11%

Later, we wish to loosely group the subjects into two classes: low AHI and high AHI. A low AHI in this study is arbitrarily defined as an $\mathrm{AHI} < 10$. There were 14 subjects with high AHIs, with a mean AHI of 26 and a standard deviation of 19.8. The remaining 23 subjects with low AHIs had a mean AHI of 3.4 and a standard deviation of 2.2.

B. Electrocardiogram Processing

A Hilbert-transform-based R peak detector was used to find the R peak locations in each subject's ECG [23]. The accuracy of the detector is estimated at approximately 98% [24]. The R peak locations are used both to derive RR-based features which may directly provide information about sleep stage, and in the calculation of an ECG-derived respiration (EDR) signal. No attempt was made to distinguish NN beats (normal sinus rhythm) from others. We noted that the ECG signal was slightly clipped in several subjects' recordings. This clipping will cause errors in the locations of some of the R peak locations and hence the RR interval series. To investigate the impact of such errors we simulated the effects of ECG clipping on the RR interval series by comparing the RR interval series derived from the unclipped ECG signals with the RR interval series from the same ECG signals after clipping at 80%. The details of the comparison are contained in Appendix II. We concluded that the RR interval series error introduced through clipping is negligible compared to the overall physiological variability of the RR intervals and hence does not introduce any significant error into our calculation of RR-based features.

C. Electrocardiogram-Derived Respiration (EDR)

Even though, we will subsequently use a directly acquired measure of respiratory effort (inductance plethysmograph), we decided to determine the utility of a respiratory estimate directly acquired from the ECG. It has been previously shown by several researchers that the magnitude of the ECG signal is amplitude modulated by respiration [25], [26]. Other factors may also cause changes in amplitude such as variations in electrode contact resistance (or capacitance) caused by movement, or a change in the electrical axis of the heart caused by altered body position. Hence our processing is aimed toward extracting the modulation that is the result of respiration and rejecting any electrode or body position influences. We label the derived estimate of respiration as the EDR signal.

We have found that a useful EDR signal can be constructed by tracing the envelope of the T peaks, or for a more noise robust estimate, integrating several samples around each T wave peak. Previous researchers have focused on calculating an EDR by using the QRS complexes [11]. However, as mentioned above, for our data sets it was found that the R peak was clipped in several subjects' recordings, so the T wave was used instead to derive the EDR. A description of the methods used in deriving the EDR signal from the ECG T waves is contained in Appendix III.

D. RR Interval Series Processing

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). This normalized RR interval series is denoted as $\mathrm{RR}_{\mathrm{norm}}$. However, since we may want to calculate spectral features in cycles/s as well as cycles/interval, we retain both normalized and raw RR series.

The RR interval series exhibits significant variation over the entire night's sleep. An interesting marker of changes in sleep state may be the *relative* changes in the RR interval series rather than the absolute value. We quantified the relative changes in the RR series by detrending the $\mathrm{RR}_{\mathrm{norm}}$ series with a 15-min moving average. We denote this deviant of the RR series as $RR_{detrend}$. The detrended RR is simply the current $\mathrm{RR}_{\mathrm{norm}}$ interval length minus the average $\mathrm{RR}_{\mathrm{norm}}$ length over the previous 15 min. This may help to account for underlying variation in the ECG due to circadian rhythm.

E. Outlier Correction in RR and EDR Signals

The RR interval series and the EDR signal both exhibit outlying points due to noise and QRS misdetections. To correct the impulse noise in both signals, we used the following technique. A smoothed estimate of the signal to be corrected is obtained using a median filter 5 samples in length. The difference of the signal from its median filtered equivalent is measured. If the absolute value of this difference at any point is above a certain threshold the signal at that point is replaced with its equivalent median value. The use of a median filtered equivalent signal essentially allows comparison of the errors to other points in the immediate signal locality, which is advantageous, as a fixed threshold will not work when the signal to be corrected contains any large amplitude drift component. Standard median filtering with a filter length of say, 3 samples, will also be reasonably effective at impulse noise removal, however, the signal will be altered (although perhaps negligibly) in regions containing no noise. Therefore, there is a tradeoff between preserving the signal in regions of no noise, and correcting for the effect of outliers in other regions.

F. Inductance Plethysmograph Preprocessing

Features directly related to respiration can be determined by analysis of the ribcage respiratory effort signal. This signal is also processed as follows. First, the signal is low pass filtered with a tenth-order Butterworth filter with a cutoff of 0.8 Hz, to remove HF noise and variation above respiratory frequencies. Since, the ribcage respiratory effort will in general be uncalibrated in terms of absolute tidal volume, we decided to normalize it for each subject, and consider only relative differences.

The ribcage signal is normalized by first detecting the turning points and then calculating the difference between sequential

peaks and troughs. The median peak-to-trough amplitude over the entire record is then determined and the signal is normalized by dividing through by this value, so that the median peak-to-trough amplitude is unity.

III. Methods: Classifier Design and Feature Generation

A. Feature Extraction

Given the set of ECG and respiration signals described above, we now consider the design of an automated sleep staging system based on those signals. In designing our sleep stager, we decided to extract features from each 30-s epoch which are consistent with those suggested by the literature.

RR-Interval Series Features: Spectral representations of the RR interval series have been widely used previously for a variety of applications [6]. To calculate a power spectral density estimate, the data ($\mathrm{RR}_{\mathrm{norm}}$ intervals falling within the epoch) from the epoch is zero-meaned, windowed (using a Hanning window), and the square of its discrete Fourier transform (DFT) is taken as a single periodogram estimate of the interval-based power spectral density. The x-ordinate of this estimate is in cycles/interval, which can be converted to cycles/s by dividing by the mean RR for the epoch. From this spectral estimate, five features are calculated:

- the logarithm of the normalized LF (power in the 0.05–0.15 Hz band),
- the logarithm of the normalized HF (power in the 0.15–0.5 Hz band), where normalization is achieved by dividing by the total power in the VLF, LF, and HF bands (0.01–0.5 Hz),
- the LF/HF power ratio,
- the mean respiratory frequency, which is defined by finding the frequency of maximum power in the HF band, and
- the logarithm of the power at the mean respiratory frequency.

In addition to the RR spectral features, we also used a range of temporal RR features for each 30-s epoch. These features were:

- the mean $\mathrm{RR}_{\mathrm{norm}}$;
- the standard deviation of $\mathrm{RR}_{\mathrm{norm}}$;
- the difference between the longest and shortest $\mathrm{RR}_{\mathrm{norm}}$ interval in the epoch;
- the mean value of the $\mathrm{RR}_{\mathrm{detrend}}$ in the epoch.

The difference between longest and shortest $\mathrm{RR}_{\mathrm{norm}}$ within the epoch is an attempt to quantify some of the dynamic behavior within the epoch (perhaps waking epochs are more dynamic than sleep, etc.) The mean $\mathrm{RR}_{\mathrm{detrend}}$ in one epoch is an attempt to examine the short-time variation in the RR interval series. Since each $\mathrm{RR}_{\mathrm{detrend}}$ value is a measure of the present $\mathrm{RR}_{\mathrm{norm}}$ relative to the previous 15 min of $\mathrm{RR}_{\mathrm{norm}}$, the mean $\mathrm{RR}_{\mathrm{detrend}}$ of an epoch is a measure of whether the heart rate in the present epoch is less than or greater than it has been over the last 15 min. This allows the discrimination of sudden rises in the heart rate, indicating short arousals, which may not rise significantly above the heart rate of other epochs of sleep.

ECG Derived Respiratory Features: The EDR epoch is taken as the EDR points corresponding to the R peaks falling within the epoch. The spectrum is calculated as for the RR interval series. From the EDR spectrum, the VLF (0.01–0.05 Hz), LF (0.05–0.15 Hz), HF (0.15–0.5 Hz) powers, respiratory frequency, and the power at respiratory frequency are estimated. The standard deviation of each epoch's EDR was also calculated.

RR-EDR Cross-Spectral Features: The VLF (0.01–0.05 Hz), LF (0.05–0.15 Hz), HF (0.15–0.5 Hz) powers were calculated from the cross-spectrum of the RR interval series and EDR for each epoch.

Ribcage Respiratory Effort Features: As with the RR interval series and the EDR, we calculate the ribcage respiratory effort spectrum as the square of the DFT of the ribcage respiratory effort signal for that epoch, windowed with a Hanning window. From the spectrum we calculate the logarithm of the power in the 3 bands—VLF (0.01–0.05 Hz), LF (0.05–0.15 Hz), and HF (0.15–0.5 Hz). The definition of these bands is taken directly from the corresponding definitions for ECG signals. Furthermore we estimate the respiratory frequency as the frequency of peak power in the range of 0.05 Hz–0.5 Hz, and also the logarithm of the power at that frequency.

In addition we derive several time domain features from the ribcage respiratory effort signal. The first is an estimate of its envelope power. We find the standard deviation of the peak values for the epoch, and similarly the standard deviation of the troughs. We then find the mean of the two values and divide by the standard deviation of the ribcage respiratory effort signal for the epoch. Essentially we are measuring the average top and bottom envelope powers as a fraction of the total signal power for the epoch. We denote this feature "*Envelope Power*." The second time domain feature attempts to measure a breath-by-breath correlation. We define a breath cycle as the time from the trough of one breath to the trough of the next. We find the cross-correlation of the adjacent breaths. Clearly in most cases the breaths will be of different lengths, in this case the shorter is padded with zeros to make it of equal length. We find the maximum cross-correlation value and divide it by the maximum of the energy of either breath alone to normalize the maximum cross-correlation value. The maximum cross-correlation values, for all pairs of adjacent breaths in the epoch, are then averaged. We denote this feature "*Breath-by-Breath Correlation*." The third time domain feature is a further measure of breath-by-breath variation. We take the standard deviation of the time between peak locations, similarly we take the standard deviation of the time between trough locations. We then take the mean of these two deviations. We denote this "*Breath Length Variation*". Finally we derive a second estimate of the respiratory frequency, using nonspectral means. We calculate the mean time between adjacent peaks and between adjacent troughs. The frequency of respiration is calculated as the inverse of this time. We denote this feature "*Time Domain Respiratory Frequency*." One final note to make in this section is that all estimates of respiratory frequency were further normalized by subtracting (from each epoch's estimate of the frequency) the median value of that parameter over all epochs for the entire night. This was deemed a necessary step as the mean respiratory frequency will vary from subject to subject. The median was subtracted as it is more robust than the mean to outliers.

The complete list of features for each 30-s epoch is given in Table II, and we will use the indices from this table in referring to possible feature combinations later.

TABLE II
FEATURE LIST

1	RR LF band	RR Interval based features
2	RR HF band	
3	RR standard deviation	
4	RR respiratory freq	
5	RR respiratory power	
6	LF/HF Ratio	
7	Longest Shortest RR difference	
8	Detrended RR mean	
9	RR mean	
10	EDR VLF band	EDR based features
11	EDR LF band	
12	EDR HF band	
13	EDR standard deviation	
14	EDR respiratory frequency	
15	EDR respiratory power	
16	RR-EDR Cross spectrum VLF band	EDR-RR Interval based features
17	RR-EDR Cross spectrum LF band	
18	RR-EDR Cross spectrum HF band	
19	Ribcage Respiratory effort VLF band	Ribcage Effort based features
20	Ribcage Respiratory effort LF band	
21	Ribcage Respiratory effort HF band	
22	Ribcage Respiratory effort respiratory frequency	
23	Ribcage Respiratory effort respiratory power	
24	Envelope Power	
25	Breath by breath correlation	
26	Breath length variation	
27	Time domain respiratory frequency	

B. Classifier Model: Quadratic Discriminant Classifier

Following the feature extraction stage described above, each 30-s epoch now has an associated set of 27 features—9 RR-based, 6 EDR-based, 3 cross-spectral-based and 9 ribcage respiratory effort based. The tool that we will use for classification is a quadratic discriminant classifier (QDC), based on Bayes' rule. In deriving a decision rule for a QDC, gaussianity of the feature vector distributions, and independence between successive epochs is theoretically assumed. Neither gaussianity nor independence will necessarily be satisfied. Note that in deriving features above, we have attempted to ensure that each feature has an approximately Gaussian distribution. This can be ensured, for example, by using the logarithm of the spectral powers, rather than their absolute values. Neglecting the dependence between successive epochs will not negate the authenticity of the classification results, however, classification accuracy may be improved if the dependence between epochs is considered as a post-processing step.

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. The feature vector in this case contains at most 27 elements, which are a selection the features described in the previous section. Using Bayes' rule we wish to find the class i which will maximize the posterior probability

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

Maximizing the left-hand side of (1) is equivalent to maximizing its logarithm. Therefore, assuming a normal distribution for the feature vector, $p(\mathbf{x}|\omega_i)$ becomes

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

where $\mathbf{\Sigma}_i$ is the covariance matrix of the ith class, and $\boldsymbol{\mu}_i$ is the mean vector of the ith class. Substituting (2) into the natural logarithm of (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

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

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.

C. Feature Subset Selection

In theory, with quadratic or linear discriminant classifiers, the addition of features containing little or no relevant information in the classification process will not degrade the performance of the classifier. One could simply include all features in the classification process and features containing no information will be "ignored" by the classifier. In practice this is rarely true—null features add "noise" to the system, and the removal of these redundant features can greatly improve results. However, with 27 features to choose from, we are allowed 2^{27} feature subset combinations, so it is obviously not feasible to search all possible combinations. Various algorithms exist which allow efficient searching of the feature subset combinations.

In this study, a form of the sequential forward floating search (SFFS) algorithm was exploited to identify the feature subset that will maximize the classification performance criterion [27]. The performance criterion we chose to maximize was Cohen's kappa statistic (κ) [28] described in Appendix I. The κ-coefficient is a measure of interrater agreement and takes into account the prior probability of a-specific class occurring. The two raters under comparison are our sleep staging system and the expert polysomnograph annotators.

The SFFS algorithm operates as follows. Three passes are made with the ordinary sequential forward selection (SFS) [27], so that three features are selected. One pass of the SFS simply adds the feature that most improves performance to the already selected features. Next, "unselection" of a selected feature is considered. The feature is found which most improves performance by its removal, and it is unselected. However, if no improvement is seen by the removal of any features then no features are unselected. Following the unselection phase the SFS is run again to select another feature. The cycle of a selection phase (with the SFS), followed by a possible unselection phase, is repeated until either the number of features required is reached,

or until the SFS phase fails to select a feature immediately followed by the failure of the unselection phase to remove a feature, in which case it is impossible for the selected feature subset to change and the algorithm must terminate.

The advantage of the SFFS over the SFS, or other greedy feature selection algorithms, is its ability to avoid nesting. Nesting occurs in greedy selection algorithms if a feature is selected early on that is not a member of the optimal feature subset, as it cannot be removed. Another algorithm, the *plus l, takeaway r* algorithm, can also avoid nesting. Its operation is similar to the SFFS and it provides similar results but has a longer execution time as it always removes l features, whereas the SFFS judiciously decides whether to remove a feature or not. Indeed the SFFS may not find the optimal feature subset, as it is inherently a sub-optimal search, but will often yield results comparable with those of an exhaustive search, with significantly less search time.

D. Design of a Subject-Specific System

While ideally we wish to develop a subject-independent cardiorespiratory-based sleep stager, a subject-specific system also has utility, both potentially in multi-night studies, and in providing proof-of-concept as to whether sleep staging can be provided by the ECG and respiration dataset alone.

As with all the systems described here, the quadratic discriminant classifier model is used 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 37 subjects. These probabilities are calculated as: $\mathrm{P(W)} = 0.26$, $\mathrm{P(R)} = 0.13$, $\mathrm{P(S)} = 0.61$. 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 of a class 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 containing insufficient observations is simply eliminated from the training data. To test the system the remaining 80% of the subject's epochs are presented to the classifier.

In Section IV, we present the overall accuracy (the percentage of correctly classified epochs from the test set), the absolute error from the true sleep efficiency, and Cohen's kappa statistic κ. A κ value above 0.7 is typically taken to indicate a high-degree of intersystem reliability. The accuracies and κ obtained for each of the 37 subjects are averaged to give a mean accuracy and κ. Each subject's accuracy and κ is itself derived from an ensemble of ten classifier runs, with differing selections of training data each time. The accuracies are derived from an ensemble average so as to remove any bias caused by the random selection of the training data.

E. Design of a Subject-Independent System

To construct a subject-independent classifier, features from the other 36 subjects were pooled together to form the training data for the classifier, again training a 3-class—W, R, and S—classifier by estimating the class prior probabilities, covariance matrices, and means. This was repeated 37 times, leaving one subject out of the training data each time. In each case the remaining subject was used to test the system. Obtained accuracies and κ, from each of the 37 runs, are averaged for an overall estimate of performance.

F. Design of an EEG Comparative System

To gain a perspective on the results of the subject-specific and subject-independent systems, two further systems were designed using spectral and time domain features from the EEG in place of the cardiorespiratory features described. These systems were designed in accordance with standard approaches outlined in the literature [29]–[34], which recommend using EEG spectral features for sleep staging. The EEG spectral features used are: average power in the delta (0.75–3.75 Hz), theta (4–7.75 Hz), alpha (8–12 Hz), spindle (12.25–15 Hz), and beta (15.25–30 Hz) frequency bands [34].

The powers in the designated frequency bands were calculated using a periodogram estimator. The 30-s EEG epoch was windowed using a sliding 2-s Hanning window with a 1-s overlap into 29 segments. The periodogram was constructed by averaging the square of the DFT of each segment over all 29 segments. The relevant frequency bands were then integrated to give the resulting band power.

The time domain features were the Hjorth parameters of activity, mobility and complexity [29]. They were derived from the entire 30-s epoch. Letting $\mathbf{x}$ denote the EEG epoch containing N samples, the Hjorth parameters are defined as

$$\mathrm{Activity}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\mathbf{x}})^2$$

$$\mathrm{Mobility}(\mathbf{x}) = \sqrt{\frac{\sigma(\mathbf{x}')}{\sigma(\mathbf{x})}}$$

$$\mathrm{Complexity}(\mathbf{x}) = \frac{\mathrm{Mobility}(\mathbf{x}')}{\mathrm{Mobility}(\mathbf{x})}$$

where $\mathbf{x}'$ is the first derivative of $\mathbf{x}$, $\sigma(\mathbf{x})$ is the standard deviation of $\mathbf{x}$, and $\mu_{\mathbf{x}}$ is the mean of $\mathbf{x}$. We also note that the activity is equal to the variance of $\mathbf{x}$.

Using the same training and classifier paradigm as outlined above, the subject-specific and subject-independent classifiers were designed and tested.

IV. Results

A. Subject-Specific Results

Table III details the results for the subject-specific sleep staging system for all subjects, and for subjects broken down by low and high AHI indices, after presenting all 27 features to the features selection algorithm.

In Table IV, we list the features selected by the SFFS algorithm. The indices listed refer to the feature list defined in Table II.

B. Subject-Independent Results

In Table V we present the results for the subject-independent sleep staging system for all subjects after presenting all 27 features to the feature selection algorithm.

TABLE III
SUBJECT SPECIFIC RESULTS

Subject Group	Mean Cohen's Kappa Coefficient κ	Mean Accuracy	Average Sleep Efficiency Error
All	0.56 ± 0.11	79% ± 5.4%	3.3%
Low AHI	0.6 ± 0.1	81% ± 4.6%	2.5%
High AHI	0.51± 0.09	77% ± 5.5%	4%

TABLE IV
SUBJECT SPECIFIC FEATURES

Subject group	Features Selected (in order of selection)
All	9, 27, 21, 8, 18, 20, 2, 16
Low AHI	22, 9, 23, 8, 15, 19, 2, 16
High AHI	8, 2, 9, 23, 27, 20

TABLE V
SUBJECT-INDEPENDENT RESULTS

Subject Group	Mean Cohen's Kappa Coefficient κ	Mean Accuracy	Average Sleep Efficiency Error
All	0.32 ± 0.11	67% ± 7.8%	11%
Low AHI	0.33 ± 0.1	68% ± 7.3%	11.5%
High AHI	0.31 ± 0.08	69% ± 7%	10%

TABLE VI
SUBJECT-INDEPENDENT FEATURES

Subject group	Features Selected (in order of selection)
All	22, 8, 20, 2, 4, 23, 5, 25, 9, 27, 21, 1, 11, 16, 17
Low AHI	19, 25, 4, 5, 8, 9, 16, 22, 12, 15, 11, 23
High AHI	8, 22, 2, 23, 25, 20, 14, 4, 16, 15

TABLE VII
EEG SUBJECT SPECIFIC RESULTS

Subject Group	Mean Cohen's Kappa Coefficient κ	Mean Accuracy	Average Sleep Efficiency Error
All	0.75 ± 0.12	87% ± 6.8%	2.7%
Low AHI	0.76 ± 0.12	87% ± 7.4%	3%
High AHI	0.73 ± 0.1	87% ± 5.8%	2.2%

TABLE VIII
EEG SUBJECT INDEPENDENT RESULTS

Subject Group	Mean Cohen's Kappa Coefficient κ	Mean Accuracy	Average Sleep Efficiency Error
All	0.68 ± 0.15	84% ± 8%	6.4%
Low AHI	0.7 ± 0.16	84% ± 9.4%	7.9%
High AHI	0.68 ± 0.13	84% ± 7.7%	5%

Table VI lists the features selected by the features selection algorithm in the subject-independent case.

C. Low AHI Versus High AHI

We wish to investigate the difference in performance between subjects with low apnea-hypopnea indices (AHI) and those with high AHIs. We repeat the above-mentioned subject-specific and subject-independent experiments with the subjects split into low AHIs ($<$10 apneas or hypopneas per hour) and high AHIs. There were 14 subjects with high AHIs the mean AHI was 26 and the standard deviation was 19.8. The remaining 23 subjects with low AHIs had a mean AHI of 3.4 and a standard deviation of 2.2.

D. Comparative EEG Results

Tables VII and VIII summarize the results of the subject-specific and subject-independent systems when trained using the 8 EEG features described earlier (no feature selection algorithm was used). As for the cardio-respiratory scoring system, we provide results broken down by high and low AHI class.

V. DISCUSSION AND CONCLUSION

Subject-specific and subject-independent simplified cardiorespiratory sleep staging systems have been designed and compared to a standard EEG-based sleep staging system using a database of 37 overnight polysomnographs.

The cardiorespiratory subject-specific system performs less well across all subjects ($\mathrm{Accuracy} = 79\%$, $\kappa = 0.56$) than its EEG counterpart (87%, $\kappa = 0.75$). However, its performance suggests that ECG and respiration together represent valid physiological signals for estimating sleep stage with a reasonable degree of accuracy. In fact, considering that $\kappa \geq 0.7$ is generally accepted as indicating a high degree of agreement, the cardiorespiratory subject-specific system performs well. To place the classification performance in context, consider that optimum EEG-based subject-independent systems typically have performances in the 80–85% range [30]–[33] (averaged over both normal and pathological populations).

However, in the transition to a subject-independent system, the cardiorespiratory-based system is less successful. Some degradation in performance should be expected; such degradation is seen even in the EEG-based system by a 0.07 decrease in the κ coefficient. However, in the cardiorespiratory-based system we observe a more significant drop in the κ coefficient of 0.24 (across all subjects). Heuristically, this appears to be primarily due to the fact that the distribution of our chosen cardiorespiratory features exhibits larger intersubject variability, as compared to the EEG-based features. These variations may be caused by inadequate choice of normalization strategy, or more likely by real intersubject physiological variations.

We have considered whether the presence of significant SDB is a detrimental factor in the performance of our systems. Intuitively, it is plausible that the micro-arousals associated with SDB could adversely affect the scoring system. As a benchmark, we compared the performance of the EEG-based systems on low-AHI and high-AHI subjects, and found that there was no statistically significant difference. For the subject-specific cardiorespiratory scoring system, there is a statistically significant $(\mathrm{p} < 0.05)$ decrease in performance in high AHI subjects as compared to low AHI subjects. This is not unexpected, as apneic events will disturb both the sinus rhythm and the respiration, although the same difference in performance is not observed in the subject-independent results. The subject-independent system performs similarly on both subject groups. This indicates that the influence of SDB-related cardiorespiratory

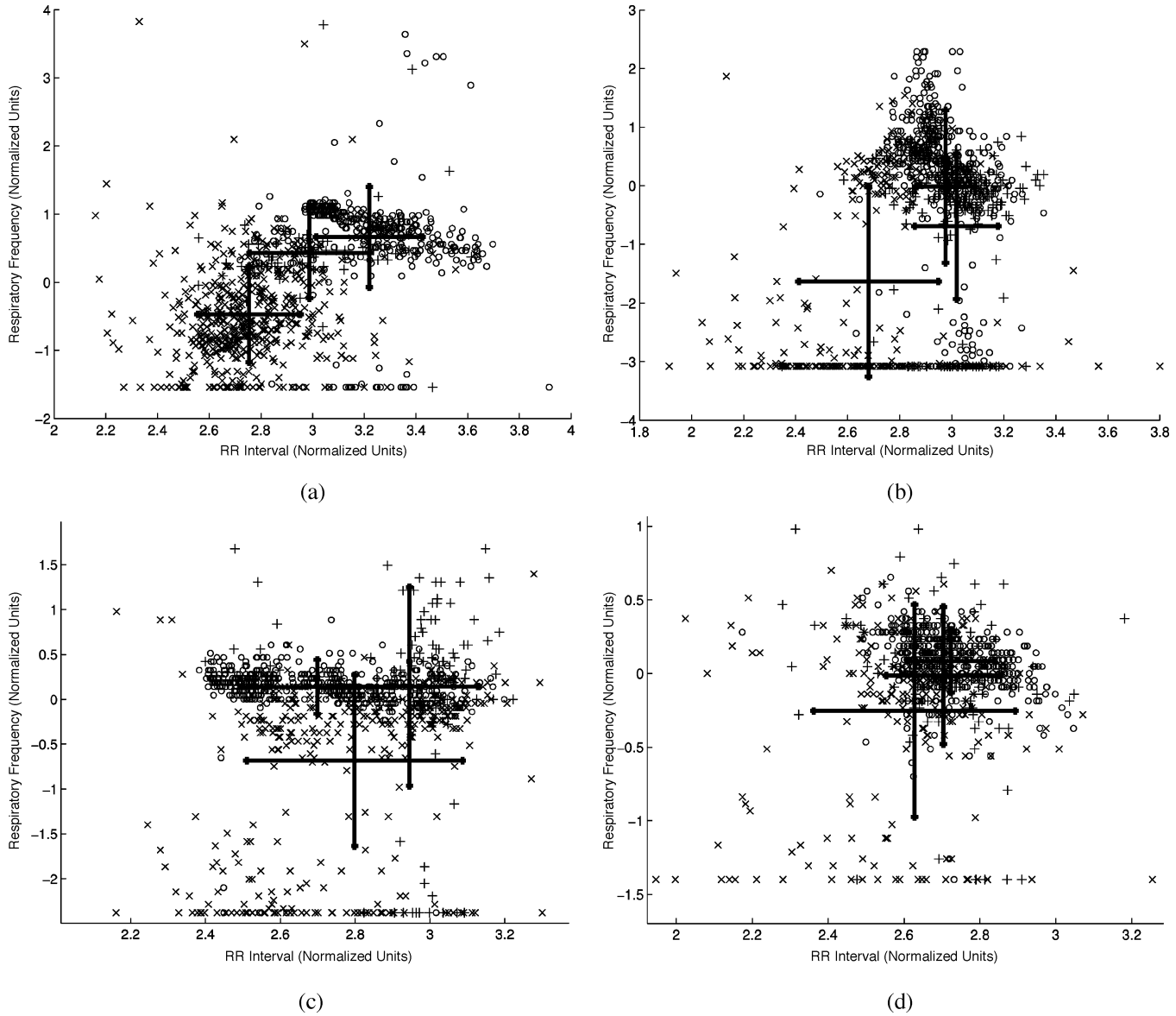

Fig. 1. (a) Plot of mean RR interval versus respiratory frequency for Subject 15. The values for Wake epochs are denoted with "x," non-REM sleep epochs are marked with "o," and R epochs have a "+" symbol. Also shown is the mean of each class and the width of plus and minus one standard deviation. There is reasonable separation between three classes. (b) Plot of mean RR interval versus respiratory frequency for Subject 4. There is separation of classes in RR interval, but not in respiratory frequency. (c) Plot of mean RR interval versus respiratory frequency for Subject 25. There is separation of classes in respiratory frequency, but not in RR interval. (d) Plot of mean RR interval versus respiratory frequency for Subject 6. There is minimal separation of classes.

events is not the sole limiting factor in cardiorespiratory sleep staging.

It is instructive to consider the wide intersubject variability of the features available from the cardiorespiratory signals. Fig. 1 shows four scatter plots (for four different subjects) of the epoch mean RR interval (feature 9) against the epoch mean respiratory frequency (feature 22). The subject in Fig. 1(a) has good interclass separation in both RR interval and respiratory frequency. Fig. 1(b) shows good separation in the RR interval only. Fig. 1(c) shows good separation of respiratory frequency, while Fig. 1(d) shows poor interclass separation for both features. In the subject-specific system the subjects in Fig. 1(a)–(c) will achieve good results, the subject of Fig. 1(d) will not. However, if we were to train the system using the data in Fig. 1(a)–(c), the results would be poor for the subject in Fig. 1(d).

We note that features 8 and 9, the mean $\mathrm{RR}_{\mathrm{detrend}}$ and the mean $\mathrm{RR}_{\mathrm{norm}}$ are consistently chosen by the feature selection algorithm. In particular we note that the mean $\mathrm{RR}_{\mathrm{norm}}$ is chosen above the mean $\mathrm{RR}_{\mathrm{detrend}}$ in the subject-specific systems, but this is reversed in the subject-independent case. It seems that the mean $\mathrm{RR}_{\mathrm{norm}}$ discriminates well in the subject-specific systems but the discriminating qualities are not consistent across all subjects. However, the mean $\mathrm{RR}_{\mathrm{detrend}}$ does seem to generalize well because it measures relative local change in heart rate and identifies arousals through a brief rise in heart rate, even if the heart rate is lower than some earlier epochs of sleep. Similarly, the onset of sleep is found by a slowing of the heart rate after some arousal, even when the heart rate is still faster than some epochs of waking. In addition, we note that while a large number of features are chosen by the feature selection algorithm in each instance, very comparable results can be obtained using the first five or six features chosen.

In this paper we utilize ECG clinical data which contains some clipping of the R peaks due, most likely, to a inadequate

choice of amplifier settings during the signal capture. We should state, however, that our results do not justify the relaxation of the quality applied to ECG recording in polysomnography. On the contrary, if research into ECG-based PSG techniques is to succeed, further high signal quality databases are required. Hence, we strongly encourage the continuation and improvement of quality ECG recording in PSG.

A possible point of interest is the delay inherent in some inductance plethysmography devices. Although there was no delay associated with the device used in this study, some methods may contain a delay in recording, relative to the ECG, of 2 s or 3 s. However, we argue that even when such a delay exists it is insignificant since we are using a 30-s epoch, and since only our interpretation of transitional epochs (epochs on the boundary of a sleep state change) will be affected by such a delay.

A small number of non-EEG-based sleep staging systems have been described in the literature. In the small study described in [35], 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 [36], 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, and do not give any results for a subject-independent system.

There are many potential confounding factors which we have not attempted to consider in this study. Specifically, it is plausible that pathologies such as congestive heart failure, cardiac dysrhythmias, and chronic obstructive pulmonary disorder, or medications such as beta-blockers or ACE inhibitors will influence both cardiac and respiratory dynamics. These effects may well lead to misclassification of sleep stage if not considered. However, in this first study, we are interested in evaluating potential performance in a general population being considered for obstructive sleep apnea, and without significant co-existing morbidity. We conclude that measurement of ECG and respiration can provide information related to sleep stages, and its performance on a subject-specific basis is quite impressive. However, further work will be required to improve the performance of a truly subject-independent automated cardiorespiratory sleep stager.

Appendix I
Cohen's Kappa Coefficient

As a measure of system performance we will use Cohen's Kappa Coefficient (κ) [28]. κ is a measure of interrater agreement, where the two raters in our case are the expert sleep technician (who scored the polysomnograph recordings) and the automated sleep staging system. κ is a chance-adjusted measure of agreement which varies from $\kappa = 1$ for perfect agreement to $\kappa = 0$ for a performance no better than chance. The need for such a measure is evident when we consider the relative proportions of the sleep stages, W:S:R, whose ratios are approximately 25:65:10. Therefore, with complete ignorance we could score all stages as S and achieve 65% accuracy, which may appear to be quite a reasonable performance. However, in this instance $\kappa = 0$, which is a better measure of performance.

TABLE IX
Confusion Table for Calculating Cohen's Kappa Coefficient

	Rater B				
Rater A	1	2	…	m	Total
1	$n_{11}(p_{11})$	$n_{12}(p_{12})$	…	$n_{1m}(p_{1m})$	$n_{1.}(p_{1.})$
2	$n_{21}(p_{21})$	$n_{22}(p_{22})$	…	$n_{2m}(p_{1m})$	$n_{2.}(p_{2.})$
:	:	:	⋱	:	:
m	$n_{m1}(p_{m1})$	$n_{m2}(p_{m1})$	…	$n_{mm}(p_{mm})$	$n_{m.}(p_{m.})$
Total	$n_{.1}(p_{.1})$	$n_{.2}(p_{.2})$	…	$n_{.m}(p_{.m})$	$n(1)$

For completeness, consider the definition of κ, and an example of its calculation.

Assume that 2 raters, A and B, are allowed to classify n observations into one of m classes. We construct Table IX, where each entry (i, j) is the number of times Rater A classified an observation as class i when Rater B classified as class j. Alternatively, we can express this as a fraction of the overall total by dividing by n. This fraction is shown in parenthesis in Table IX.

The total proportion of observer agreement (p_{o}) is the sum of the diagonal of Table IX

$$p_o = \sum_{i=1}^{m} p_{ii}. \tag{4}$$

The proportion of observations that would be classified by chance by both Rater A and Rater B into class i is

$$\Pr(\text{Rater A and Rater B classify as } i) = p_{i.}p_{.i}.$$

Therefore, the total proportion of agreement expected by chance (p_{e}) is given by

$$p_e = \sum_{i=1}^{m} p_{i.}p_{.i}. \tag{5}$$

Now we can define κ as the proportion of agreement adjusted for chance

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \tag{6}$$

We will give a small example to illustrate its use. Assume two raters classify the following 16 observations into one of three classes, say A, B and C. Tabulating the resulting classifications (as in Table IX) in Table X we see that they agree 10 times out of 16 observations giving an accuracy of 62.5%. We can now calculate κ using (4)–(6)

$$\begin{aligned} p_o &= 0.3125 + 0.25 + 0.0625 = 0.625 \\ p_e &= (0.375)(0.4375) + (0.375)(0.4375) + (0.25)(0.125) \\ &= 0.359\,375 \\ \kappa &= \frac{(0.625 - 0.359\,375)}{(1 - 0.359\,375)} = 0.4146. \end{aligned}$$

Appendix II
RR Interval Series Error

To investigate the error caused to the RR interval series by clipping in the ECG, we simulated the error using ten unclipped overnight ECG recordings. We ensured that the ECG signals

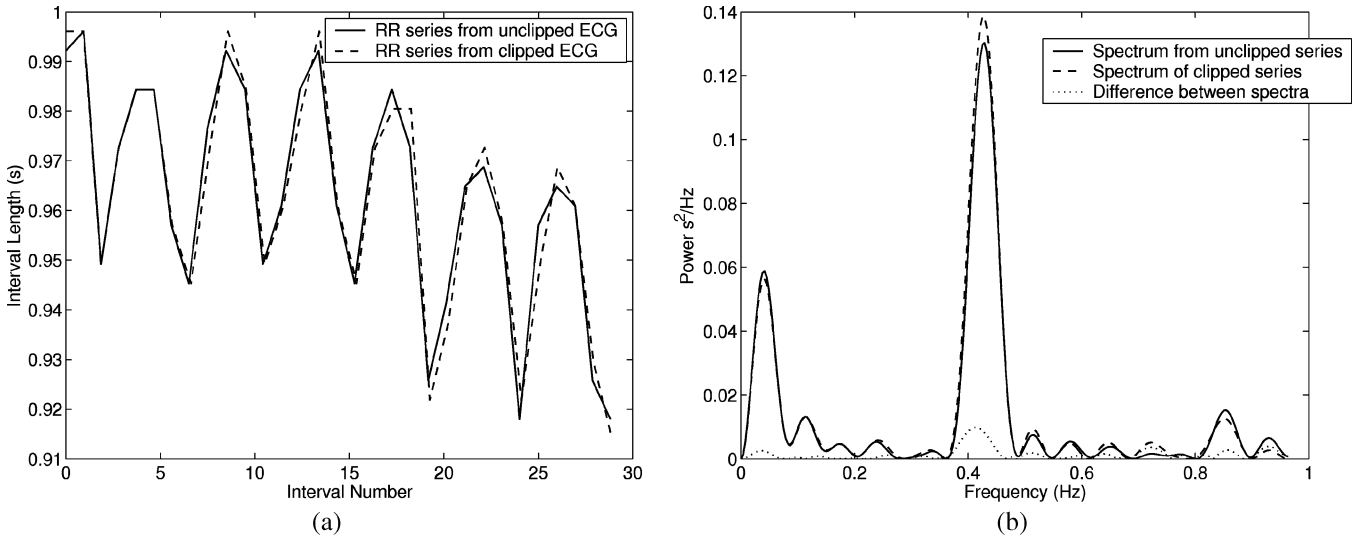

Fig. 2. (a) Plot of an epoch of the RR interval series versus time. The solid plot is the RR interval series derived from an unclipped section of ECG. The dashed plot is from the same section of ECG after clipping at 80% its median R peak amplitude. (b) Plot of the spectra of the given RR interval series from (a). Also shown in the dotted plot is the difference between the spectra powers.

TABLE X
EXAMPLE CONFUSION MATRIX FOR CALCULATING K

	Rater B			
Rater A	A	B	C	Total
A	5 (0.3125)	0 (0)	1 (0.0625)	6 (0.375)
B	2 (0.125)	4 (0.25)	0 (0)	6 (0.375)
C	0 (0)	3 (0.1875)	1 (0.0625)	4 (0.25)
Total	7 (0.4375)	7 (0.4375)	2 (0.125)	16 (1)

were unclipped by complete visual examination of the modulation in the ECG R peak amplitudes over the night's recording. (A clipped ECG will exhibit periods of zero modulation in the R peak amplitude.) Next, we found the R peak locations in each ECG signal and derived the corresponding RR interval series. The ECG signals were then zero-meaned and severely clipped at 80% of their median R peak amplitude. The error between each "true" RR interval, derived from the unclipped ECG, and the new RR interval, derived from the clipped signal, was computed, using our standard QRS detection algorithm. The mean RR interval error was 10^{-4} samples and had a standard deviation of 0.97 samples; for our data sampled at 256 Hz. We conclude that the clipping introduces a nonbiased error of small magnitude in the RR interval series. Table XI shows the histogram of all RR interval errors measured in samples. It is clear from the histogram that after the clipping the RR interval-derived is rarely more than 1 sample in error. To put this error in perspective the standard deviation of the difference between consecutive RR intervals was calculated as 9.1 samples over all 10 recordings. Hence the standard deviation of the RR interval error introduced is approximately 10% of the variation in the RR interval series itself. In addition, we argue that this type of noise introduced to the RR interval series will have little effect on the RR interval spectrum, as it will be spread uniformly across the spectrum. Also, since the RR interval error is zero-meaned we would not expect it to excessively perturb the mean RR interval of an epoch. Fig. 2(a) shows two RR interval epochs one from a section of unclipped ECG, and the other from the same section of ECG after clipping. Fig. 2(b) shows the corresponding interval-based spectra.

TABLE XI
RR INTERVAL ERROR HISTOGRAM

Count	555	1796	49992	172967	49328	2207	542
RR interval error (samples)	≤-3	-2	-1	0	1	2	≥3

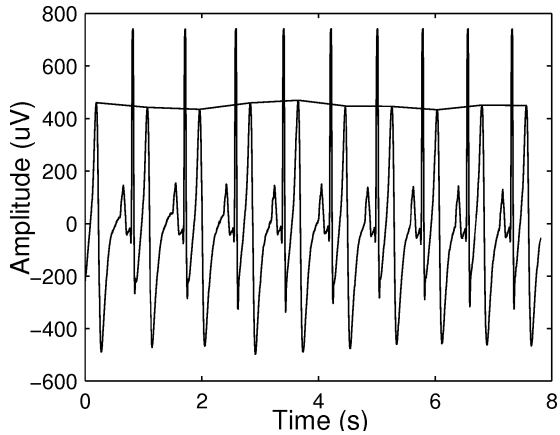

Fig. 3. Calculation of an ECG-derived respiration signal by measuring the amplitude of the T wave in the normalized baseline-reduced ECG for each cardiac cycle.

APPENDIX III
ECG DERIVED RESPIRATION SIGNAL

Since our goal in deriving the EDR is to consider relative changes in respiration amplitude or frequency, it is appropriate to attempt to normalize the ECG signal prior to calculating the EDR. This is due to the fact that the ECG amplitude may be affected by body position, and slow variations in electrode contact

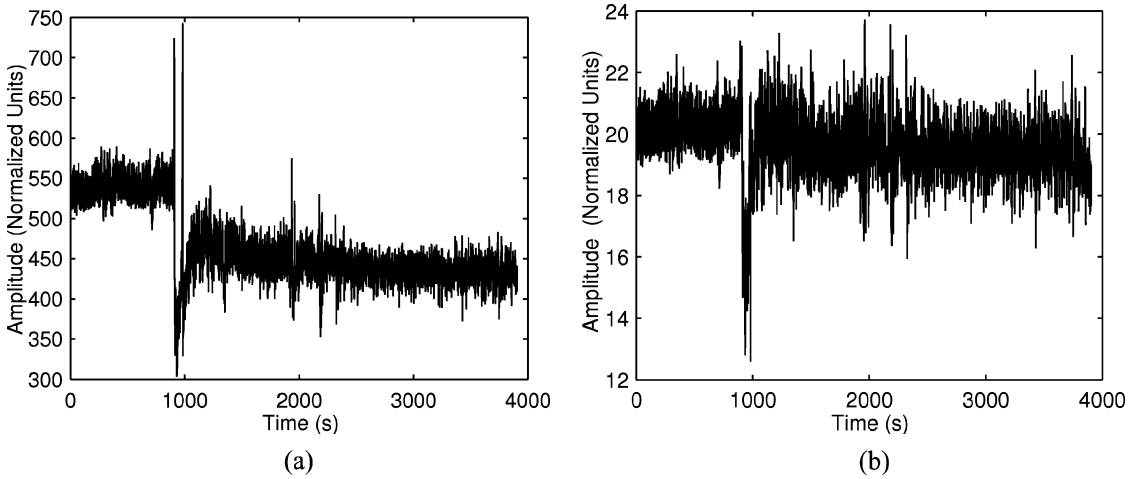

Fig. 4. (a) Unprocessed ECG-derived respiration signal resulting from calculation in Fig. 3. (b) Corrected ECG-derived respiration signal after removal of baseline drift.

impedance over a night's recording. To normalize the overall ECG amplitude, we attempt to ensure that the signal has approximately the same average power over time. This is achieved by using a sliding window of 2-min duration, with an overlap of 1 min from one window to the next. The standard deviation of the raw ECG signal under each window is calculated, and the signal is then linearly scaled by the standard deviation between consecutive window center locations so that the standard deviation of every section of ECG is approximately equal to unity. To be specific, if the standard deviation under the window centered at sample $t_{\rm d}$ is σ_1, and the standard deviation under the next window centered at sample $t_{\rm d+N}$ is σ_2, with the ECG sample points in between denoted as $s(t_{\rm d+}k)$, for $k = 0 \ldots (N-1)$, then the corrected ECG points are, $\hat{s}(t_{\rm d+}k) = s(t_{\rm d+k})/\{\sigma_1 + k[(\sigma_2 - \sigma_1)/N]\}$, for $k = 0 \ldots (N-1)$. In this way, the signal is linearly scaled between the two window centers and has a standard deviation approximately equal to unity. This is repeated for every consecutive pair of window centers. The choice of window length is arbitrary but must not be so small as to remove modulation effects at frequencies of interest (i.e., the EDR).

A further potential confounding factor in calculating the EDR is the presence of baseline drift in the original ECG signal. The most effective strategy appears to be removal of as much baseline drift in the original ECG, followed by subsequent post-processing of the EDR itself.

Baseline removal in the original ECG is achieved as follows. Using the normalized ECG signal described above, a median filter of width 3 s is centered at every R peak. The value resulting from the median filter centered at each R peak should be an approximation of the value of the PQ segment for that R peak, which is hopefully close to the isoelectric line for that signal. This value is used as an estimate of the baseline corresponding to that R peak. The signal is then linearly detrended between the baseline estimations at each R peak, i.e., the baseline between each R peak is represented piecewise linearly and this piecewise linear estimate is subtracted from the previously power-normalized ECG signal. The EDR signal is now estimated from the normalized, and baseline removed ECG signal by searching for the T wave peaks following every R peak and either choosing the peak value of the T wave peak, or the area under the peak and its neighboring samples. Fig. 3 illustrates how the EDR is derived from the T wave peak directly, which is appropriate for signals with relatively low noise. In this study we arbitrarily integrate 11 samples about the T wave peak. However, despite the use of ECG signal normalization, and baseline removal as described above, sudden baseline changes were still observed in the EDR.

These changes are probably due to T wave morphology changes, such as rotation of the electrical axis of the heart caused by altered body position.

A final step was taken to remove these artifacts. To obtain an estimate of its baseline, the EDR signal was passed twice through a median filter of length 4 s. The resulting baseline estimate was subtracted from the original. The first pass produces an estimate of the EDR baseline. The second pass removes oscillatory information that may still exist from the EDR in the first baseline estimate. The resulting EDR 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). Fig. 4 shows a section of EDR signal both before and after this final post-processing step.

ACKNOWLEDGMENT

The authors would like to thank C. Guilleminault of Stanford University, and P. Hanly and D. Lukic of St. Michael's Hospital Toronto, ON, Canada, for provision of data.

REFERENCES

[1] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," *New Engl. J. Med.*, vol. 328, pp. 1230–1235, 1993.

[2] T. Gislason and B. Benediktsdottir, "Snoring, apneic episodes, and nocturnal hypoxemia among children 6 months to 6 years old. An epidemiologic study of lower limit of prevalence," *Chest.*, vol. 107, pp. 963–966, 1995.

[3] T. Young, L. Evans, L. Finn, and M. Palta, "Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women," *Sleep.*, vol. 20, pp. 705–706, 1997.

[4] K. Dingli, E. L. Coleman, M. Vennelle, S. P. Finch, P. K. Wraith, T. W. Mackay, and N. J. Douglas, "Evaluation of a portable device for diagnosing the sleep apnoea/hypopnoea syndrome," *Eur. Respir. J.*, vol. 21, no. 2, pp. 253–259, 2003.

[5] J. A. Reichert, D. A. Bloch, E. Cundiff, and B. A. Votteri, "Comparison of the NovaSom QSG, a new sleep apnea home-diagnostic system, and polysomnography," *Sleep Med.*, vol. 4, pp. 213–218, 2003.

[6] F. Roche, J. M. Gaspoz, and I. Court-Fortune *et al.*, "Screening of obstructive sleep apnea syndrome by heart rate variability analysis," *Circulation*, vol. 100, pp. 1411–1415, 1999.
[7] F. Roche, D. Duverney, and I. Court-Fortune *et al.*, "Cardiac interbeat interval increment for the identification of obstructive sleep apnea," *Pacing Clin. Electrophysiol.*, vol. 25, pp. 1192–1199, 2002.
[8] F. Roche, V. Pichot, and E. Forza *et al.*, "Predicting sleep apnea from heart period: a time-frequency wavelet analysis," *Eur. Respir. J.*, vol. 22, pp. 937–942, 2003.
[9] P. K. Stein, S. P. Duntley, P. P. Domitrovich, P. Nishith, and R. M. Carney, "A simple method to identify sleep apnea using Holter recordings," *J. Cardiovasc. Electrophyiol.*, vol. 14, pp. 467–473, 2003.
[10] K. Dingli, T. Assimakopoulos, P. K. Wraith, I. Fietze, C. Witt, and N. J. Douglas, "Spectral oscillations of RR intervals in sleep apnea/hypopnea syndrome patients," *Eur. Respir. Jour.*, vol. 22, pp. 943–950, 2003.
[11] P. de Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley, "Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 6, pp. 686–696, Jun. 2003.
[12] M. Rechtschaffen and A. Kales, *Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Los Angeles: UCLA Brain Information Services/Brain Research Institute, 1968.
[13] C. W. Whitney, D. J. Gottlieb, S. Redline, R. G. Norman, R. R. Dodge, E. Shahar, S. Surovec, and F. J. Nieto, "Reliability of scoring respiratory disturbance indices and sleep staging," *Sleep.*, vol. 21, no. 7, pp. 749–757, Nov. 1998.
[14] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset," *Sleep*, vol. 23, no. 7, pp. 901–908, 2000.
[15] T. Penzel, J. W. Kantelhardt, L. Grote, J. H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 10, pp. 1143–1151, Oct. 2003.
[16] Z. Shinar, A. Baharav, Y. Dagan, and S. Akselrod, "Automatic detection of slow-wave-sleep using heart rate variability," *Comput. Cardiol.*, pp. 593–596, 2001.
[17] F. Versace, M. Mozzato, G. De Min Tona, C. Cavellero, and L. Stegagno, "Heart rate variability during sleep as a function of the sleep cycle," *Biol. Psychol.*, vol. 63, pp. 149–162, 2003.
[18] Y. Ichimaru, K. P. Clark, J. Ringler, and W. J. Weiss, "Effect of sleep stage on the relationship between respiration and heart rate variability," *Comput. Cardiology*, pp. 657–660, 1990.
[19] T. Penzel, A. Bunde, J. Heitmann, J. W. Kantelhardt, J. H. Peter, and K. Voigt, "Sleep stage-dependent heart rate variability in patients with obstructive sleep apnea," *Comput. Cardiol.*, pp. 249–252, 1999.
[20] G. Calcagnini, G. Biancalana, F. Giubilei, S. Strano, and S. Cerutti, "Spectral analysis of heart rate variability signal during sleep stages," in *Proc. IEEE Engineering in Medicine and Biology Society 16th Annu. Int. Conf. (Engineering Advances: New Opportunities for Biomedical Engineers)*, Baltimore, 1994, pp. 1252–1253.
[21] D. W. Hudgel, R. J. Martin, B. Johnson, and P. Hill, "Mechanics of the respiratory system during sleep in normal humans," *J. Appl. Physiol.*, vol. 56, pp. 133–137, 1984.
[22] J. W. Kantelhardt, T. Penzel, S. Rostig, H. F. Becker, S. Havlin, and A. Bunde, "Breathing during REM and non-REM sleep: correlated versus uncorrelated behavior," *Physica. A.*, vol. 319, pp. 447–457, 2003.
[23] D. Benitez, P. A. Gaydecki, A. Zaidi, and A. P. Fitzpatrick, "The use of the Hilbert transform in ECG signal analysis," *Comput. Biol. Med.*, vol. 31, no. 5, pp. 399–406, 2001.
[24] R. B. Shouldice, L. M. O'Brien, C. O'Brien, P. deChazal, D. Gozal, and C. Heneghan, "Detection of obstructive sleep apnea in pediatric subjects using surface lead electrocardiogram features," *Sleep*, vol. 27, no. 4, pp. 784–792, 2004.
[25] G. Moody, R. Mark, A. Zoccola, and S. Mantero, "Derivation of respiratory signals from multi-lead ecg," *Comput. Cardiol.*, pp. 113–116, 1985.
[26] S. Travaglini, C. Lamberti, and J. De Bie, "Respiratory signal derived from eight-lead ecg," *Comput. Cardiol.*, pp. 65–68, 1998.
[27] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recogni. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.
[28] R. Bakeman and J. M. Gottman, *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1986.
[29] J. M. Hjorth, "The physical significance of time domain descriptors in EEG analysis," *Electroencephalogr. Clin. Neurophysiol.*, vol. 34, no. 3, pp. 321–325, 1973.
[30] J. M. Flexer, G. Gruber, and G. Dorffner, "Continuous unsupervised sleep staging based on a single EEG signal," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2002, vol. 2415, Artificial Neural Networks—ICANN 2002, pp. 1013–1018.
[31] T. Katayama, E. Suzuki, and M. Saito, "Staging of awake and sleep based on feature map," *Syst. Comput. Jpn.*, vol. 26, no. 7, pp. 98–107, 1995.
[32] C. L. Albertario, S. M. Zendell, G. Hertz, M. M. Maberino, and S. H. Feinsilver, "Comparison of a frequency-based analysis of electroencephalograms (Z-ratio) and visual scoring on the multiple sleep latency test," *Sleep.*, vol. 18, no. 10, pp. 836–843, 1995.
[33] N. Schaltenbrand, R. Lengelle, M. Toussaint, R. Luthringer, G. Carelli, A. Jacqmin, E. Lainey, A. Muzet, and J. P. Macher, "Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients," *Sleep.*, vol. 19, no. 1, pp. 26–35, 1996.
[34] C. Guilleminault, Y. Do Kim, S. Chowdhuri, M. Horita, M. Ohayon, and C. Kushida, "Sleep and daytime sleepiness in upper airway resistance syndrome compared to obstructive sleep apnoea syndrome," *Eur. Respir. J.*, vol. 17, pp. 883–847, 2001.
[35] G. G. Haddad, H. J. Jeng, T. L. Lai, and R. B. Mellins, "Determination of sleep state in infants using respiratory variability," *Pediatr. Res.*, vol. 21, no. 6, pp. 556–562, Jun. 1987.
[36] B. H. Jansen and K. Shankar, "Sleep staging with movement related signals," *Int. J. Biomedical Comput.*, vol. 32, no. 3–4, pp. 289–297, 1993.

Stephen J. Redmond was born in Dublin, Ireland, in 1980. He received the B.E. degree in electronic engineering from the National University of Ireland, Dublin, in 2002. He is currently working towards the Ph.D. degree in the Biomedical Digital Signal Processing Group, National University of Ireland, Dublin.

His research interests include pattern recognition, biomedical signal processing, and machine intelligence.

Conor Heneghan (M'90) was born in Dublin, Ireland, in 1968 and received the B.E. degree in electronic engineering from University College Dublin, in 1990 and the Ph.D. degree in electrical engineering from Columbia University, New York, NY, in 1995.

He is currently a Senior Lecturer in the Department of Electronic and Electrical Engineering at University College Dublin. He was previously Director of Tele-Informatics at the New York Eye and Ear Infirmary and a Postdoctoral Research Scientist at Boston University, Boston, MA. His research interests include signal and image processing for biomedical applications and signal processing for communications.

Dr. Heneghan is also a member of the Institution of Electrical Engineers (IEE) and the Institution of Engineers of Ireland (IEI). He is a member of the IEEE Engineering in Medicine and Biology Society, the Signal Processing Society, and the Communications Society. He is a reviewer for IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and IEEE TRANSACTIONS ON SIGNAL PROCESSING.