RECOGNITION OF SLEEP STAGES FROM SENSOR DATA

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

JIALEI YANG

Dr. James Keller, Thesis Supervisor

DECEMBER 2015

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

RECOGNITION OF SLEEP STAGES
FROM SENSOR DATA

presented by Jialei Yang,

a candidate for the degree of master of science,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor James Keller

Professor Marjorie Skubic

Professor Mihail Popescu

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Professor James Keller for the continuous support of my master's study and related research, for his patience, motivation, and immense knowledge. Furthermore, I would like to thank Professor Marjorie Skubic and Professor Mihail Popescu for their thoughtful advice and encouragement which helped me in all the time of research. My sincere thanks also goes to the members of the bed sensor team, who shared with me a lot of their research experiences and suggestions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Sleep is understood as a reversible state of unconsciousness, characterized by a decrease of activity and alertness [1]. It is an essential activity for humans to maintain health. The lack of sleep or low quality of sleep will affect normal activities and cause physical and mental issues. The ability of monitoring sleep quality continually can help find sleep disorders instantly. Furthermore, some research showed a relation between sleep and other diseases (Parkinson's disease [2], Alzheimer's disease [3]). Therefore, the study of sleep is highly important.

Sleep is usually divided into two main stages: rapid eye movement (REM) and non-rapid eye movement (NREM). NREM is further divided into three subclasses: N1, N2 and N3 by the American Academy of Sleep Medicine (AASM) [4]. Polysomnography (PSG) is a type of sleep study used to measure these sleep stages and diagnose different types of sleep disorders. The PSG system connects to various physiological signals including electroencephalogram (EEG), electromyogram (EMG), electrooculogram (EOG), electrocardiogram (ECG) and respiratory airflow. Trained experts give a sleep stage every 30 seconds based on the obtained signals. The system is usually placed at a sleep lab. So the subjects require not only the connection of various sensors and electrodes but also must spend the night in a bed that is different from their own [5]. These settings are inconvenient and also may affect a subject's sleep patterns. Therefore, a home-use and more efficient

system is required to monitor sleep.

The aim of this work is to study the possibility of sleep stage recognition using data from three different sensors: a bed sensor that produces a ballistocardiogram (BCG) signal, an ECG and an oro-nasal airflow sensor. The work focused on recognition of Awake, REM and NREM stages.

The Hydraulic Bed Sensor (MUHBS) proposed in [6] is a non-invasive sensor that can capture the BCG signal. The sensor is placed under the mattress. So this sensor can be used at home without any connection to the body.

Although a PSG system would be an ideal ground truth, access to a sleep lab was not obtained. So a Mindo-Hydra wearable EEG device was used as the ground truth. Electroencephalogram (EEG) is one of the key physiological indicators to identify sleep stages. The processing of the EEG signal provided the sleep stages detected by its automatic detection algorithm [7]. Data were collected with a healthy subject by sleeping on the bed sensor and wearing the EEG device simultaneously.

Previous work [], [8] have shown that heart rate, respiration and body movement information can be extracted from the bed sensor. From them, heart rate variability (HRV) and respiratory variability (RV) parameters can be calculated. HRV and RV have been applied to sleep analysis in many studies [9], [10], [11], [12] and showed good results. So these two types of parameters were used to solve the sleep stage recognition problem of the MUHBS. In addition to them, cepstral analysis was also applied to the BCG signal and the

linear frequency cepstral coefficients (LFCC) were used to represent the cepstral parameters.

However, comparing with the golden standard PSG system, the recognized sleep stages from the EEG system were less accurate. This may cause the performance of developed methods not to be truly reflective of actual sleep stages. For the purpose of verifying the extracted features and developed methods, two other databases: the MIT-BIH Polysomnographic Database (MITBPD) [13] and the Sleep-EDF Database (Expanded) [14] were also studied here. Subjects of the MITBPD were apnea patients and subjects of the latter database were healthy people. These two databases include several physiological signals and the sleep stages identified by experts. Hence, the accuracy of the ground truth could be ensured. In order to calculate the same HRV and RV parameters as the bed sensor dataset had, only the ECG signal from the MITBPD and the respiration signal from the Sleep-EDF database were used, respectively.

In this study, the support vector machine (SVM) classifier and threshold comparison classifier were applied to the three databases. HRV, RV and LFCC features were used with the bed sensor data; HRV and LFCC features were applied to the MITBPD data; RV features were used with the Sleep-EDF data. To the best of my knowledge, it was the first time LFCC features from BCG and ECG signals were employed for the sleep stage recognition problem. The results indicated this type of feature could improve the classification results, and the performances of the MITBPD were better than all of the

previous studies using the same database. In addition, it was found the apnea patients had different HRV patterns in sleep stages compared with previous research of healthy people. Finally the simple threshold comparison classifier with few rules was shown to be efficient for the sleep-EDF database.

This thesis is organized as follows: In the next section, the background and related work are introduced. In section 3, the detailed methods including data collection, pre-processing, feature extraction, classification and evaluation are described. Section 4 displays the experiments and their results. The discussions of the results are also presented in this section. Finally, a conclusion is given in section 5.

2. Background

2.1. Sleep

2.1.1. Sleep Stages

The criteria of sleep stages were first standardized in 1968 by Allan Rechtschaffen and Anthony Kales (R&K scoring manual) [15]. In 2004, the AASM commissioned a revision of sleep scoring rules, covering not only sleep stages but also the scoring of arousals, respiratory events, sleep related movement disorders and cardiac abnormalities [16]. The revised scoring manual was published in 2007 [4].

Both manuals divided sleep into two main stages: rapid eye movement (REM) and non-rapid eye movement (NREM). The difference is R&K divided NREM into four stages: N1 and N2 as light sleep, N3 and N4 as deep sleep while AASM combined N3 and N4 into stage N3. In NREM sleep, the activity of body slows down. The stage is characterized by low temperature, slow breathing and slow cardiac rhythms. On the contrary, the REM sleep is dominated by intense cerebral activity, irregular breathing and rapid and irregular cardiac activity [17].

Humans usually sleep following a sleep cycle alternating between NREM and REM. Each cycle lasts about 90 minutes and there are four to six cycles each night. During each sleep cycle, the sleep stages go from light to deep then go back to light and REM. The first REM

stage usually appears about 70 minutes after sleep onset. Then it occurs every 90 minutes with gradually longer duration [18]. During a night's sleep, REM stages usually account for 20~25% of the total time in adults. Such sleep cycles are usually depicted by a graph called hypnogram, in which the stages of sleep are plotted as a function of time. Figure 2.1 shows an example hypnogram of a normal person.

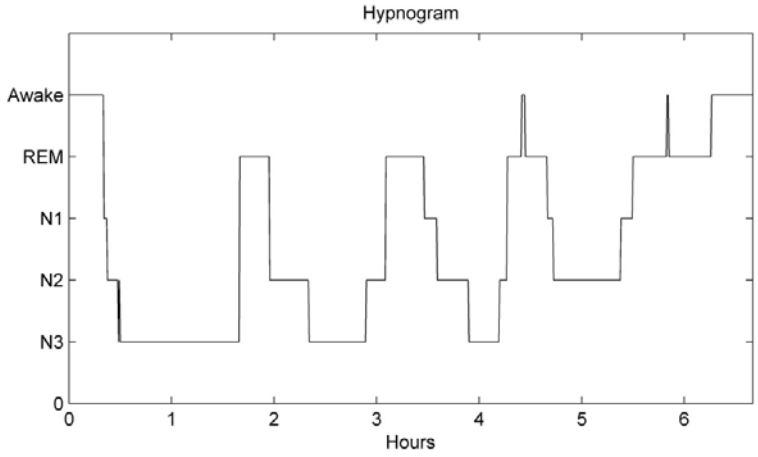

Figure 2.1: An example hypnogram

The sleep goes from awake to deep and the first REM period occurs about 80 min after fall asleep. Along with each sleep cycle, the proportion of REM stages increases while the proportion of deep stage (N3) decreases.

Both the R&K manual and the AASM manual recommended a epoch-by-epoch scoring method where each epoch is 30 seconds. A sleep stage is assigned to each epoch by observing specific patterns in the EEG, EMG and EOG signals. The ground truth of three datasets used in this work all followed this rule. Hence, the algorithms developed here also divided the acquired signals into 30 second (30s) windows and each 30s epoch can be classified to one sleep stage (Awake, REM or NREM).

2.1.2. Sleep Quality Parameters

In order to assess sleep quality reliably, a collection of parameters are defined. The

following list shows definitions of the sleep quality parameters [19] that have been used in this work.

- Time in bed (TIB): The duration of time from "lights out" to final awakening.
- Total sleep time (TST): The amount of actual sleep time in a recording.
- Sleep efficiency (SE): The ratio of total sleep time to time in bed.

$$\frac{\text{TST}}{\text{TIB}} \times 100$$

- Sleep onset latency (SL): The duration of time from "lights out", or bedtime, to the onset of sleep.
- Percentage of REM stages (%SR): The percentage of REM stages based on total sleep time (TST).

The general definition of sleep onset is the first non-awake stage (REM or NREM) after a subject goes to bed. However, one may change between Awake and sleep very frequently in the beginning of sleep. So [19] gives some examples of specific criteria to detect sleep onset. One of those criteria was used in this work: the first epoch recognized as a sleep stage (REM or NREM) is defined to be sleep onset. However, if the first stage is N1, the epoch is judged as sleep onset only if it followed by N1 or other non-awake stages for 3 minutes. Figure 2.2 shows the detected sleep onset in three situations. For (a), the first Non-awake stage is N1 at 5.5 min and there is no awake stage in the following 3 minutes (5.5-8.5 min). So sleep onset is 5.5 minutes; For (b), The first non-awake stage is N1 at 5.5 min. However, there is an Awake stage at 7 min, so the 3 min requirement is interrupted.

The N2 stage at 7.5 min is the first non-awake stage after the Awake stage at 7 min. So sleep onset is 7.5 minutes; For (c), The first non-awake stage is N2 at 5.5 min. So no 3 min requirement is needed and sleep onset is 5.5 minutes.

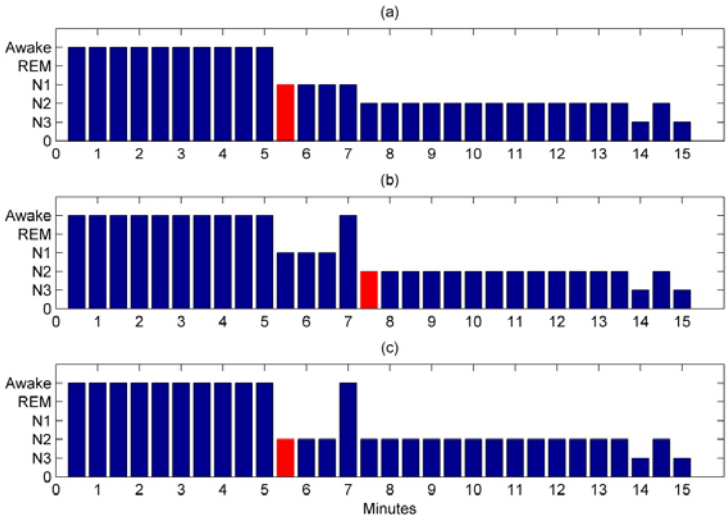

Figure 2.2: Detected sleep onset in three situations

Red bars indicate the detected sleep onset. (a) Sleep onset is 5.5 minutes. (b) Sleep onset is 7.5 minutes. (c) Sleep onset is 5.5 minutes.

2.2. Ballistocardiography (BCG)

The BCG is a non-invasive method developed with the aim to study cardiac activity by measuring body movements caused by the contraction of the ventricles and the blood flow in the systemic arterial tree [20]. The Committee on Ballistocardiographic Terminology decided to keep the terminology of the ballistocardiogram waves proposed by Starr (Figure 2.3) [21], [22].

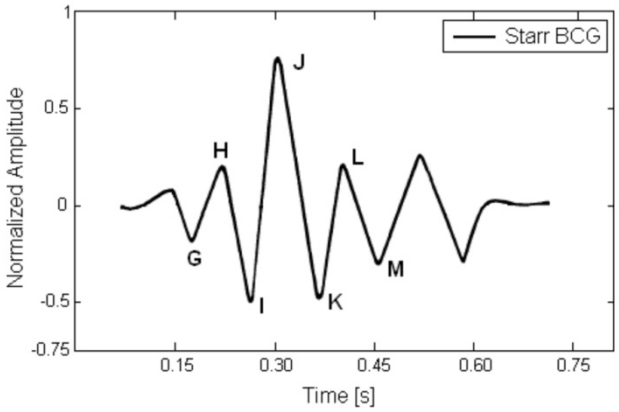

Figure 2.3: A typical shape of BCG waveform proposed by Starr

The waveform composes of 7 waves labeled with G to M. The J wave is usually the largest headward wave.

Although the shape of a BCG waveform may vary among different types of systems and different monitoring positions, most of the waveforms have similar shapes as depicted in the Starr BCG. Figure 2.4 shows an example waveform produced by the MUHBS.

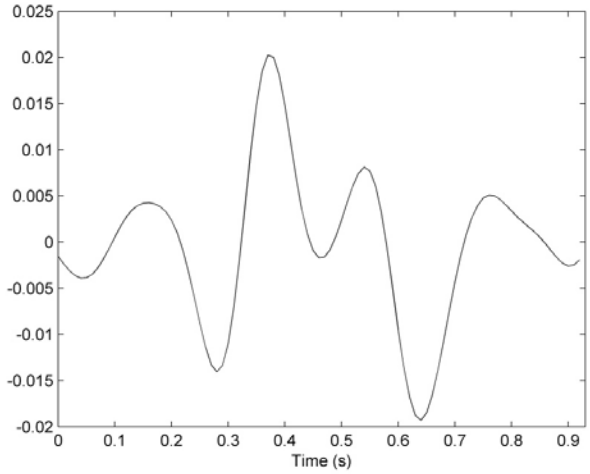

Figure 2.4: An example of BCG waveform from MUHBS

The waveform has a similar shape to the Starr BCG. The largest peak in the MUHBS BCG corresponding to the J peak in the Starr BCG is crucial for calculating heart rate. The differences between these J peaks are called JJ intervals which corresponds to RR intervals in the ECG signal, or also called beat-to-beat intervals. The detection of JJ intervals leads to the possibility for HRV calculation.

2.3. Heart Rate Variability (HRV)

In section 2.1.1, it was mentioned that irregular cardiac activity is a characteristic of REM sleep while slow and steady cardiac rhythms usually occur in NREM sleep. Thus, heart rate can be an indicator to separate different sleep stages.

Heart rate regulation is predominantly governed by the autonomic nervous system [23]. The autonomic nervous system is the part of the nervous system that controls automated body functions such as heart rate, respiratory rate and blood pressure. The system is divided into two parts: parasympathetic and sympathetic. In general, the parasympathetic nervous system predominates during rest by slowing heart rate, respiratory rate and lowering blood pressure. The sympathetic nervous system is responsible for stimulation of "fight-or-flight" by increasing heart rate, respiratory rate and blood pressure [24]. Previous research [25] also found that sympathetic nervous activity increases during REM and parasympathetic nervous activity increases during NREM. Thus, in essence, heart rate variability (HRV) provides a noninvasive method to estimate the autonomic function.

Variations in heart rate can be quantified by many methods: time domain methods, frequency domain methods, rhythm pattern analysis and non-linear methods. Two of the most common approaches, time domain and frequency domain methods, are used in this work. The implementation of these methods first requires the detection of heart beat intervals (HBI). Then these methods are applied to a certain period of intervals, which is 30 seconds in this work, to evaluate the variations.

2.3.1. Time Domain Methods

From a series of HBI, some statistical measures can be calculated to reflect the variations in heart rate. These measures may be divided into two classes, (a) those derived from direct measurements of the beat-to-beat intervals, and (b) those derived from the differences between beat-to-beat intervals [26]. Table 2.1 shows a list of commonly used statistical measures.

Table 2.1 List of time domain HRV measures

Measures	Description
SDNN	Standard deviation of all RR intervals.
RMSSD	The square root of the mean of the sum of the squares of differences between adjacent RR intervals.
pNN50	The percentage of adjacent intervals differing by more than 50 ms.
SDSD	Standard deviation of differences between adjacent intervals.

2.3.2. Frequency Domain Methods

In addition to direct statistic measures from HBI series, spectral analysis has suggested that power in specific frequency bands can be related to parasympathetic and sympathetic nervous system activity [27]. Previous research [28] demonstrated that power in a high frequency (HF) band (0.15–0.40 Hz) is related to parasympathetic nervous system activity and power in a low frequency (LF) band (0.04–0.15 Hz) is related to both sympathetic and parasympathetic influence. So the ratio of LF to HF is often used in order to cancel out the parasympathetic. Thus, the use of HF, LF and LF/HF can reflect the autonomic nervous system activity.

Frequency domain analysis is performed by calculating the power spectral density (PSD)

of the HBI. Then the power of specific frequency bands is calculated. Table 2.2 shows the list of frequency measures and their frequency ranges.

Table 2.2 List of frequency domain HRV measures

Measures	Description
VLF	Power in very low frequency range: 0-0.04Hz
LF	Power in low frequency range: 0.04-0.15 Hz
HF	Power in high frequency range: 0.15-0.4 Hz
LF/HF	Ratio of LF and HF.

2.4. Respiratory Variability (RV)

Similar to heart rate, respiration can also reflect the autonomic nervous system activity. Respiration is less rigorously controlled during sleep than in the waking state. Breathing in REM sleep is usually irregular compared with that in NREM sleep [29]. So respiratory variability parameters can also be useful for sleep stages recognition.

Unlike HRV, measures of RV haven't been well defined. Since the goal is to evaluate the variations in respiration, similar statistics calculated for HRV can be applied to breath-to-breath intervals. A time-domain method may be more appropriate here because the breath intervals in 30s usually only have 5 to 10 values corresponding to 10-20 breaths per minute. So other methods such as frequency domain analysis may not give a reliable estimate.

Besides, respiration activity consists of two steps: inspiration and expiration, so some other measures which can reflect the relation between these two processes may also be useful, such as: the ratio of time of expiration and inspiration and the ratio of amplitude of

expiration and inspiration.

2.5. Related Work

Many researchers have been working on automatic sleep stage recognition with the aim of developing a convenient and comfortable sleep scoring system with the ability of in-home monitoring. Most of the studies used the BCG signal from various bed sensors because of its non-invasive property. Others used the ECG, respiratory inductive plethysmography (RIP) and radar sensors. Three categories of features were studied in these works: HRV, RV and body movements. The most common performance measures reported in the literature are accuracy and Cohen's kappa coefficient (k). The kappa coefficient measures the inter-rater agreement which is described in [30] in detail. The advantage of the kappa coefficient is that it has a lower sensitivity to an imbalanced dataset. So it is suitable for the sleep stage recognition problem since each sleep stage accounts for different proportions of total sleep. Table 2.3 shows a commonly used scale which interprets the meaning of the specific kappa value.

Table 2.3: Meaning of kappa value

Kappa	Agreement
<0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

In the following section, some of the previous work is described.

a) Adnane, M. et al. [31] proposed a method for sleep-awake detection using the ECG. The method was tested on the MITBPD using a support vector machine (SVM) classifier. Three methods were employed to extract features from beat-to-beat intervals. They were HRV, detrended fluctuation analysis (DFA) and a new method they proposed called windowed DFA (WDFA). A subject-specific scheme was adopted, where 20% of a subject's data were randomly chosen as training set and the remaining part (80%) was used to test. A mean accuracy of 79.31% (12 features, k=0.41) and 79.99% (10 features, k=0.43) were reported.

b) Mendez, M. O. et al. [11] reported a three stage classification system to classify Awake, REM and NREM based on the bed Emfit sensor. Frequency domain HRV measures were calculated as feature sets. ECG signals from 17 subjects and BCG data from 6 subjects were recorded to train and test a Hidden Markov Model (HMM). A post-processing step was implemented by considering epochs with body movements (obtained from bed sensor) as Awake stages. The final results of three stages classification was 83% (k=0.42).

c) Tataraidze, A. et al. [12] presented a three stage classification method using respiratory signals. A set of 33 RV features were extracted from the RIP signal. Data from 29 subjects without sleep-related breathing disorders were collected and trained with a bagging method. However, what was the base classifier was not mentioned in the paper. A leave-one-subject-out cross-validation procedure was used for testing the

classification performance. Furthermore, four heuristics based on knowledge were applied. They were: 1) First 20 minutes were scored as wakefulness; 2) If an epoch did not belong to one of the nearest stages, it was scored as previous stage; 3) All REM epochs during the first 60 minutes of recordings were scored as previous stage; 4) If an interval between REM epochs was less than 15 minutes, all epoch included in the interval were scored as REM. The accuracy was 77.85±6.63% (k=0.59±0.11) without heuristics and 80.38±8.32% (k=0.65±0.13) after heuristics were applied.

d) Park, K. S. et al. [10] reported a threshold comparison method for sleep stages classification using the BCG signal obtained from load cell or polyvinylidenefluoride (PVDF) film sensors. Their method used threshold comparison classifiers with a hierarchical structure to classify Awake, REM, Light (N1 and N2) and deep (N3) stages. The dataset consisted of ten normal subjects and ten patients with Obstructive Sleep Apnea Syndrome (OSA). HRV features were calculated in each epoch. Thresholds were determined from the smoothed values of these parameters. They reported the accuracy of 77.1±3.3% in accuracy (k=0.58±0.06).

All of the above methods achieved the accuracy about 80%, and the kappa values ranged from 0.42 to 0.65 which covered moderate and substantial agreements according to table 2.3. However, after analyzing the reported confusion matrixes of these works, it was found that NREM sleep usually could be separated from other stages while the Awake and REM sleep were easy to mix together. Furthermore, the sleep stage recognition problem is really

sensitive to the dataset. Several proposed methods were used in this work without success. Thus, the sleep stage recognition problem needs to be further studied.

3. Methods

3.1. Dataset Description

3.1.1. Bed Senor Dataset

3.1.1.1.Hydraulic Bed Sensor

The MU bed sensor system consists of four hydraulic transducers. The detailed construction of each transducer and their placement is described in [6] and [8], respectively. Four transducers were placed vertically under a mattress (Figure 3.1).

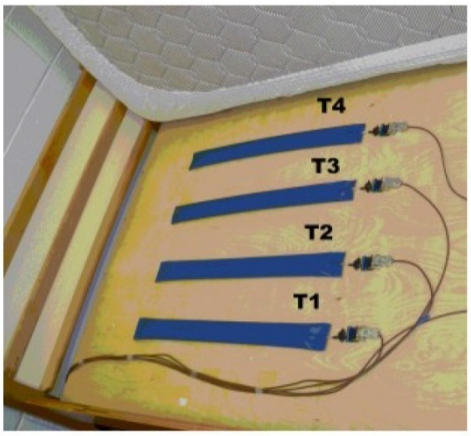

Figure 3.1: MU hydraulic bed sensor transducers placement (picture from [8])
Four blue transducers are labeled T1 to T4.

The main body of each transducer is filled with a certain volume of water and the end of the body is connected to a pressure sensor. Each transducer will produce two signals to the computer. One of the signal is the raw signal which comes from the Analog-to-Digital converter (ADC) connected to the pressure sensor. This raw signal is further filtered and amplified by an Amplifying/Filtering Card (AFC) to obtain a filtered signal with a

sampling frequency of 100 Hz. Figure 3.2 shows 30 seconds' signals of the four transducers when a healthy person is laying on the bed.

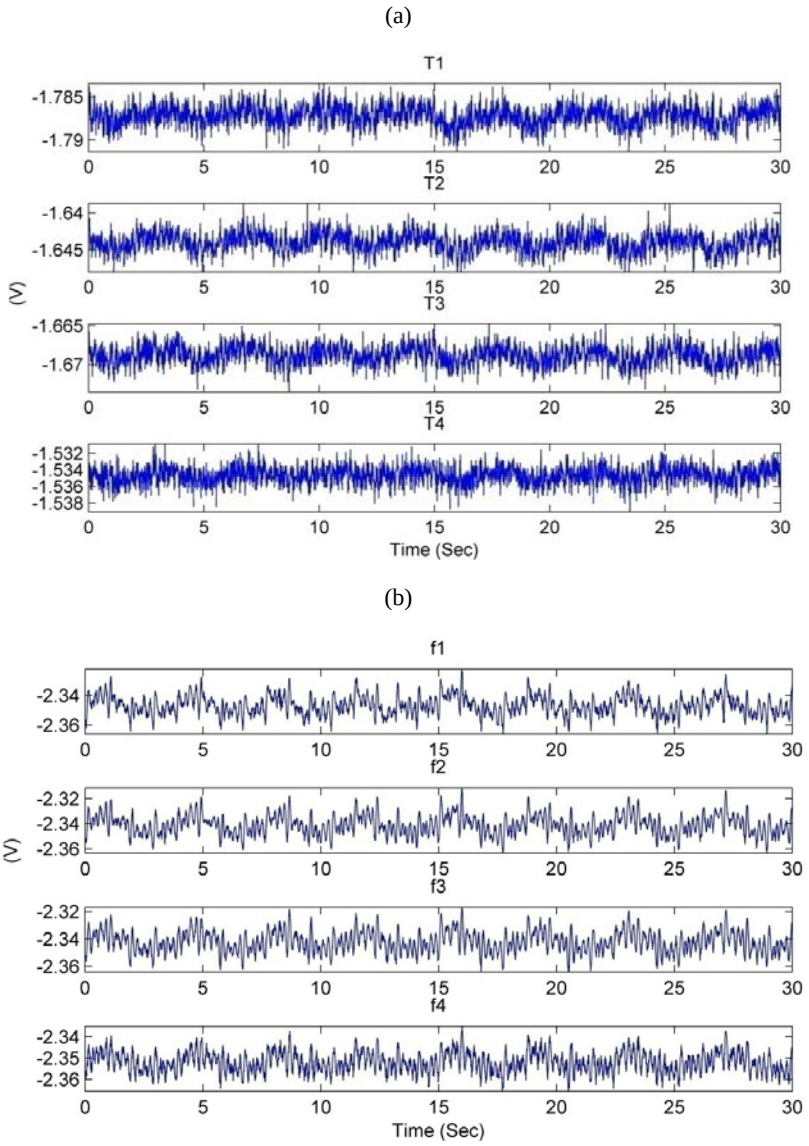

Figure 3.2: Thirty seconds of BCG signals for the four transducers
(a) raw signals. From top to bottem are signals from transducer 1 to 4; (b) hardware filtered signals. From top to bottem are signals from transducer 1 to 4

3.1.1.2.Mindo-Hydra Wearable EEG Device

The EEG device used as the ground truth was developed by the MINDO company [32]. The device consists of a wearable EEG band (Figure 3.3), along with software on Android

tablet which can communicate with the EEG band via blue tooth.

Figure 3.3: Mindo-Hydra wearable EEG band (picture from Mindo-Hydra wearable EEG device manual)

The sampling rate for the EEG signal is 128 Hz. For every 30s epoch, the software gives a predicted sleep stage. The system classifies sleep into four stages: Awake, REM, light (N1 and N2) and deep (N3). The outputs of the sleep stages is displayed on the android and is saved in a txt file. Figure 3.4 shows the interface of the android software with detected sleep stages.

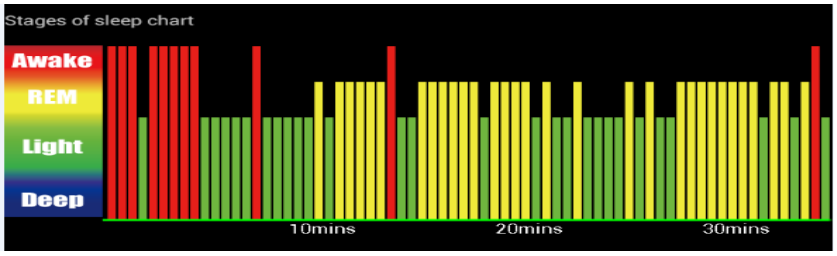

Figure 3.4: Interface of the android software with detected sleep stages for about 30 minutes. Each bar is 30s.

The exact performance of the EEG sleep stages recognition software with this EEG band is unknown. However, in [7], the same research center reported a hierarchical classification algorithm for sleep stage classification using forehead (FP1 and Fp2) EEG signals. The structure of their hierarchical classification is shown in Figure 3.5.

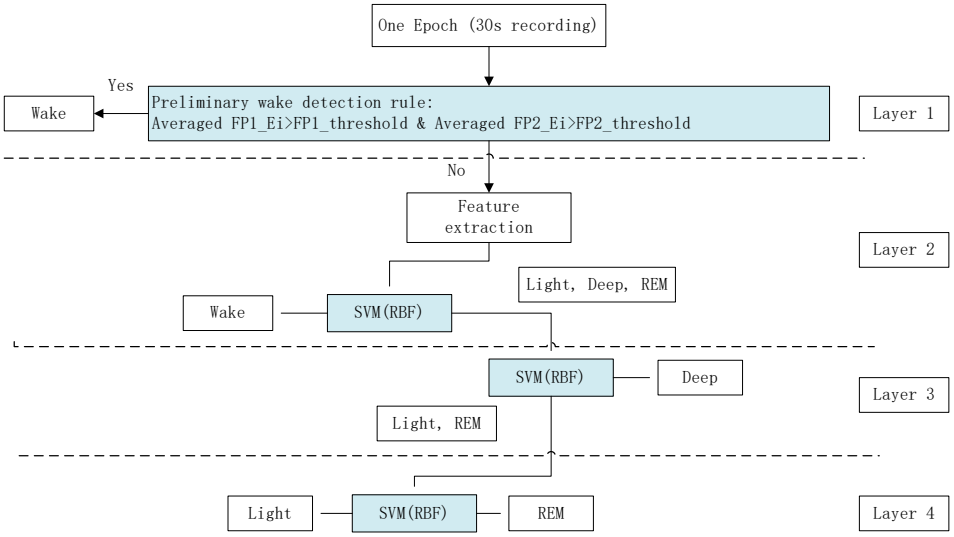

Figure 3.5: The structure of the hierarchical classification proposed in [7]

There are totally 4 layers including one preliminary Awake detection and 3 other layers to classify each sleep stage. The *Ei* in the rule of the first layer represents the *i*th 30s epoch.

There are totally 4 layers including one preliminary Awake detection and 3 other layers to classify each sleep stage (Awake, deep, light and REM). The preliminary Awake detection compared the power in frequency domain with thresholds. In other three layers, forward and backward feature selection method was adopted to find the best feature sets with SVM classifiers. The features were extracted from the frequency domain of the EEG signal. After they tested the proposed algorithm on 12 subjects (12 male; mean age 23±4 years), they reported the results as shown in table 3.1.

Table 3.1: Reported performance of sleep stage classification via EEG signals.

Accuracy (%)	**kappa**	**Sensitivity (%)**			
		Wake	**Light**	**Deep**	**REM**
75.36	64.59	84.36	63.76	93.82	75.84

3.1.1.3.Data Collection

Only one healthy subject (female; age 23) participated in this experiment. The subject slept with both bed sensor and EEG band for a whole night. Eight nights worth of data were

collected (mean sleep time: 5.80±1.17 hours). Both Night 4 and Night 5 have two segments with a gap in between.

During the eight nights' data collection, two storage systems were used to save signals from the bed sensor. One of them was the ground truth system which saved the bed sensor signals to a PC; the other one was an embedded system where the bed sensor signals were saved to an SD card. Since the bed sensor and EEG device were two standalone systems, it was important to synchronize their time in order to bring the ground truth into correspondence with the BCG signal. The time of the EEG device was determined by the Android tablet. For the ground truth system, it was easy to adjust time on the PC to the same as that on the tablet. The procedure of data collection with the ground truth system is listed below:

1) Check transducers' position and connect cables.
2) Synchronize time on PC and Android tablet.
3) Wear EEG band and test its configuration. Open the Android software. Look at EEG waves showed on screen and blink eyes. If blink pattern is shown in waves, consider EEG system is set up correctly. Close software.
4) Start bed sensor by clicking start button on Lab View software on PC.
5) Start EEG device by clicking start button on Android software on tablet.
6) Lie on bed.

For the embedded storage system, the time couldn't be adjust, so a different data collection

procedure was applied to synchronize the time. The procedure of data collection with the embedded system is:

1) Check transducers' position and connect cables.
2) Wear EEG band and test its configuration. Open the Android software. Look at EEG waves showed on screen and blink eyes. If blink pattern is showed in waves, consider EEG system is set up correctly. Close software.
3) Start embedded system by pushing button on device.
4) Start data collection on both devices by sitting on the bed and clicking start button on Android software together.
5) Lie on bed.

Note that, for embedded system, recording on the SD card only starts when a certain amount of weight is added on the bed sensor (sit on bed in the step 4). So the times were synchronized by starting the bed sensor and EEG device together. An example below (Figure 3.6) shows the hypnogram given by the EEG sleep detection system.

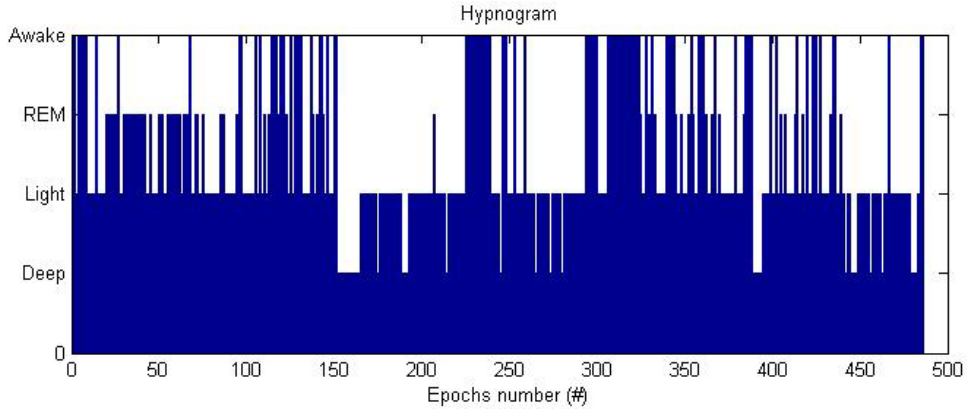

Figure 3.6: One night's hypnogram given by the EEG sleep detection system

To be note that, the abscissa axis here is "Epochs number" instead of "hours" and each epoch is 30s. Following hypnogram plots in this thesis will all use "Epochs number" as abscissa axis.

3.1.1.4.Data Structure of the bed sensor dataset

In this work, the objective was to detect Awake, REM and NREM, so the four stages from the EEG device were further combined to three classes: Awake, REM and NREM (light and deep). Table 3.2 shows number of epochs of each sleep stages for each night.

Table 3.2: Number of epochs of three sleep stages: Awake, REM and NREM in the bed sensor dataset

	Awake	**REM**	**NREM**	**Total**	**Awake(%)**	**REM(%)**	**NREM(%)**
Night1	104	67	315	486	21.40%	13.79%	64.81%
Night2	138	203	230	571	24.17%	35.55%	40.28%
Night3	206	249	127	582	35.40%	42.78%	21.82%
Night4a	118	179	196	493	23.94%	36.31%	39.76%
Night4b	117	49	176	342	34.21%	14.33%	51.46%
Night5a	410	62	160	632	64.87%	9.81%	25.32%
Night5a	91	61	114	266	34.21%	22.93%	42.86%
Night6	327	378	55	760	43.03%	49.74%	7.24%
Night7	424	141	174	739	57.37%	19.08%	23.55%
Night8	57	41	600	698	8.17%	5.87%	85.96%
Total	1992	1430	2147	5569	35.77%	25.68%	38.55%

3.1.2. MIT-BIH Polysomnographic Database (MIPBPD)

3.1.2.1.Data Set Description of the MITBPD

The MIT-BIH Polysomnographic Database is a collection of multiple physiologic signals during sleep. The database contains over 80 hours' worth of four-, six-, and seven-channel polysomnographic recordings, each with an ECG signal annotated beat-by-beat. The 18 PSG records were collected from 16 male subjects with or without apnea syndrome. The mean age of the subjects was 40 (range:32-56) [33]. Each record contains the raw data, the header files, the QRS annotation files and the sleep stage annotation files. The raw files contain the original PSG signals with a 250 Hz sampling rate. All recordings include an

ECG signal which was the only signal used in this project. The QRS annotation files give the location of detected QRS wave (heart beat) in the ECG signal and the sleep stage annotation files give the sleep stages for each 30 second epoch according to the criteria of Rechtschaffen and Kales (R&K). Figure 3.7 and Figure 3.8 show plots of 30 seconds' ECG signal with the detected QRS waves and a plot of one night's hypnogram, respectively.

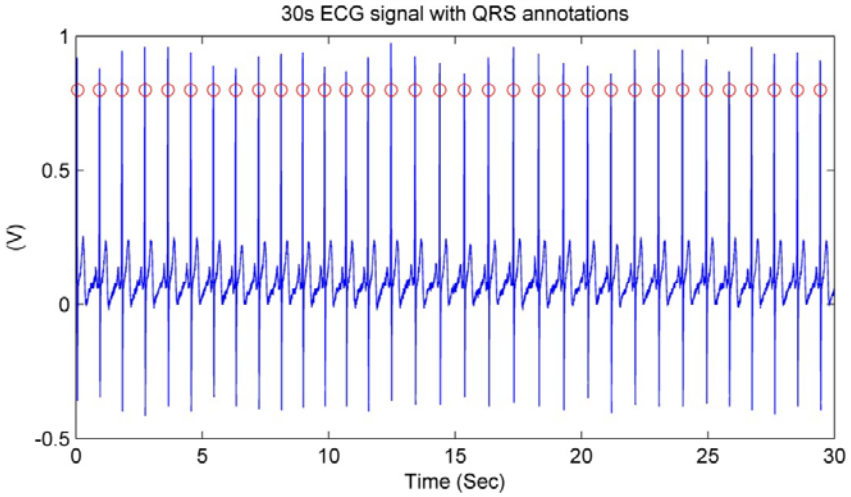

Figure 3.7: A 30s ECG signal with detected QRS waves

The red circles indicate the positions of detected QRS waves given by the QRS annotation file.

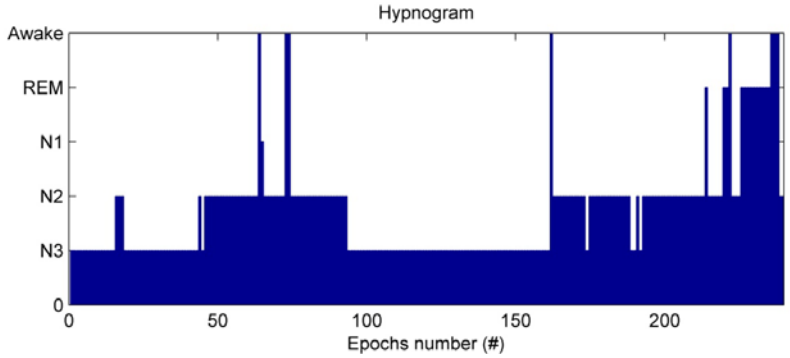

Figure 3.8: One night's hypnogram given by the sleep stage annotation file.

3.1.2.2.Data Structure of the MITBPD

The MITBPD used R&K standard to divide sleep stages. They are: Awake, REM, N1, N2, N3, N4 and MT (motion time). These stages were further combined to Awake (Awake and

MT), REM and NREM (N1, N2, N3 and N4). In some of the 30s epochs, no QRS annotations were given; these epochs were removed from the recordings. Table 3.3 shows number of epochs of each sleep stage for each record.

Table 3.3: Number of epochs of three sleep stages: Awake, REM and NREM in the MITBPD

	Awake	REM	NREM	Total	Awake(%)	REM(%)	NREM(%)
Slp01a	8	13	219	240	3.33%	5.42%	91.25%
Slp01b	180	25	155	360	50.00%	6.94%	43.06%
Slp02a	52	77	231	360	14.44%	21.39%	64.17%
Slp02b	108	29	133	270	40.00%	10.74%	49.26%
Slp03	133	74	495	702	18.95%	10.54%	70.51%
Slp04	162	23	535	720	22.50%	3.19%	74.31%
Slp14	322	36	356	714	45.10%	5.04%	49.86%
Slp16	316	65	313	694	45.53%	9.37%	45.10%
Slp32	394	0	246	640	61.56%	0.00%	38.44%
Slp37	75	11	612	698	10.74%	1.58%	87.68%
Slp41	229	90	461	780	29.36%	11.54%	59.10%
Slp45	119	81	556	756	15.74%	10.71%	73.54%
Slp48	214	31	515	760	28.16%	4.08%	67.76%
Slp59	140	35	283	458	30.57%	7.64%	61.79%
Slp60	286	31	384	701	40.80%	4.42%	54.78%
Slp61	124	79	517	720	17.22%	10.97%	71.81%
Slp66	175	0	264	439	39.86%	0.00%	60.14%
Slp67x	72	0	82	154	46.75%	0.00%	53.25%
Total	3109	700	6357	10166	30.58%	6.89%	62.53%

3.1.3. The Sleep-EDF Database (Expanded)

3.1.3.1.Data Set Description of the sleep-EDF database

The Sleep-EDF Database (Expanded) consists of two studies [33]. One of the studies was used in this project. The original collection of this study came from 79 healthy Caucasians aged 25-101, without any sleep-related medication [34]. Among them, the subjects currently provided in the database are 10 males and 10 females aged 25-34. Each subject

was recorded on two subsequent day-night periods. Subjects wore a modified Walkman-like cassette tape recorder described in [35] for about 20 hours in their homes. Several signals were recorded including a respiration signal. The respiration signal was obtained from an oral-nasal respiration air flow [35]. The oral-nasal airflow signal was then sampled at 1 Hz (Figure 3.9).

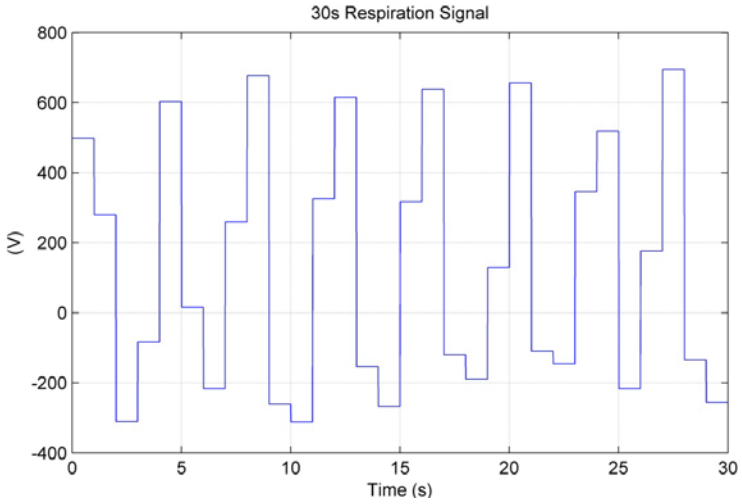

Figure 3.9: A 30s respiration signal.

The sleep stages in each 30s epoch were given by annotation files according to the R&K standard. Because each 20 hour recording contains both daily living activities and sleep, data during the sleep time need to be extracted. Epochs from 20 minutes before the first non-awake stage to the last non-awake stage were considered as sleep time. Respiratory signals and detected sleep stages in this range were kept and data outside this range were excluded. An example of one night's hypnogram is shown in Figure 3.10.

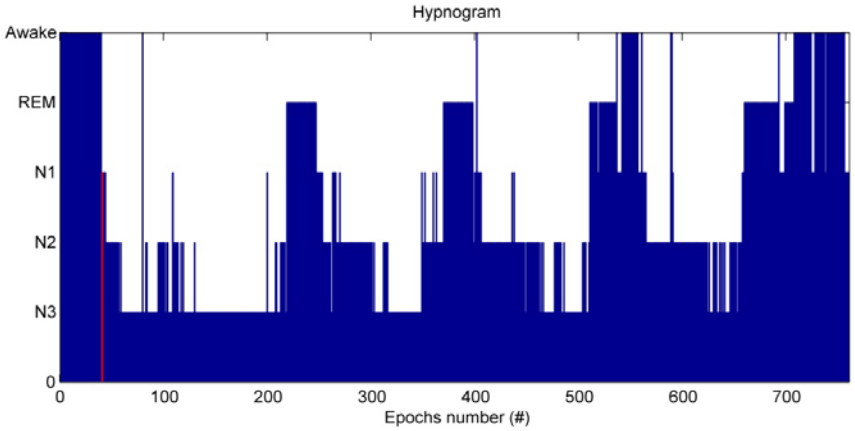

Figure 3.10: One night's hypnogram given by the sleep-EDF

The red epoch in the plot is the first non-awake stage (N1) given by the annotation file. Forty Epochs (20 minutes) before this epoch were kept and considered as sleep time.

3.1.3.2.Data Structure of the sleep-EDF database

The Sleep-EDF Database (Expanded) used the R&K standard to divide sleep stages into 7 categories: Awake, REM, N1, N2, N3,N4 and MT (motion time). These stages were further combined to Awake (Awake and MT), REM and NREM (N1, N2, N3 and N4). A data selection process was implemented on this database. The detailed description of this process will be discussed in section 3.2.3. After data selection, 21 recordings among 40 were selected. Table 3.4 shows number of epochs of each sleep stage for these 21 recordings.

Table 3.4: Number of epochs of three sleep stages: Awake, REM and NREM in the sleep-EDF database

	Awake	REM	NREM	Total	Awake(%)	REM(%)	NREM(%)
SC4001	108	125	528	761	14.19%	16.43%	69.38%
SC4002	103	215	729	1047	9.84%	20.53%	69.63%
SC4011	77	170	776	1023	7.53%	16.62%	75.86%
SC4012	82	176	848	1106	7.41%	15.91%	76.67%
SC4031	60	209	603	872	6.88%	23.97%	69.15%
SC4041	120	196	839	1155	10.39%	16.97%	72.64%
SC4042	105	270	745	1120	9.38%	24.11%	66.52%
SC4061	62	102	599	763	8.13%	13.37%	78.51%
SC4062	113	187	636	936	12.07%	19.98%	67.95%
SC4071	44	198	654	896	4.91%	22.10%	72.99%
SC4101	75	207	742	1024	7.32%	20.21%	72.46%
SC4102	64	199	749	1012	6.32%	19.66%	74.01%
SC4121	96	258	618	972	9.88%	26.54%	63.58%
SC4122	210	199	488	897	23.41%	22.19%	54.40%
SC4131	75	172	701	948	7.91%	18.14%	73.95%
SC4141	107	233	584	924	11.58%	25.22%	63.20%
SC4142	101	213	558	872	11.58%	24.43%	63.99%
SC4151	92	208	572	872	10.55%	23.85%	65.60%
SC4161	136	260	668	1064	12.78%	24.44%	62.78%
SC4162	99	195	629	923	10.73%	21.13%	68.15%
SC4181	58	118	708	884	6.56%	13.35%	80.09%
Total	1987	4110	13974	20071	9.90%	20.48%	69.62%

3.2. Pre-processing

In this section, pre-processing steps for three databases are explained separately. The main goal was to get the raw signals ready for feature extraction. Each of the three database had its own process according to the quality and properties of the signals.

3.2.1. Bed Senor Dataset

3.2.1.1. Ground Truth Filtering

The ground truth for this dataset was the EEG sleep detection system. In order to make the errors of ground truth have the least influence on this work, several steps were implemented to filter the sleep stages or delete the unreliable sleep stages of the ground truth. The process was aimed at keeping the ground truth as accurate as possible. Three rules were set according to common sense or theories of the sleep cycle.

1) For every three 30s epochs, if the middle one is different from the other two and the other two are the same, change the middle one to the same stage as the other two epochs. But the middle epoch doesn't change if it is an Awake stage.
2) For every three 30s epochs, if the sleep stages are all different, the middle epoch is removed from the recording. This rule also doesn't apply to Awake stage.
3) If REM stages appear in the first 60 minutes, these stages are removed from the recording.

The first two rules were set because a person usually sleep in a certain stage for at least a few minutes. So if the middle epoch of three consecutive epochs is different from other two, it is more likely that this epoch should belong to the same stage as other two. However, if three consecutive epochs all belong to different stages, it is hard to determine which stage the middle one actually belongs to, so this epoch should be deleted. But these two rules don't apply to the Awake stage because a person can wake up from any sleep stage and keep waking for any duration. The third rule was set because the first REM stage usually appears 70 minutes after sleep onset. So any given REM stages in the first 60 minutes is

possible to be misclassified and should be deleted. Figure 3.11 displays the first 90 minutes (180 epochs) of one recording's original ground truth and the processed ground truth.

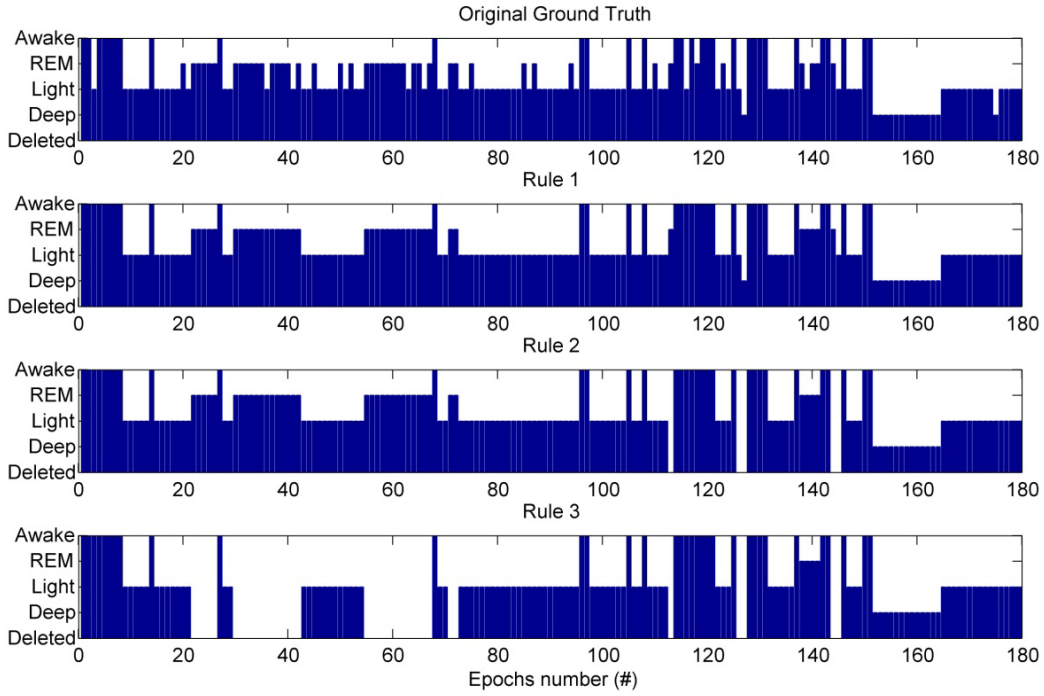

Figure 3.11: The ground truth filtering process

From top to bottom are the original ground truth and the processed ground truth after appling each rule.

The original ground truth shows frequent changes among Awake, REM and light sleep. After applied rule 1, these epochs were smoothed. For example, scattered REM epochs around epoch 85 were filtered to light sleep. After applied rule 2, five epochs were deleted such as epoch around 115. The other two epochs around it were light and Awake, but the epoch itself was REM. So this epoch was deleted because the consecutive three epochs belonged to three different stages. The last plot shows the REM stages before epoch 120 were deleted such as the epochs between 20-60.

3.2.1.2.BCG Signal Filtering

The hardware filtered BCG signal contains information of heart rate, respiration and

restlessness. In order to separate these waveforms, the hardware filtered BCG signals were input to a filtering process. Let's call the hardware filtered BCG signal BCG_f. It was then processed as follow:

1) A 6th-order bandpass Butterworth filter with a lower cutoff frequency of 0.7 Hz and a higher cutoff frequency of 10 Hz was applied on the BCG_f. The obtained signal BCG_{hr} has heart rate information.
2) A 6th-order lowpass Butterworth filter with a cutoff frequency of 0.7 Hz was applied on the BCG_f. The obtained signal BCG_r has respiration information.

The cutoff frequencies are chosen based on the fact that a typical respiratory rate is below 0.5 Hz and the frequency content of BCG is not higher than 10 Hz [36]. Figure 3.12 shows an example of BCG_f, BCG_{hr} and BCG_r in a 30s epoch.

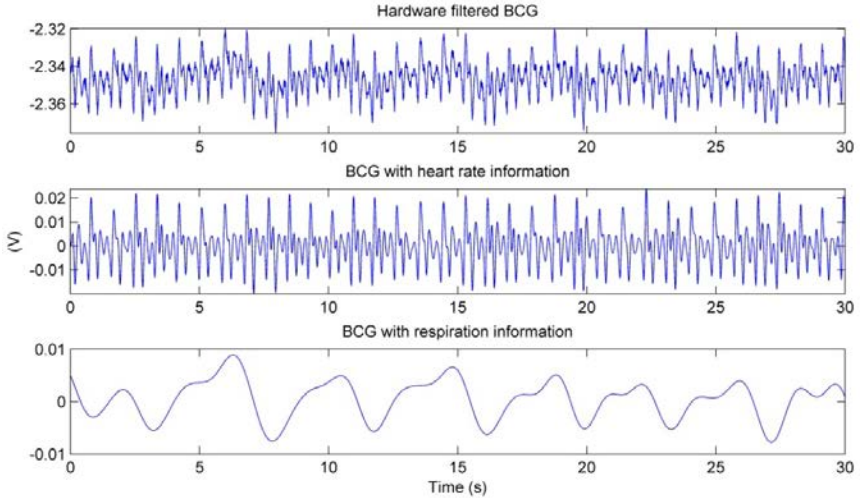

Figure 3.12: A 30s epoch orignal and filtered BCG signal
From top to bottom are BCG_f, BCG_{hr} and BCG_r.

3.2.1.3.Heart Beat Interval Calculation

Before calculating HRV parameters, heart beat intervals (HBI) should be extracted from

BCG_{hr}. An existing algorithm in [8] was used to detect heart beats. The algorithm detected heart beats within each 30 second window (3000 samples for 100Hz signal). Two examples of the results of detection are shown in Figure 3.13. One of the example has 100% accuracy while the other example has several misclassified beats.

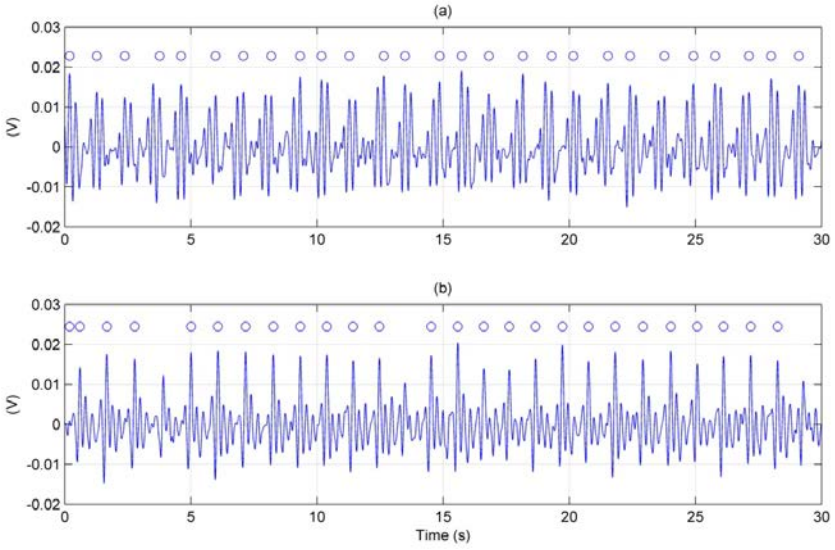

Figure 3.13: Two examples of heart beats detection results
(a) 27 beats among 27 beats were detected correctly; (b) Several beats were misclassified including three false negative beats (4, 13 and 28) and 1 false positive beat before the 1st beat.

Let's assume the time stamp of detected heart beat is x(n), where n is a serial number of beats. Then its corresponding heart beat interval is:

$$I(n)=x(n+1)-x(n) \qquad n=1, 2, 3...$$

Figure 3.14 shows two plots of beat-to-beat intervals corresponding to Figure 3.13.

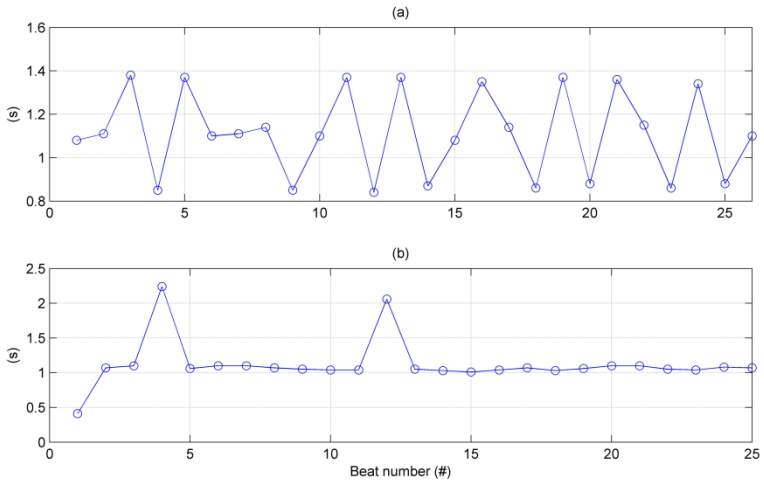

Figure 3.14: Two examples of beat-to-beat intervals corresponding to Figure 3.13
(a) 26 interval values derived from 27 detected heart beats; (b) 25 interval values derived from 26 detected heart beats.

In figure 3.14(b), there are three abnormal intervals. One of them is lower than 0.5 sec and the other two are higher than 2 sec. These intervals are caused by the first three misclassified beats in Figure 3.13(b). So, the misclassified beats will lead to larger or shorter intervals compared with normal HBI and such errors will directly impact the reliability of HRV parameters. In order to remove these abnormal beat-to-beat intervals, the proposed algorithm in [8] ran a post-processing to remove unreliable intervals. Figure 3.15 shows the HBI series of Figure 3.14(b) after post-processing.

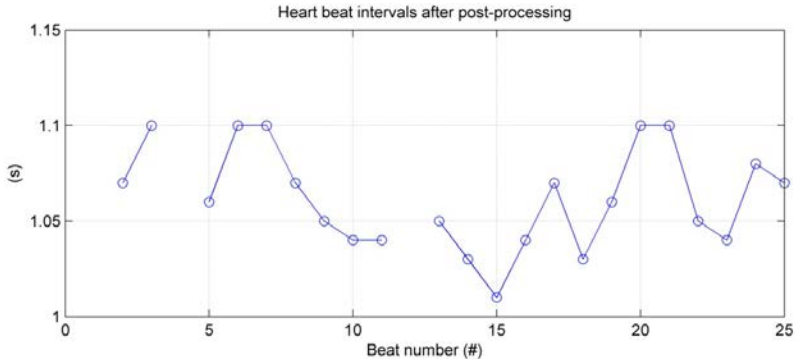

Figure 3.15: The output heart beat intervals from the algorithm
Because the three abnormal intervals were removed, the 22 output intervals are separated into 3 blocks.

Because the unreliable intervals were removed, the output intervals from the algorithm may be discontinuous in each epoch. In Figure 3.15, the outputs were separated to three blocks and intervals in each block are successive. Hence, totally 22 beat-to-beat intervals were calculated in this 30 seconds' epoch.

Based on beat-to-beat intervals, differences of successive intervals are defined as:

$$D(n)=I(n+1)-I(n) \quad n=1, 2, 3...$$

Since only intervals in one block are successive, D(n) is calculated in each block. For example, Figure 3.15 has three blocks. So, we calculate differences of interval values in each block and the values in all three blocks are used as a representative of this 30s epoch. Figure 3.16 shows examples of successive interval differences.

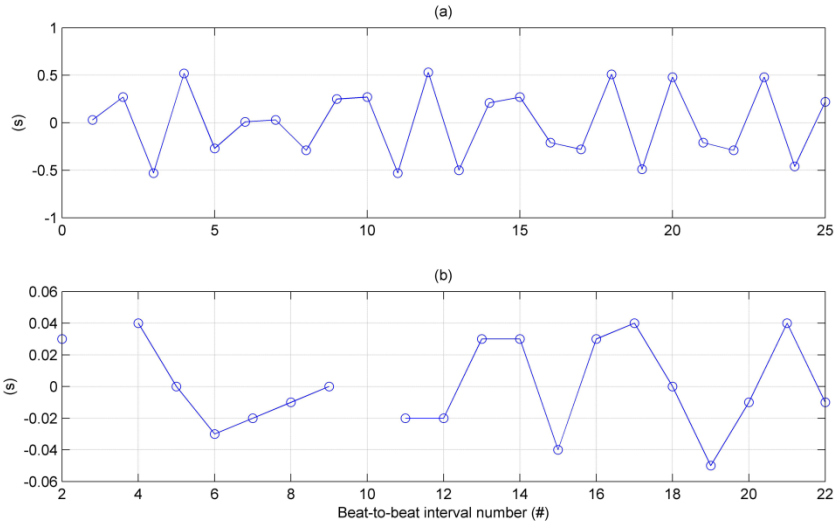

Figure 3.16: Two examples of successive interval differences corresponding to Figure 3.14 (a) and 3.15, respectively. (a) 25 differences of intervals continuously; (b) 19 differences of intervals in 3 blocks.

3.2.1.4.Respiration Cycle Interval Calculation

After obtaining the respiration signal, BCG_r, the peak and trough points can be easily

detected. Figure 3.17 shows one example of detected respiration cycles.

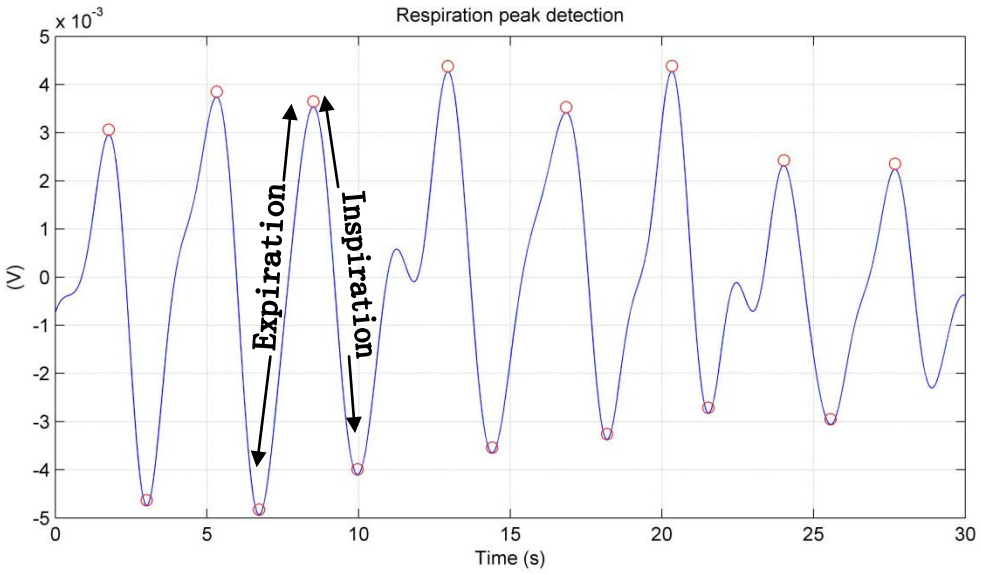

Figure 3.17: A 30s BCG and detected peaks and troughs labeled with red circles
Expiration is defined as trough to peak and inspiration is defined as peak to trough.

Similar to HRV, RV parameters were also calculated based on breath-to-breath intervals. Again, let's assume location of peaks are $x_p(n)$ and troughs are $x_t(n)$, where n is the series number. Then the breath-to-breath intervals are represented by peak-to-peak intervals:

$$I(n)=x_p(n+1)-x_p(n) \quad n=1, 2, 3...$$

The differences of breath intervals are then defined as:

$$D(n)=I(n+1)-I(n) \quad n=1, 2, 3...$$

Besides, expiration and inspiration activity are defined as trough to peak and peak to trough, respectively (see figure 3.17). The expiration intervals and inspiration intervals are defined as:

1) if the first breath process in this epoch is inspiration:

$$I_{Ex}(n)=x_p(n+1)-x_t(n) \quad n=1, 2, 3...$$

$$I_{In}(n)=x_t(n)-x_p(n) \quad n=1, 2, 3...$$

2) if the first breath process in this epoch is expiration:

$$I_{Ex}(n)=x_p(n)-x_t(n) \quad n=1, 2, 3...$$

$$I_{In}(n)=x_t(n+1)-x_p(n) \quad n=1, 2, 3...$$

Moreover, the amplitude of respiration is also a useful parameter. The amplitude of peaks and troughs are defined as $A_p(n)$ and $A_t(n)$, respectively. The differences of amplitude are defined as:

1) if the first breath process in this epoch is inspiration:

$$DA_{Ex}(n)=A_p(n+1)-A_t(n)$$

$$DA_{In}(n)=A_p(n)-A_t(n)$$

2) if the first breath process in this epoch is expiration:

$$DA_{Ex}(n)=A_p(n)-A_t(n)$$

$$DA_{In}(n)=A_p(n)-A_t(n+1)$$

where DA_{Ex} and DA_{In} are differences of amplitude of expiration and inspiration, respectively.

3.2.1.5.Epoch Removal

After ground truth filtering process, some of the epochs were removed from the recordings. Then the heart beat and respiration cycle detection algorithms were applied on the BCG signal. Because of the quality of the BCG signals and the implementation of the detection algorithms, in some of the 30s epochs, no heart beat or respiration cycle was detected.

These epochs were also removed from the recordings and didn't participate in further processes. After ground truth filtering and epoch removal, the number of epochs of each sleep stage of each night is shown in Table 3.5.

Table 3.5: Number of epochs of three sleep stages: Awake, REM and NREM in the bedsensor dataset after epochs removal

	Awake	REM	NREM	Total	Awake(%)	REM(%)	NREM(%)
Night1	78	7	262	347	22.48%	2.02%	75.50%
Night2	115	106	125	346	33.24%	30.64%	36.13%
Night3	100	111	34	245	40.82%	45.31%	13.88%
Night4a	94	81	122	297	31.65%	27.27%	41.08%
Night4b	75	16	127	218	34.40%	7.34%	58.26%
Night5a	231	21	91	343	67.35%	6.12%	26.53%
Night5b	40	11	68	119	33.61%	9.24%	57.14%
Night6	173	154	16	343	50.44%	44.90%	4.66%
Night7	312	80	130	522	59.77%	15.33%	24.90%
Night8	41	16	426	483	8.49%	3.31%	88.20%
Total	1259	603	1401	3263	38.58%	18.48%	42.94%

3.2.2. MIT-BIH Polysomnographic Database

3.2.2.1.Signal filtering

The ECG signals in the MITBPD need to be further filtered to get rid of noise. According to [13], the raw ECG signals had already been digitized at a sampling interval of 250 Hz with 12 bits/sample. In order to get a monitor-quality ECG signal, a band pass filter with lower cutoff frequency of 0.05 Hz and higher cutoff frequency of 40 Hz was applied to the signal [37]. Note that a lower cutoff frequency of 0.05 Hz was used here instead of 0.5 Hz as recommended in [37]. This was done to keep the respiration information in the lower frequency domain. Figure 3.18 shows an example of a 2s raw ECG signal and its filtered signal.

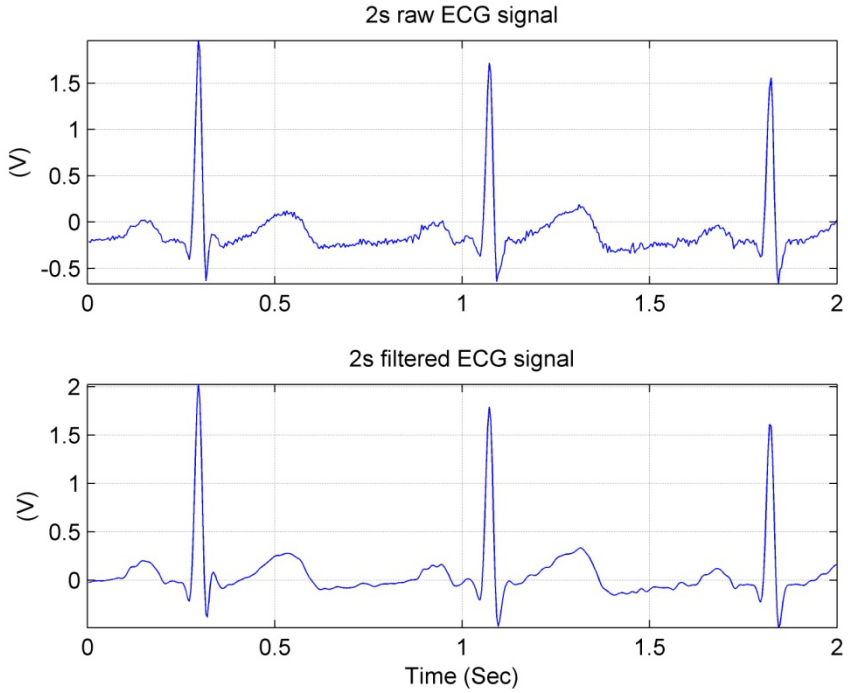

Figure 3.18: A 2s raw ECG signal and filtered ECG signal
The high frequency noise in the raw signal was removed.

3.2.2.2.Heart beat interval calculation

The QRS annotation files of the database provides the locations of heart beats. The same definition was used to define heart beat x(n), heart beat interval I(n), and differences of successive intervals D(n).

Because the QRS annotations have a high accuracy, only a simple step of pre-processing was used to remove the errors of beat detection. In some of the situations, there were two detected points near a single real heart beat, so an extra small interval value would be calculated. The pre-processing found these points and deleted one of them. Figure 3.19 shows an example.

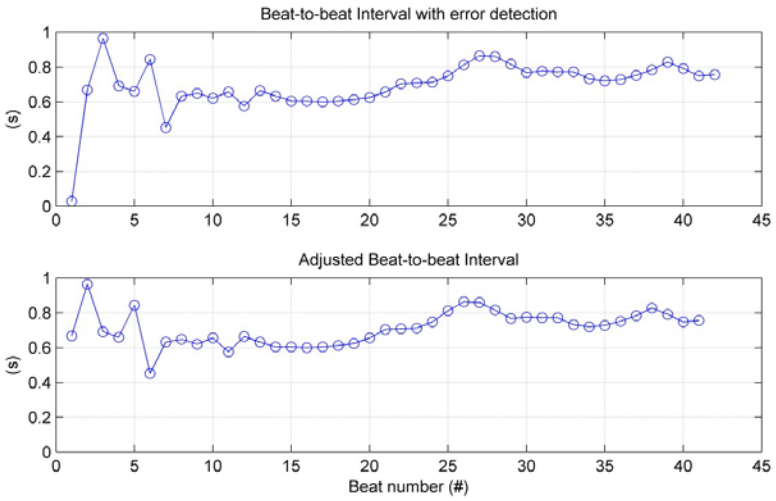

Figure 3.19: Beat-to-beat intervals of one epoch from the MITBPD

The top figure shows the first interval value is very small (nearly 0) which means two close points were both labeled as heart beats. The bottom figure shows the removal of this interval.

3.2.3. Sleep-EDF Database (Expanded)

3.2.3.1.Data selection

As mentioned in the database description section, the respiration signals in the Sleep-EDF database were resampled at 1 Hz and the sampling frequency of the provided signals is 100 Hz. In order to calculate RV parameters, the peaks $x_p(n)$ and troughs $x_t(n)$ were first detected by finding local maximums and minimums (Figure 3.20).

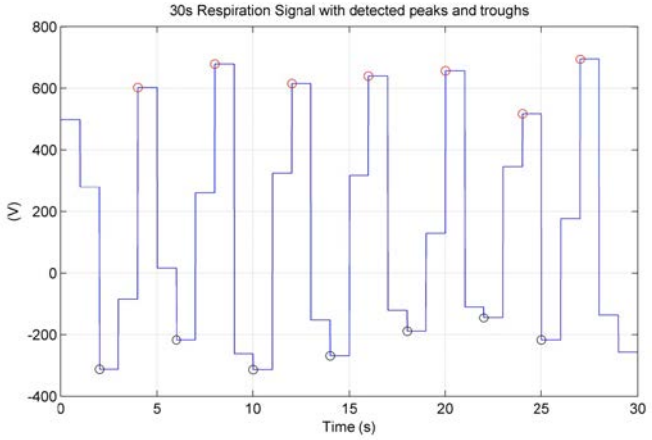

Figure 3.20: A 30s respiration signal with detected peaks and troughs labeled with red and black circles

Then the breath-to-breath interval I(n) was calculated using the same definitions as in section 3.2.1.4. Because the signal was resampled from 100Hz to 1 Hz, the values of the calculated breath-to-breath intervals were integers bigger than 2 seconds. According to [29], a healthy adult breathes 12 - 15 times per minute at rest. That means most of the breath-to-breath intervals should be in the range of 4 - 5 seconds. Figure 3.21 displays the histogram of breath-to-breath intervals of one recording.

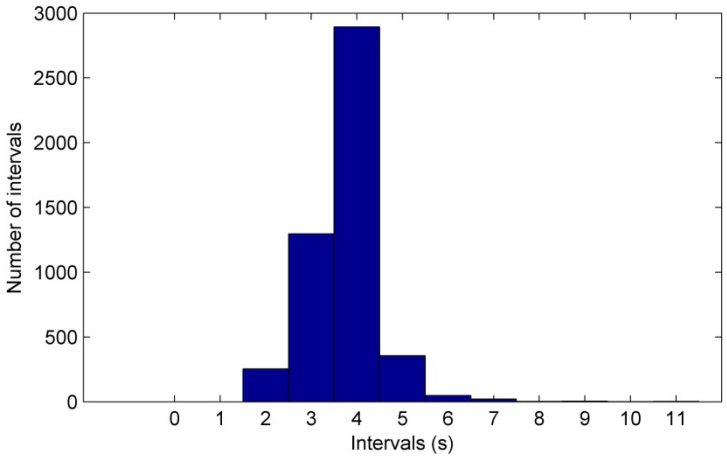

Figure 3.21: Histogram of breath-to-breath intervals of one recording

The majority of the intervals are 4 seconds. But about 1/3 of the intervals are 3 seconds (20 breaths per minutes) and some others have lower or higher values.

Although some of the intervals in Figure 3.21 are out of the 4-5 second range, considering the possible errors caused by resampling and physiological differences among different individuals, this distribution is still acceptable. However, the breath-to-breath intervals of some of the recordings gathered around 2s or have a wider distribution. These recordings may be too noisy. Figure 3.22 shows one example.

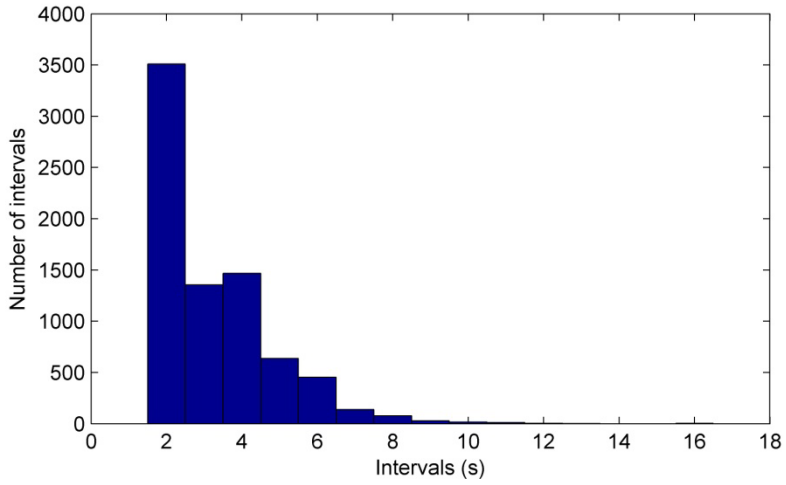

Figure 3.22: Histogram of breath-to-breath intervals of one noisy recording
Most of the intervals are 2 seconds which means the respiratory rate can reach 30 times per minutes.

Because the original respiration signals are unavailable, the cause of such problem couldn't be detected. It could be the body movements, connection with the sensor or even the subject did have such high respiratory rate. In order to keep the whole dataset reliable, a data selection process was implemented to select recordings with respiratory rate in a normal range. The selection rule is: If the proportion of I(n)> 6s or I(n)< 3s is bigger than 10%, then this recording was deleted.

This range selected the recordings where the majority of the respiratory rates were in the range of 10 to 20 times per minute, which allows some variability. After selection, 21 recordings were kept in the dataset.

3.3. Feature Extraction

3.3.1. Heart Rate Variability (HRV) Features

HRV features were calculated on the bed sensor dataset and MITBPD. All these features

were computed based on the beat-to-beat intervals I(n) or the successive differences of beat-to-beat intervals D(n) within a 30 second epoch. In section 2.3, some of the HRV parameters were listed. Besides these parameters recommended by [26] which show variations in heart rate directly, some other simple statistics were also calculated. These statistics reflect the range of I(n) sequence in one epoch. Some of these features are mean of heart beat intervals (mHBI), maximum of heart beat intervals (maxHBI) and minimum of heart beat intervals (minHBI). The list of features and their descriptions are described below:

1) RMSSD: the square root of the mean of the squares of differences between adjacent beat-to-beat intervals D(n).

$$\sqrt{\frac{\sum_{n=1}^{N} D(n)^2}{N}}$$

2) pNN50: the percentage of differences of adjacent intervals D(n)>0.05s.

3) mHBI: mean of heart beat intervals:

$$\frac{\sum_{n=1}^{N} I\,(n)}{N}$$

4) SDSD: Standard deviation of differences between adjacent intervals.

$$\sqrt{\frac{1}{N-1}\sum_{n=1}^{N} |D\,(n)-\mu|^2}$$

where $\mu = \frac{1}{N}\sum_{n=1}^{N} D\,(n)$

5) SDNN: Standard deviation of beat-to-beat intervals.

$$\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}|I(n)-mHBI|^2}$$

6) maxHBI: maximum value of beat-to-beat intervals:

$$\max_{1\le n\le N} I(n)$$

7) minHBI: minimum value of beat-to-beat intervals:

$$\min_{1\le n\le N} I(n)$$

8) max_minHBI: maximum value of beat-to-beat intervals subtract minimum value of beat-to-beat intervals:

$$\max_{1\le n\le N} I(n) - \min_{1\le n\le N} I(n)$$

9) CV: coefficient of variance:

$$\frac{1}{mHBI}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}|I(n)-mHBI|^2}$$

10) Frequency domain features: LF, HF, Total power, LF/HF

In order to calculate frequency domain features, the power spectral density (PSD) of beat-to-beat intervals I(n) needs to be calculated first. Due to the uneven sampling property of I(n), it was first interpolated using a Spline interpolation method [38] with 4Hz sampling rate to get a uniform distribution of values. The obtained sequence was padded with trailing zeros to length 256. Then the PSD was calculated using the Fast Fourier Transform (FFT) based estimation with 256 sampling points. Next, LF, HF and the total power components were then calculated. Their corresponding frequency range are: LF--[0.04 0.15], HF--[0.15

0.4], total power--[0 0.4].

3.3.2. Respiratory Variability (RV) Features

The RV features were calculated on the bed sensor dataset and the Sleep-EDF database. All these features were computed based on the breath-to-breath intervals I(n) or the successive differences of breath-to-breath intervals D(n) within a 30 second epoch. Most of the RV features referred to the statistics applied to heart beat intervals and the features extracted in [12]. The selected features and their descriptions are listed below:

1) mDI: Mean of the differences of intervals:

$$\frac{1}{N}\sum_{n=1}^{N} D\,(n)$$

2) MADI: Max absolute differences of intervals:

$$\max_{n\leq N}(|D\,(n)|)$$

3) Respiratory rate (RR): The respiratory rate is the number of breaths taken in 60 seconds.

$$RR = \frac{60}{N}\sum_{n=1}^{N}\frac{1}{I(n)}$$

4) SDRR: Standard deviation of respiratory rate:

$$\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}|\frac{60}{I(n)} - RR|^2}$$

5) RMSSD: the square root of the mean of the squares of differences between adjacent breath-to-breath intervals.

$$\sqrt{\frac{\sum_{n=1}^{N} D(n)^2}{N}}$$

6) CV: coefficient of Variance

$$\frac{1}{\mu}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}|I(n)-\mu|^2}$$

where $\mu = \frac{1}{N}\sum_{n=1}^{N} I\,(n)$

7) Median of respiratory rate:

$$Q_2(\frac{60}{I(n)})$$

where Q_2 means median.

8) IQR: Inter quartile range of respiratory rate.

$$Q_3(\frac{60}{I(n)}) - Q_1(\frac{60}{I(n)})$$

where Q_3 and Q_1 are the third quartile and the first quartile, respectively.

9) MAD: Mean absolute deviation value of respiratory rate.

$$\frac{1}{N}\sum_{n}|\frac{60}{I(n)} - RR|$$

10) Ratio of mean of differences of amplitude of expiration and inspiration:

$$\frac{\sum_{n=1}^{N} DA_{Ex}(n)}{\sum_{n=1}^{N} DA_{In}(n)}$$

where DA_{Ex} and DA_{In} are differences of amplitude of expiration and inspiration.

11) Ratio of Mean of expiration intervals and inspiration intervals:

$$\frac{\sum_{n=1}^{N} I_{Ex}(n)}{\sum_{n=1}^{N} I_{In}(n)}$$

where I_{Ex} and I_{In} are differences of expiration and inspiration intervals.

3.3.3. Linear Frequency Cepstrum Coefficients (LFCC)

The cepstrum is defined as the inverse Discrete Fourier Transform (DFT) of the logarithm of the power spectrum of a signal. In speech recognition, cepstrum analysis plays an important role. After some experiments, it was found useful here to help detect sleep stages. The LFCC feature was calculated on both the bed sensor dataset (BCG_{hr}) and MITBPD (filtered ECG). It has the same processing steps as the famous MFCC feature [39], the only difference is LFCC uses linear-frequency filter bank and MFCC uses a mel-frequency filter bank. The calculation steps are described below.

1) The first step was to generate the power spectrum. This was completed by first cutting the whole 30s signal to short time frames using a sliding window. The sliding window was 2s long (200 points for BCG_{hr}; 500 points for ECG) with 80% overlap. For each time frame, the sequence was then padded with trailing zeros to length 256 and 512 for BCG_{hr} and ECG, respectively. Next, the DFT was applied with k sampling points (k is 256 for BCG_{hr} and 512 for ECG) in each time frame to obtain the complex spectrum. Finally, the power spectrum was obtained by taking the square of the absolute value of the complex spectrum. Figure 3.23 displays the spectrogram of one 30s epoch of the ECG signal. Figure 3.24 shows the power spectrum of the first time frame of this epoch.

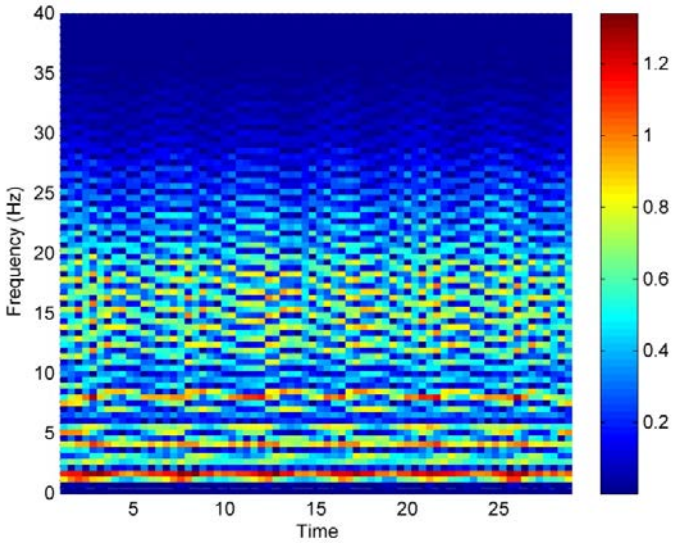

Figure 3.23: An example spectrogram of one 30s epoch of the ECG signal
The amplitude of the power spectrum is represented by color of each point.

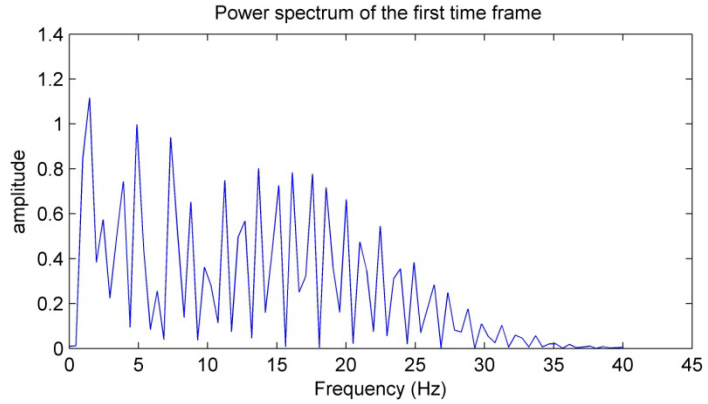

Figure 3.24: Power spectrum of the first time frame of the example epoch in Figure 3.23.

2) The second step was to compute a linear filter bank. The filter bank consists of 26 filters. The BCG_{hr} and ECG signals have frequencies ranging from 0.7 to 10 Hz and 0.05 to 40 Hz, respectively. According to their own frequency ranges, 28 points were generated, linearly spaced between the minimum and maximum frequency. Then these frequencies were rounded to the nearest FFT bins obtained from DFT. Twenty-six triangular filters were built through these points. Figure 3.25 show the filter banks for the two databases.

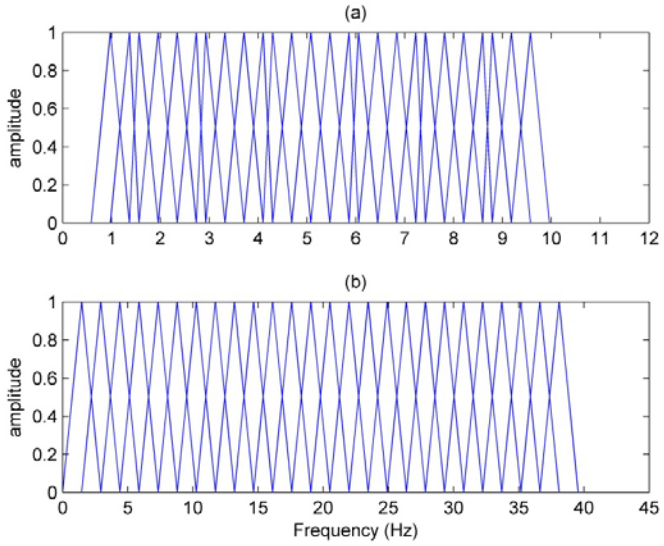

Figure 3.25: Filter banks, each bank contains 26 triangular filters

(a)Filter bank for BCG_{hr}, filters range from 0.7 to 10Hz;(b) Filter bank for ECG, filters range from 0.05 to 40Hz

3) Then, each filter was multiplied with the power spectrum and summed to get 26 bank energies for each time frame.

4) The next step was to get the log bank energies. This was completed by taking the logarithm of each of the 26 bank energies. Figure 3.26 depicts the 26 log bank energies of the example time frame in Figure 3.24.

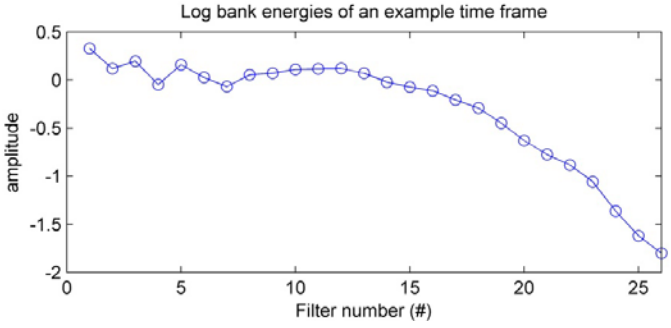

Figure 3.26: Twenty-six log bank energies of one example time frame.

5) The Discrete Cosine transform (DCT) of the 26 log energies was taken to give 26 cepstral coefficients for each time frame. The very first coefficient was discarded

because it was the DC term. So for each time frame, there were 25 coefficients.

6) To combine information in all time frames, for each coefficient, compute its mean value and standard deviation along the time axis. That gave 25 means of LFCC and 25 standard deviations of LFCC. In the end, there were 50 features from LFCC. They were 25 mean (mLFCC) and 25 standard deviation (stdLFCC) values of the coefficients. Figure 3.27 display the obtained features: mLFCC and stdLFCC of the example epoch.

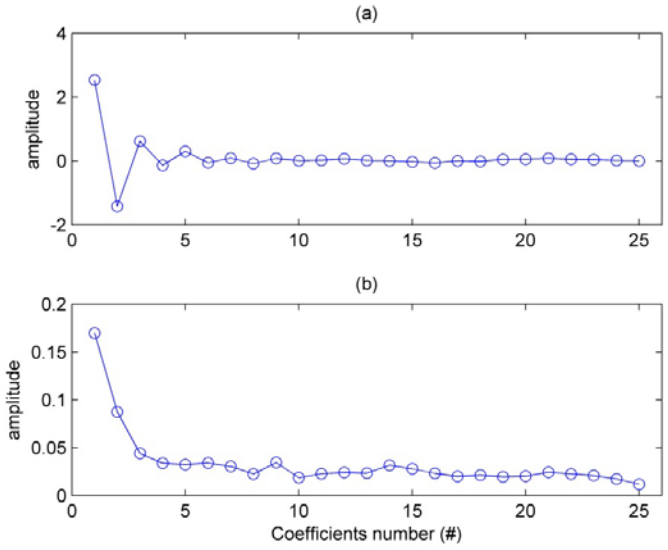

Figure 3.27: LFCC features of the example epoch of ECG

(a) Mean value of each coefficient along time axis (mLFCC); (b) Standard deviation value of each coefficient along time axis (stdLFCC).

3.3.4. Further Processes of Features

Sleep in each night is a time series event. So the features extracted in each epoch are related to their previous and later ones. After extracting features listed above, some further processing was applied on the features. These processes showed an ability to improve the performance of sleep stage recognition. The two processing methods are described below.

3.3.4.1.Feature Smoothing

The features calculated from BCG, ECG and respiration signal reflect the body conditions in a certain 30s epoch. Reference [24] pointed out that a healthy heart rate is not fixed but rather varies in milliseconds in response to moment-to-moment physiological changes; same for the breath cycle. Hence, feature values calculated based on these physiological indexes are different in each epoch. This kind of normal variation is not the variation we are looking for. They are not caused by the change of sleep stages. Therefore, feature smoothing was applied to help remove these variations and get the trend of the useful values.

The other reason to apply feature smoothing is because epochs of the same sleep stages usually connect together as a block and then transit to another block of stages. They don't transit very frequently except for the Awake stage. Thus, feature smoothing can help find a block of places that may belong to the same stage.

The smoothing methods used in the work were Robust Locally Weighted Scatterplot Smoothing (RLOWESS) [40]. The span of the smoothing was varied in different experiments.

3.3.4.2.Feature Detrending

Different from feature smoothing, feature detrending was used to remove the trend of the features and to detect spikes. It is useful for Awake detection because of the sudden change of physiological indexes when someone wakes up in the middle of the night. Different

from linear trends removal, the detrending process here is aimed at removing nonlinear trends. The detrended features were obtained by subtracting the smoothed features from the original ones. The smoothed features were obtained by applying RLOWESS with a 30 point sliding window.

3.4. Weighted Support Vector Machine (wSVM)

A Support Vector Machine (SVM) was used in this research because it was employed in many of previous sleep studies and showed good performance. Besides, comparing with other classification methods such as: K-nearest neighbors and Naive Bayes, it also had better results with the three databases in this work.

The idea of the SVM is to find a separating hyperplane to maximize the margin between two classes. Given a set of *l* data points

$$\{(x_i, y_i), i = 1,2, \dots, l\}, x_i \in R^N, y_i \in \{-1, +1\},$$

suppose that all data points satisfy the following constraints (separable case):

$$w^T x_i + b \geq 1 \quad , y_i = 1$$

$$w^T x_i + b \leq -1 \quad , y_i = -1$$

The distance between these two hyperplanes, or so called the margin is $\frac{2}{\|w\|}$. To maximize the margin, the problem can be written as:

$$\min \frac{1}{2}\|w\|^2 \qquad s.t., y_i(w^T x_i + b) \geq 1$$

This is a quadratic programming optimization problem, and can be expressed as:

$$\max_a \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j x_i{}^T x_j$$

$$s.t., a_i \geq 0, i = 1, \ldots, l$$

$$\sum_{i=1}^{l} a_i y_i = 0$$

The hyper-plane is obtained by solving above optimization problem. However, this is the separable case, for the non-separable case, positive slack variables, ξ_i, i=1,...,l and a penalty C was introduced. In order to maximize the margin, the optimization problem now becomes:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i \quad s.t., y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

Again, it is equivalent to solving:

$$\max_a \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j x_i{}^T x_j$$

$$s.t., C \geq a_i \geq 0, i = 1, \ldots, l$$

$$\sum_{i=1}^{l} a_i y_i = 0$$

A kernel function $K(x_i, x_j)=x_i{}^T x_j$ is usually used to solve non-linear problem by mapping inputs into high dimensional feature spaces. In this work, the Radial Basis Function (RBF) was used as the kernel. The RBF kernel defines as:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|}, \gamma > 0$$

Like other classifiers, SVM may bias the results when the training data are imbalanced. In the sleep datasets, the proportions are different in each sleep stage. So the class weighting

scheme was used in the SVM classifier.

Class weighting was accomplished by assigning different penalty *Cs* to the two classes based on the ratio of number of samples in each class. Considering the penalty for the positive class is C_1, then the penalty C_2 for the negative class is:

$$C_2 = \frac{N_1}{N_2} C_1$$

where *N1* and *N2* are number of samples in class 1 and class 2, respectively. The SVM classifier and the class weighting scheme was implemented by libsvm library[41].

3.5. Threshold Comparison Classifier

The basic theory of the sleep stage recognition problem is higher variability of cardiac and respiratory activities in REM and Awake while much stable physiological indexes in NREM sleep. So it was easy to think of setting thresholds on some of the features and trying to find out if such thresholds can separate different sleep stages. Thurs, this method was tested on the MITBPD and Sleep-EDF database. The bed sensor database was not included because of its unreliable ground truth and the as yet unclear relation between features and sleep stages.

The advantages of this method are discussed as follows: 1) Implementation of the method reflects the meaning of the feature directly. So the features and thresholds can be selected and adjusted according to theory or common knowledge. 2) Each night was treated as a complete and independent sample. So the classifier would not be influenced by the

differences of the physiological indexes among different subjects or the changes of these indexes for the same subject in different nights. For example, the high or low values of heart rate and respiratory rate are relative values. A reasonable method would be to compare them within one night or even a shorter time period to define the ranges of high or low values instead of defining them using all subjects or all nights of the same subject.

The detailed implementation and experiments will be discussed in section 4.

3.6. Performance Evaluation

3.6.1. Performance Measurements

It was hard to evaluate the performance of the sleep stage classification problem. Because not only does the problem have imbalanced classes, but also the proportion of each class is different each night. As mentioned in section 2.5, the most common measurements used in the literature for this problem were accuracy and Cohen's kappa coefficient. Accuracy is calculated in nearly all classification problems and it expresses the proportion of samples that were classified to the right classes. But it is not sufficient in imbalanced problems because one can simply classify all samples to the majority class and still get a high accuracy. Thus, the kappa coefficient was used due to its lower sensitivity to an imbalanced dataset. Besides accuracy and kappa, other measurements calculated to evaluate the performance were: sensitivity, specificity and precision. The formulas of these measurements are described as follows.

First, the confusion matrix of the results is defined as:

Table 3.6: Confusion matrix

		True Condition	
		Class 1	**Class2**
Predicted Condition	**Class1**	True Positive (TP)	False Positive (FP)
	Class 2	False Negative (FN)	True Negative (TN)

The sensitivity is:

$$Se=TP/(TP+FN)$$

The specificity is:

$$Sp=TN/(FP+TN)$$

The precision is:

$$Pr=TP/(TP+FP)$$

The accuracy is:

$$Acc=(TP+TN)/(P+N)$$

The kappa value is:

$$k = \frac{p_o - p_e}{1 - p_e}$$

$$p_o = ACC,\ p_e = \frac{TP + FP}{P + N} \cdot \frac{TP + FN}{P + N} + \frac{TN + FN}{P + N} \cdot \frac{TN + FP}{P + N}$$

where P_o is the proportion of observed agreement and P_e is the proportion of agreement expected by chance.

Sensitivity measures the proportion of positive samples that are detected while specificity

measures the proportion of negative samples that are detected. Precision expresses the proportion of detected positive samples that are really positives. It will also be influenced by the imbalance of the dataset. The benefit of sensitivity and specificity is that they are not affected by the sample numbers of each class. The goal of a good classifier is to make both sensitivity and specificity as big as possible.

For the three sleep quality measures: sleep efficiency (SE), sleep onset (SL) and percentage of REM stages (%SR), the absolute errors were calculated for evaluation. Here,

$$E_{se} = |SE\text{-}SE'|$$

$$E_{sl} = |SL\text{-}SL'|$$

$$E_{\%sr} = |\%SR - \%SR'|$$

where SE', SL' and %SR' are the estimated results; SE, SL and %SR are the actual values. In section 2.1.2, the criteria for SL calculation was set. The criteria involves the N1 stages. In this work, the subclasses of NREM stages (N1~N3) were not separated, so N1 was not obtained. Thus, the criteria for estimated SL was modified and defined as the first detected sleep stage (REM and NREM) that continues for at least 3 minutes.

3.6.2. Validation and Model Selection

Cross-validation was used to validate the classifier. Two strategies were engaged to separate the training and validation parts. The first strategy was to put all recordings together and randomly pick training and validation sets. The same proportion of each class was kept. The second strategy was leave-n-nights-out. The training and validation sets

were chosen night-by-night. It means n nights' recordings were used to test and the remaining were used as the training set.

The SVM classifier with RBF kernel had two free parameters that need to be selected. They were the penalty C and γ for the RBF kernel. The grid search method was employed to complete this mission. Grid search is a simple method that searches for the optimal pair of parameters through the hyperparameter space of a learning algorithm. It is an exhaustive and expensive method, but it can usually avoid falling into local optimum. The grid search method was measured by cross-validation on the training set. The parameters with the best cross-validation measurement were picked. The measurement here was the kappa coefficient. Cross-validation applied on the training set used the same two strategies as the cross-validation of the whole dataset.

So the whole validation and model selection process had two cross-validation stages. The first layer of cross-validation was applied on the whole dataset, then in each iteration, the training set was applied with the second layer of cross-validation to select parameters with grid search (Figure 3.28).

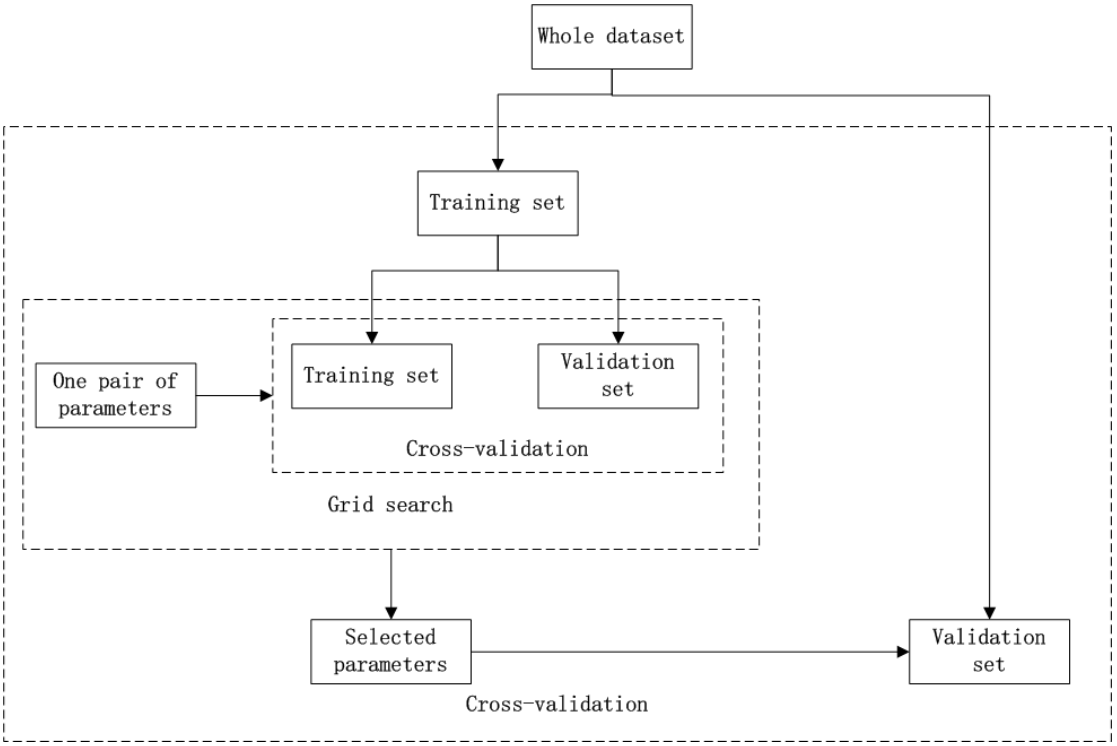

Figure 3.28: The structure of validation and model selection process

The process includes two cross-validation stages. One is for the whole dataset and the other one is for grid search.

Since the two parameters C and γ were both real-valueded without bound, a human defined sub-space was set as the search range. A heuristic [42] was used in order to lock the rough position of γ. The reciprocal of the median of the pair-wise distances of all training samples was calculated and defined as γ_0. The search range for γ was then set as $\gamma_0 \cdot 2^k, k \in \{-3, -2, \dots, 3\}$. The search range for C was first set as $2^k, k \in \{-5, -2, \dots, 15\}$ according to [43]. In order to reduce operation time, for each experiment, the whole dataset was first sent to the grid search process for model selection. The selected parameters for all experiments showed a range of $2^k, k \in \{-4, -2, \dots, 2\}$. Thus, the search range for penalty C was narrowed to $2^k, k \in \{-5, -2, \dots, 3\}$.

4. Experiments, Results and Discussions

This chapter contains four major parts: section 4.1 describes the experiments using a previous proposed method; section 4.2-4.4 present the experiments on the three databases: bed sensor dataset, MITBPD and sleep-EDF, respectively. The experiments include REM detection, Awake detection, Awake&REM detection and three stage (Awake, REM and NREM) classification based on the structure and performance of each database. Each section contains experiments, results and discussions. Table 4.1 lists all experiments.

Table 4.1: List of experiments

Section	Experiment No.	Database	Classifier	Short description
4.1	1	Bed sensor	Threshold comparison classifier	REM detection using method proposed in [10]
	2	MITBPD		Awake and REM detection using method proposed in [10]
4.2.1	1	Bed sensor	SVM classifier	REM detection using LFCC
	2			REM detection using smoothed LFCC
	3			REM detection using smoothed HRV and smoothed RV
	4			Three stage classification using smoothed LFCC
4.2.2	5			REM detection using HRV and RV features
	6			REM&Awake vs. NREM using HRV and RV features
4.3.1.1	1	MITBPD	Threshold comparison classifier	Awake detection using smoothed RMSSD
	2			Awake detection using smoothed HF.
	3			Awake detection using detrended mean value of the first coefficient of LFCC (mLFCC1)
4.3.1.3	4		SVM classifier	Awake detection using HRV
	5			Awake detection using smoothed LFCC
	6			Awake detection using smoothed HRV and smoothed LFCC
	7			Decision boundary adjustment based on Exp.6
	8			Combine results with Exp.7 and Exp.3
4.3.2	9			Awake detection using mean value of LFCC(mLFCC)
4.4	1	Sleep-EDF	Threshold comparison classifier	Awake&REM vs. NREM with detrended respiratory rate (RR)
	2			A post-processing was applied to Exp.1
	3			Three stage classification based on Exp. 2

4.1. Experiments Using Previous Proposed Method

The method proposed in [10] was implemented and applied on the bed sensor dataset and the MITBPD. Their method used threshold comparison classifiers with a hierarchical structure to classify Awake, REM, Light (N1 and N2) and deep (N3) stages. HRV features were extracted from the BCG signal. The four class classification with a 77.1±3.3%

accuracy and 0.58±0.06 kappa value was a very good result in the literature, especially the kappa value.

According to the paper, the hierarchical system consisted of three layers: Awake detection, REM detection and deep sleep detection. All features were extracted on every 30s epoch. The body movements and heart rate was used in the first layer. Thresholds for the first layer were: 1) If the body movements are more than 15 sec, the epoch is estimated as Awake. 2) If the heart rate is higher than the average of last three minutes consecutively for more than 15 sec, this epoch is also estimated as Awake. Features in the second layer were: heart rate, SDNN and alpha value of detrended fluctuation analysis (DFA) of heart beat intervals (HBI). The third layer employed SDNN, alpha value of DFA of HBI, LF/HF and correlation of heart rate series with its shifted version (rRR). Thresholds for the second and third layers were the smoothed value of each feature over ten epochs. If the features in one epoch were all higher (the second layer) or lower (the third layer) than thresholds, this epoch was assigned to REM (the second layer) or deep (the third layer) sleep.

Only the first and second layers were implemented since the emphasis of this work were Awake, REM and NREM. Because some detailed parameters were not given in the paper, the implementation might not be exactly the same as their proposed method.

- **Experiment 1: REM detection with bed sensor dataset**

The method was first applied on the bed sensor dataset. The unfiltered ground truth version was used because the threshold comparison classifiers needed continuous epochs.

However, the first layer of their system, the Awake detection, needed consecutive beat-to-beat intervals and the current beat detection algorithm for the BCG signal couldn't do that. Thus, only the second layer which was the REM detection was implemented. The performance measures are shown in Table 4.2, in which REM is the positive class and NREM is the negative class. Awake stages are not included in the table.

Table 4.2: Performance measures on REM detection with the bed sensor dataset using a previous method

	TP	**TN**	**FP**	**FN**	**Se (%)**	**Sp (%)**	**Pr (%)**	**Acc (%)**	**k**
Night1	7	269	15	59	10.61	94.72	31.82	78.86	0.07
Night2	7	155	14	170	3.95	91.72	33.33	46.82	-0.04
Night3	12	88	8	155	7.19	91.67	60.00	38.02	-0.01
Night4a	12	150	9	125	8.76	94.34	57.14	54.73	0.03
Night4b	4	145	11	34	10.53	92.95	26.67	76.80	0.05
Night5a	6	112	12	45	11.76	90.32	33.33	67.43	0.03
Night5a	6	93	6	51	10.53	93.94	50.00	63.46	0.05
Night6	27	37	1	273	9.00	97.37	96.43	18.93	0.02
Night7	9	162	7	128	6.57	95.86	56.25	55.88	0.03
Night8	1	514	29	32	3.03	94.66	3.33	89.41	-0.02
Mean					8.19	93.75	44.83	59.04	0.02

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa)

The table demonstrated that their proposed method for REM detection was not suitable for the bed sensor dataset. The average kappa was only 0.02 which barely reached the range of slight agreement (0.01-0.2). Thus, it was more like a random guessing.

- **Experiment 2: Awake detection and REM detection with the MITBPD**

The MITBPD was then tested with their method. For this database, the body movements were not provided, so only heart rate was used in the Awake detection. Table 4.3 displays the performance of the Awake detection.

Table 4.3: Performance measures on Awake detection with the MITBPD using a previous method

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k
Slp01a	3	216	10	5	37.50	95.58	23.08	93.59	0.25
Slp01b	11	149	25	169	6.11	85.63	30.56	45.20	-0.08
Slp02a	9	292	10	43	17.31	96.69	47.37	85.03	0.19
Slp02b	24	147	9	84	22.22	94.23	72.73	64.77	0.18
Slp03	24	533	30	109	18.05	94.67	44.44	80.03	0.16
Slp04	16	515	38	145	9.94	93.13	29.63	74.37	0.04
Slp14	26	360	32	290	8.23	91.84	44.83	54.52	0.00
Slp16	43	338	40	267	13.87	89.42	51.81	55.38	0.04
Slp32	39	209	37	349	10.05	84.96	51.32	39.12	-0.04
Slp37	2	602	17	71	2.74	97.25	10.53	87.28	0.00
Slp41	27	503	48	196	12.11	91.29	36.00	68.48	0.04
Slp45	15	596	41	98	13.27	93.56	26.79	81.47	0.09
Slp48	20	500	46	188	9.62	91.58	30.30	68.97	0.02
Slp59	6	296	22	128	4.48	93.08	21.43	66.81	-0.03
Slp60	12	402	12	269	4.27	97.10	50.00	59.57	0.02
Slp61	11	560	36	107	9.32	93.96	23.40	79.97	0.04
Slp66	19	243	18	153	11.05	93.10	51.35	60.51	0.05
Slp67x	6	73	4	65	8.45	94.81	60.00	53.38	0.03
Mean					12.14	92.88	39.20	67.69	0.06

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa)

The above table shows the average kappa value is 0.06 which is still in the slight agreement range. However, in [10], the reported average kappa value for Awake detection was 0.83. The lack of the body movements feature might reduce the performance to some extent, but it would not be a significant influence. The rule of their proposed method with body movements is: if the body movements are more than 15 sec, the epoch is estimated as Awake. Epochs with more than 15 sec movements would not take a great proportion of all epochs. So even using the body movements feature would not increase the sensitivity a lot. Results of the second layer, the REM detection, are shown in Table 4.4. Again REM is the

positive class, NREM is the negative class and Awake is not included. Three of the recordings: slp32, slp66 and slp67x don't have REM sleep during the night, so their measurements were not given in the table.

Table 4.4: Performance measures on REM detection with the MITBPD using a previous method

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k
Slp01a	4	193	16	9	30.77	92.34	20.00	88.74	0.18
Slp01b	7	127	19	18	28.00	86.99	26.92	78.36	0.15
Slp02a	6	215	8	71	7.79	96.41	42.86	73.67	0.06
Slp02b	4	112	11	25	13.79	91.06	26.67	76.32	0.06
Slp03	14	411	74	60	18.92	84.74	15.91	76.03	0.03
Slp04	2	480	47	21	8.70	91.08	4.08	87.64	0.00
Slp14	17	297	59	19	47.22	83.43	22.37	80.10	0.20
Slp16	6	260	53	59	9.23	83.07	10.17	70.37	-0.08
Slp32	0	213	33	0	-	-	-	-	-
Slp37	0	541	63	11	0.00	89.57	0.00	87.97	-0.03
Slp41	13	407	54	77	14.44	88.29	19.40	76.23	0.03
Slp45	11	490	65	70	13.58	88.29	14.47	78.77	0.02
Slp48	4	449	65	27	12.90	87.35	5.80	83.12	0.00
Slp59	1	250	33	34	2.86	88.34	2.94	78.93	-0.09
Slp60	7	329	51	24	22.58	86.58	12.07	81.75	0.07
Slp61	17	398	119	62	21.52	76.98	12.50	69.63	-0.01
Slp66	0	220	41	0	-	-	-	-	-
Slp67x	0	65	9	0	-	-	-	-	-
Mean					16.82	87.63	15.74	79.17	0.04

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa)

The average kappa value for this experiment is again very poor. Although the mean accuracy is 79.17%, this is because most of the epochs were classified to the major class (NREM). The sensitivity and specificity can also demonstrate that. The specificity reaches 87.63% while the sensitivity is only 16.82%.

The above experiments demonstrated that the method proposed in [10] couldn't provide a

similar performance on the bed sensor dataset or MITBPD as it did on their database.

4.2. Bed Sensor Dataset

There are totally 8 nights of data and 10 recordings (Night 4 and 5 have 2 recordings) in the bed sensor dataset. Two training strategies were applied on this dataset with an SVM classifier: put-all-recordings-together and leave-n-nights-out (n=1). REM detection and three stage classification (Awake, REM and NREM) were implemented with put-all-recordings-together strategy. Only REM detection and REM&Awake detection were implemented with leave-n-nights-out strategy. A discussion about these two strategies are presented at the end of this section.

Three types of features were extracted from BCG signals: HRV, RV and LFCC. The smoothed version of these features were obtained by applying RLOWESS method with a 5 point sliding window. So there were totally 6 types of features: original/smoothed HRV, original/smoothed RV and original/smoothed LFCC. These features were tested with different combinations. Some of these experiments are displayed and analyzed in the rest of this section.

4.2.1. Put-all-recordings-together

In the put-all-recordings-together strategy, all recordings were put together and training and validation sets were randomly picked. A 10-fold cross-validation was adopted as the first layer of cross-validation and a 5-fold cross-validation was used in the grid search

process. The first three experiments implemented REM detection with different feature sets and the last experiment conducted the three stage classification.

- **Experiment 1: REM detection using original LFCC features**

In the first experiment, the original LFCC features were fed to the SVM classifier for REM detection. After 10-fold cross-validation, the mean of the confusion matrix are shown in table 4.5.

Classes: Class 1: REM; Class 2: Awake & NREM

Feature Set: LFCC features

Results:

Table 4.5: Mean Confusion matrix of REM detection using original LFCC features with the bed sensor dataset

Actual / Output	REM	NREM & Awake
REM	39.1	32.7
NREM & Awake	20.9	233.3

Se=65.52%, Sp=87.71%, Pr=54.45%, Acc=83.59%

k=0.49±0.05, AUC=0.86±0.02

The experiment achieved 0.49 as the average kappa value which means a moderate agreement with the ground truth. The results also exceeded some of the previous works. But comparing sensitivity with specificity, the classifier seemed more likely to assign epochs to NREM&Awake class.

- **Experiment 2: REM detection using smoothed LFCC features**

The second experiment tested whether the smoothing process would improve the

classification results. So the smoothed LFCC features were employed in this experiment. Table 4.6 listed the confusion matrix.

Classes: Class 1: REM; Class 2: Awake & NREM

Feature Set: Smoothed LFCC features.

Results:

Table 4.6: Mean Confusion matrix of REM detection using smoothed LFCC features with the bed sensor dataset

Actual / Output	REM	NREM & Awake
REM	52	14.3
NREM & Awake	8	251.7

Se=86.67%, Sp=94.62%, Pr=78.43%, Acc=93.16%

k=0.78±0.03, AUC=0.97±0.02

The results indicate that the smoothing process improved the kappa value significantly from 0.49 to 0.78. Both sensitivity and specificity are higher than those in the previous experiment.

- **Experiment 3: REM detection using smoothed HRV and smoothed RV**

The above experiments analyzed the effect of the LFCC feature set. However, there hasn't been an explanation of the actual meaning of this feature on the BCG signal. Thus, the more meaningful features: HRV and RV were tested in this experiment. The smoothed version was used due to the improvement discovered in experiment 2.

Classes: Class 1: REM; Class 2: Awake & NREM

Feature Set: Smoothed HRV features and smoothed RV features

Results:

Table 4.7: Mean Confusion matrix of REM detection using smoothed HRV and smoothed RV with the bed sensor dataset

Actual / Output	REM	NREM & Awake
REM	48.2	26.3
NREM & Awake	11.8	239.7

Se=80.33%, Sp=90.11%, Pr=64.70%, Acc=88.31%

k=0.64±0.04, AUC=0.93±0.02

The average kappa value indicates that the smoothed HRV and RV feature set performed slightly worse than the smoothed LFCC feature set (k=0.78), but better than the unsmoothed LFCC features (k=0.49).

- **Experiment 4: three stage classification using smoothed LFCC features**

Three stage classification was implemented with one vs. rest strategy to classify Awake, REM and NREM. Three classifiers were built. They were: NREM vs. REM&Awake, REM vs. NREM&Awake and Awake vs. REM&NREM. Table 4.8 displays the confusion matrix and the measurements.

Classes: Class 1: NREM; Class 2: REM; Class 3: Awake

Feature Set: Smoothed LFCC features

Table 4.8: Mean Confusion matrix of three stages classification (Awake, REM and NREM) using smoothed LFCC features with the bed sensor dataset

Actual / Output	NREM	REM	Awake
NREM	116.3	3.3	17.2
REM	5.1	49.9	10.1
Awake	18.6	6.8	97.7

ACC =81.20%±2.90%, Kappa=0.70±0.05

Although the accuracy was just above 80%, the 0.7 kappa value means a substantial agreement between two observers (ground truth and classifier). To the best of my knowledge, the classification results beat all of the previous works.

4.2.2. Leave-one-night-out

In this part, the experiments were conducted under the leave-one-night-out strategy. This strategy was also widely used when study the sleep related problems. In fact, it is more meaningful than the put-all-recordings-together strategy, because in a practical application, sleep quality is evaluated night by night.

In the put-all-recordings-together strategy, some of the feature sets showed very good results, especially the smoothed LFCC features. However, none of these combinations of feature sets worked using the leave-one-night-out strategy. Experiment 5 shows one of the experiment with the best performance.

- **Experiment 5: REM detection using original HRV and original RV features**

This experiment used original HRV and original RV features for REM detection. Table 4.9 listed the performance measures.

Classes: Class 1: REM; Class 2: Awake & NREM

Feature Set: HRV and RV

Results:

Table 4.9: Performance measures on REM detection using original HRV and original RV features with the bed sensor dataset

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	AUC
Night1	6	203	136	1	85.71	59.88	4.23	60.40	0.04	0.68
Night2	14	225	14	92	13.21	94.14	50.00	69.28	0.09	0.63
Night3	43	89	44	68	38.74	66.92	49.43	54.10	0.06	0.47
Night4a	31	179	36	50	38.27	83.26	46.27	70.95	0.23	0.65
Night4b	10	58	143	6	62.50	28.86	6.54	31.34	-0.02	0.46
Night5a	4	245	76	17	19.05	76.32	5.00	72.81	-0.02	0.40
Night5b	6	90	17	5	54.55	84.11	26.09	81.36	0.26	0.84
Night6	54	140	48	100	35.06	74.47	52.94	56.73	0.10	0.56
Night7	20	389	52	60	25.00	88.21	27.78	78.50	0.14	0.57
Night8	12	239	227	4	75.00	51.29	5.02	52.07	0.03	0.61
Mean					44.71	70.75	27.33	62.75	0.09	0.59

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy) ,k (Kappa), AUC(Area under the curve)

Although this feature set had the best performance among all combinations of features, we can still immediately see that the results are terrible. The average kappa value was only 0.09. The average AUC value was also only 0.59 which means even adjusting the decision boundary would not lead to a better result. Compared with the results with put-all-recordings-together strategy, the performance dropped significantly.

- **Experiment 6: REM&Awake vs. NREM using original HRV and original RV**

Experiment 5 demonstrated that when we applied the leave-one-night-out strategy, the performances were really bad. One conjecture is that the REM and Awake stages were mixed together and hard to separate because they had similar physiological indexes. Thus, if these two stages were put into the same class, the results might be improved. So in this experiment, REM and Awake were both put in the positive class. However, this conjecture was not confirmed. The average kappa value was a negative number which means the

classifier performed even worse than random guessing (Table 4.10).

Classes: Class 1: REM & Awake; Class 2: NREM

Feature Set: HRV and RV

Table 4.10: Performance measures on REM&Awake vs. NREM using original HRV and original RV with the bed sensor dataset

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	AUC
Night1	78	37	224	7	91.76	14.18	25.83	33.24	0.03	0.59
Night2	150	37	87	71	67.87	29.84	63.29	54.20	-0.02	0.53
Night3	168	7	27	42	80.00	20.59	86.15	71.72	0.00	0.49
Night4a	117	33	89	57	67.24	27.05	56.80	50.68	-0.06	0.48
Night4b	52	14	113	38	57.78	11.02	31.52	30.41	-0.28	0.30
Night5a	158	36	55	93	62.95	39.56	74.18	56.73	0.02	0.49
Night5b	26	48	19	25	50.98	71.64	57.78	62.71	0.23	0.62
Night6	258	3	13	68	79.14	18.75	95.20	76.32	-0.01	0.56
Night7	241	47	83	150	61.64	36.15	74.38	55.28	-0.02	0.48
Night8	55	14	412	1	98.21	3.29	11.78	14.32	0.00	0.53
Mean					71.76	27.21	57.69	50.56	-0.01	0.51

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy) ,k (Kappa), AUC(Area under the curve)

4.2.3. Discussions of Two Training Strategies

Comparing the results from the two strategies, it was surprising to see such different performances. In fact, various methods and features were tested on the bed sensor dataset with leave-one-night-out strategy, but they all failed.

One reason might be that the characteristics in each recording was very different from others. But the recordings were all collected from the same subject. Although there was a 5 month gap between the first 3 and the remaining 5 recordings, no big changes of living habits or other health related problems occurred. So the physiological differences between different recordings should be small and not lead to such poor results.

The reliability of the ground truth is still the biggest concern. Although [7] announced a 75.36% accuracy and 0.64 kappa value, the experiments were conducted in a lab setting. When the data were collected in a home environment without monitoring, the quality of the EEG signal couldn't be verified. The movements of the body might have caused a bad connection with the sensor that led to a noisy or inaccurate EEG signal. Thus, the classification on these noisy signals would not give the correct stages. By looking at the hypnograms themselves, we could also draw the conclusion that the ground truth was not reliable. Figure 4.1 shows one example of the original and filtered ground truth. No pattern like a normal sleep cycle is shown on the plots. The implemented ground truth filtering process might smooth the stages a little bit and make it look better, but it could not fix the actual errors.

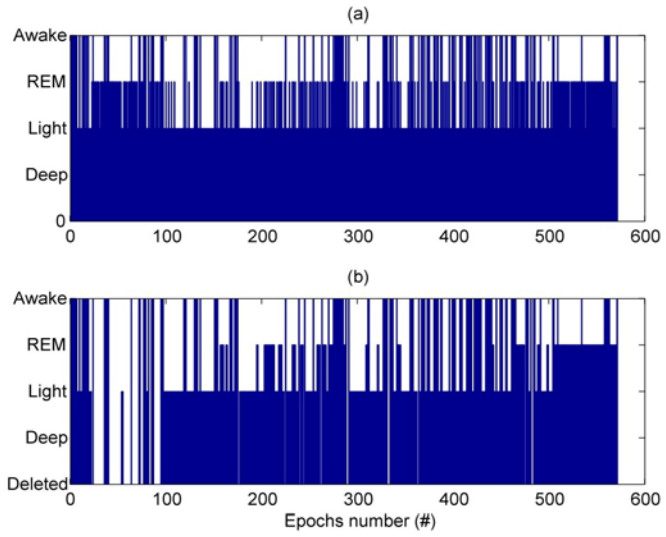

Figure 4.1: One example of the original ground truth given by EEG sleep detection device and the filtered one after applied processes described in section 3.2.1.1

(a) original hypnogram;(b) filtered hypnogram. The "Deleted" label means that epoch was removed from the recording after rule 2 or rule 3. The orginal hypnogram shows very frequent changes between Awake and REM. Althought the filtering process smoothed it a little bit, a normal sleep cycle still couldn't be seen from the hypnogram.

But why does the put-all-recordings-together perform so well? One assumption is that when we put all data together and shuffle them, nearby epochs from same stages might be put into training and testing set, respectively, allowing the classifiers to learn these variations. Meanwhile, the features used in this problem reflected the body conditions at each epoch and sudden large changes of the conditions would not happen frequently. Thus, nearby epochs might have similar values in feature space regardless of the sleep classes.

Figure 4.2 shows a distance plot of one recording with the original LFCC features. Epochs were sorted by sleep stages in the time order. The plot indicates that each stage has more than one cluster. Thus, one epoch will more likely to be classified to the right class when there are epochs from the same cluster in the training set.

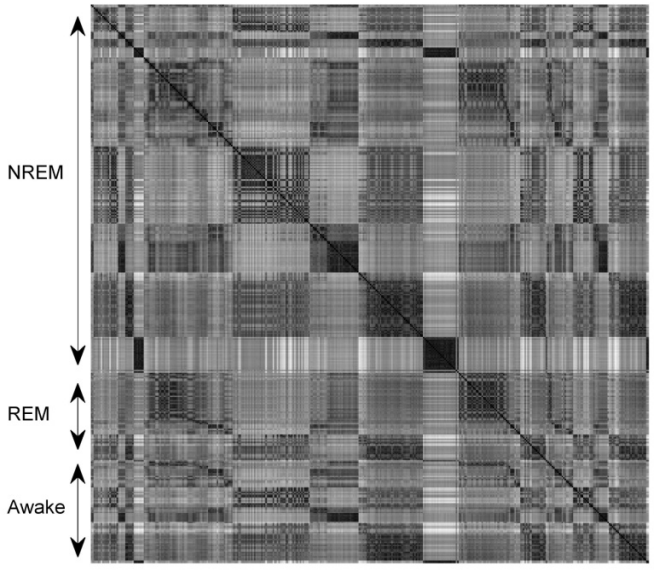

Figure 4.2: Distance plot of one recording with LFCC features

The labels on the left point out the range of each stage. Each stage were consisted of more than one cluster.

Moreover, this kind of connection would be strengthened by a smoothing process. This could explain why the smoothed feature set outperformed the original one. Figure 4.3

displays the distance plot of the same recording but with the smoothed LFCC features.

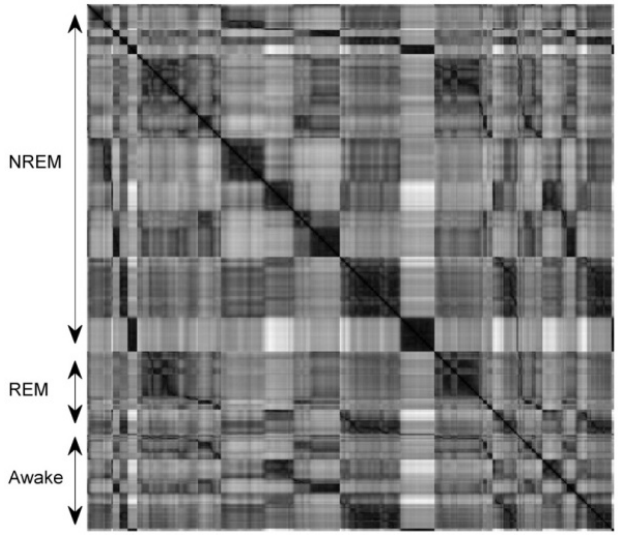

Figure 4.3: Distance plot of the same recording with smoothed LFCC features

The blocks became darker and more clear than the original LFCC features, which means the distances were smaller.

According to the discussion above, the bed sensor dataset had considerable uncertainty. So in order to evaluate the usefulness of the selected features and the classification method, they were then tested on the other two databases.

4.3. MIT-BIH Polysomnographic Database

For the MITBPD, two types of scenarios were conducted. The first scenario was the subject-independent experiments, which the training set consisted of recordings from all subjects. The second scenario was the subject-specific experiment, which the training set only consisted of recordings from the same subjects.

Because the subjects in this database suffered from sleep apnea, they woke up very frequently due to breathing issues. Thus, the Awake stages took a great proportion of the

night; meanwhile the REM stages took only a very small proportion. In addition, the occurrence of apnea in the sleep (both REM and NREM) might have affected the original characteristics of these two stages. Due to these reasons, only Awake detection was implemented with the MITBPD. Sleep efficiency was calculated in each experiment while sleep onset was not calculated because 10 out of 18 recordings didn't start with an Awake stage. So it was possible that the recordings were only a part of a night's sleep.

Two types of features were extracted from ECG signals: HRV and LFCC. The smoothed version of these features were obtained by applying RLOWESS method with a 10 point sliding window. So there were totally 4 types of features: original/smoothed HRV and original/smoothed LFCC.

4.3.1. Subject-independent Experiments

In the subject-independent scenario, two types of classifiers were adopted. They were a threshold comparison classifier and an SVM classifier.

In section 4.3.1.1, three experiments using a threshold comparison classifier are displayed. Two of the experiments used two features separately for Awake detection. The other experiment aimed at developing an additional classifier that can be combined with the SVM classifier.

In section 4.3.1.2, based on the results of the threshold comparison classifier, relationships between features and different sleep stages of this database are discussed.

In section 4.3.1.3, five experiments using SVM classifiers are described. The first three

experiments showed results of different combinations of feature sets. The next two experiments implemented two additional processes: a decision boundary adjustment process and two classifiers combination process (a threshold comparison and a SVM classifier).

4.3.1.1. Classification with A Threshold Comparison Classifier

In previous sections, it was noted that in REM or Awake stages, the variability of heart rate should be high while the parasympathetic nervous activity should decrease. RMSSD and HF are two HRV parameters which reflect the variability and parasympathetic activity. So the assumption was a higher value of RMSSD and a lower value of HF should be seen in Awake stages. Figure 4.4 and 4.5 show the box plots of these two features in Awake and sleep (REM and NREM). On each box, the central mark is the median, the edges of the box are the 25th (Q_1) and 75th (Q_3) percentiles. The red crosses indicate the outliers. Points are drawn as outliers if they are larger than $Q_3+1.5(Q_3-Q_1)$ or smaller than $Q_1-1.5(Q_3-Q_1)$. All 18 recordings were put together and the features of each recording were normalized from 0 to 1 by Min-Max scaling. Each figure includes both original and the smoothed features.

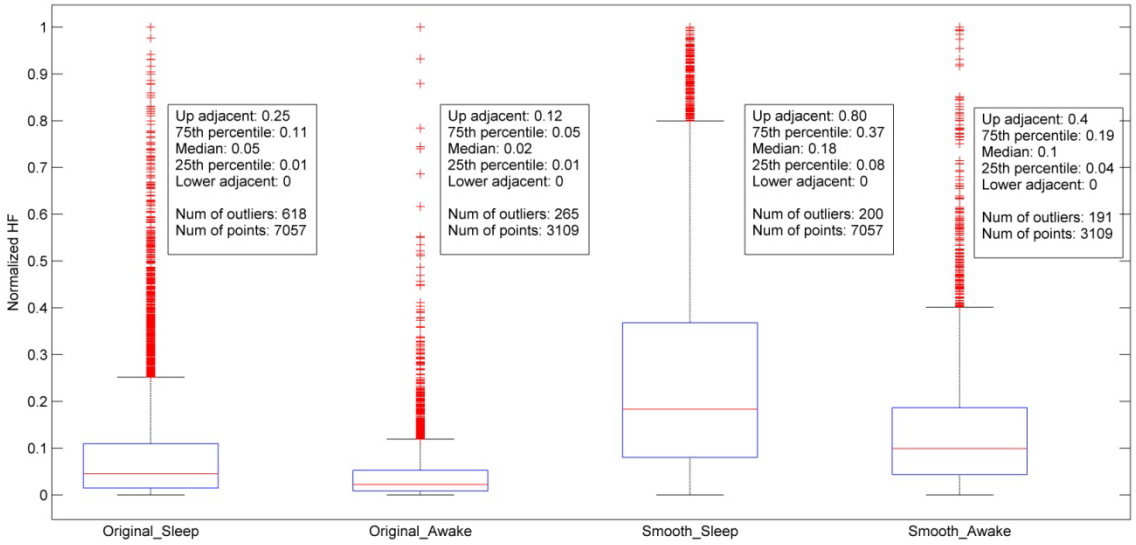

Figure 4.4: Box plots of the original and smoothed HF features in two classes

Red crosses indicate the outliers. Text boxes display information of the boxes. From left to right are the original HF in sleep, original HF in Awake, smoothed HF in sleep and smoothed HF in Awake. Both feautres have higher median in sleep than thoese in Awake.

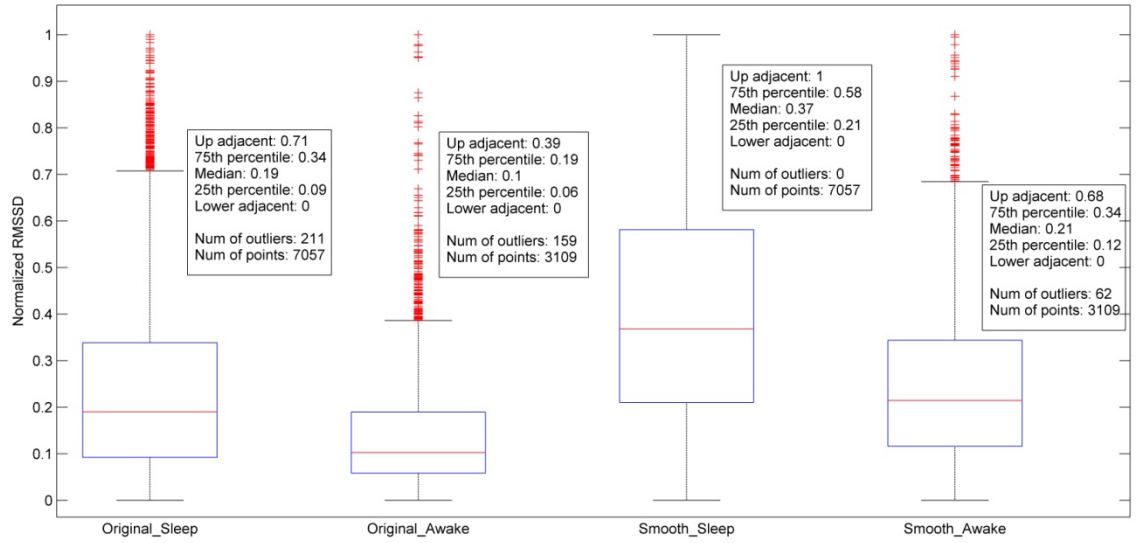

Figure 4.5: Box plots of the original and smoothed RMSSD features in two classes

Red crosses indicate the outliers. Text boxes display information of the boxes. From left to right are the original RMSSD in sleep, original RMSSD in Awake, smoothed RMSSD in sleep and smoothed RMSSD in Awake. Both features have higher median in sleep than thoese in Awake.

Both of these features show higher values in sleep than those in Awake. This matched the theories and the assumption for HF. However, RMSSD should have a higher value in

Awake compared with sleep, but did not. In fact, other time domain HRV parameters all showed a reversed pattern from the theories. They had lower value in Awake than in sleep. Research in [44] also found such patterns. Although the pattern was reversed, a threshold still could be set to separate these two stages. The smoothed features were used instead of the original features in order to reduce the outliers. Then for each recording, the threshold was set to the median value of the feature. Since Awake stages usually accounted for 30% in this database, the threshold would be lower than the actual median value of sleep (NREM and REM) stages.

- **Experiment 1: Awake detection using smoothed RMSSD**

Table 4.11 shows the results with the smoothed RMSSD.

Classes: Class 1: Awake; Class 2: NREM&REM

Feature Set: smoothed RMSSD

Method: if the value of an epoch was smaller than the threshold, this epoch was assigned as Awake

Results:

Table 4.11: Performance measures on Awake detection using smoothed RMSSD feature with the MITBPD

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	E_{se} (%)
Slp01a	5	117	115	3	62.50	50.43	4.17	50.83	0.02	46.67
Slp01b	108	108	72	72	60.00	60.00	60.00	60.00	0.20	0.00
Slp02a	44	172	136	8	84.62	55.84	24.44	60.00	0.20	35.56
Slp02b	89	116	46	19	82.41	71.60	65.93	75.93	0.52	10.00
Slp03	92	310	259	41	69.17	54.48	26.21	57.26	0.15	31.05
Slp04	128	326	232	34	79.01	58.42	35.56	63.06	0.26	27.50
Slp14	198	233	159	124	61.49	59.44	55.46	60.36	0.21	4.90
Slp16	204	235	143	112	64.56	62.17	58.79	63.26	0.27	4.47
Slp32	295	221	25	99	74.87	89.84	92.19	80.63	0.61	11.56
Slp37	74	348	275	1	98.67	55.86	21.20	60.46	0.21	39.26
Slp41	130	291	260	99	56.77	52.81	33.33	53.97	0.08	20.64
Slp45	104	363	274	15	87.39	56.99	27.51	61.77	0.24	34.26
Slp48	172	338	208	42	80.37	61.90	45.26	67.11	0.34	21.84
Slp59	88	177	141	52	62.86	55.66	38.43	57.86	0.16	19.43
Slp60	241	306	109	45	84.27	73.73	68.86	78.03	0.56	9.13
Slp61	93	329	267	31	75.00	55.20	25.83	58.61	0.17	32.78
Slp66	156	201	63	19	89.14	76.14	71.23	81.32	0.63	10.02
Slp67x	48	53	29	24	66.67	64.63	62.34	65.58	0.31	3.25
Mean					74.43	61.95	45.37	64.22	0.28	20.13

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa), E_{se} (Relative error of sleep efficiency)

The average sensitivity and specificity matched the distributions seen in figure 4.5, where 40% of the sleep epochs were misclassified to Awake. The results also show a large variation in different recordings, from the best kappa value as 0.63 (slp66) to the worst value as 0.02 (slp01a). This means that each recording had a different distribution of the feature, some of them had big differences between two classes while others were mixed together.

- **Experiment 2: Awake detection using smoothed HF**

The feature in the second experiment was the smoothed HF. Table 4.12 listed the

performance measures.

Classes: Class 1: Awake; Class 2: NREM&REM

Feature Set: smoothed HF

Method: if the value of an epoch was smaller than the threshold, this epoch was assigned as Awake

Results:

Table 4.12: Performance measures on Awake detection using smoothed HF feature with the MITBPD

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	E_{se} (%)
Slp01a	5	117	115	3	62.50	50.43	4.17	50.83	0.02	46.67
Slp01b	94	94	86	86	52.22	52.22	52.22	52.22	0.04	0.00
Slp02a	43	171	137	9	82.69	55.52	23.89	59.44	0.19	35.56
Slp02b	88	115	47	20	81.48	70.99	65.19	75.19	0.50	10.00
Slp03	90	308	261	43	67.67	54.13	25.64	56.70	0.13	31.05
Slp04	128	326	232	34	79.01	58.42	35.56	63.06	0.26	27.50
Slp14	195	230	162	127	60.56	58.67	54.62	59.52	0.19	4.90
Slp16	205	236	142	111	64.87	62.43	59.08	63.54	0.27	4.47
Slp32	285	211	35	109	72.34	85.77	89.06	77.50	0.55	11.56
Slp37	72	346	277	3	96.00	55.54	20.63	59.89	0.20	39.26
Slp41	128	289	262	101	55.90	52.45	32.82	53.46	0.07	20.64
Slp45	101	360	277	18	84.87	56.51	26.72	60.98	0.22	34.26
Slp48	176	342	204	38	82.24	62.64	46.32	68.16	0.36	21.84
Slp59	74	163	155	66	52.86	51.26	32.31	51.75	0.03	19.43
Slp60	217	282	133	69	75.87	67.95	62.00	71.18	0.42	9.13
Slp61	93	329	267	31	75.00	55.20	25.83	58.61	0.17	32.78
Slp66	154	199	65	21	88.00	75.38	70.32	80.41	0.61	10.02
Slp67x	47	52	30	25	65.28	63.41	61.04	64.29	0.29	3.25
Mean					72.19	60.50	43.75	62.60	0.25	20.13

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa), E_{se} (Relative error of sleep efficiency)

The smoothed HF feature had a very similar performance to smoothed RMSSD except for slp59 and slp60. Kappa values of these two recordings are about 0.14 higher in the previous

experiment than those in this experiment. This indicated that even though the HRV parameters were highly correlated with each other, they were still not exactly the same. So use these features together should improve the classification results.

- Experiment 3: Awake detection using mLFCC feature

In addition to the above two experiments, other features and different combinations were tried with the threshold comparison classifier but they all failed. So the threshold comparison classifier may not be suitable for this problem. However, it could be used as an additional classifier to improve the performance by combining results from this classifier with outputs from the SVM classifier. The goal was to detect as many Awake epochs as possible, but keep the specificity at a high value. That means the detected Awake epochs should have a very large possibility to actually be Awake. Then these detected epochs could be added to the results from the SVM classifier.

After some attempts, the mean value of the first coefficient of LFCC (mLFCC1) was found to be the most suitable feature. Large spikes would usually appear in the Awake stages. So feature detrending process was applied to the feature to highlight the spikes. After some tests, the threshold was set to its median value minus its standard deviation. If the feature was lower than this threshold, the epoch was classified as Awake stage. Figure 4.6 shows an example of this process and its result.

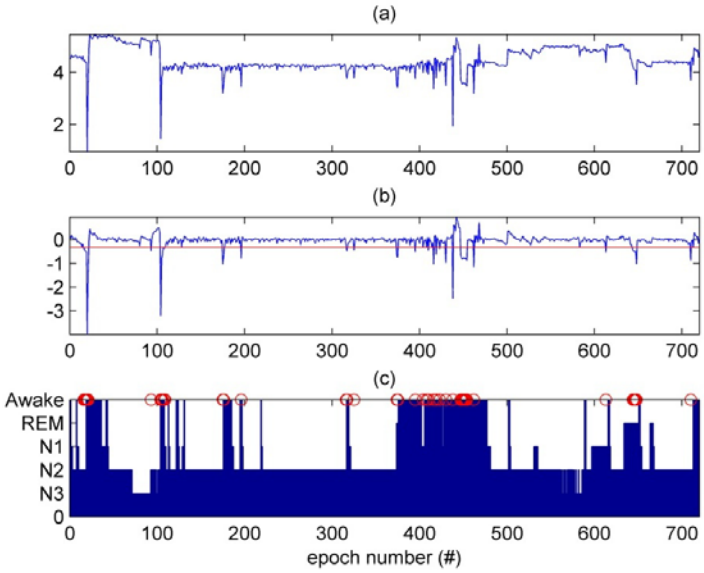

Figure 4.6: An example of Awake detection using the LFCC feature

(a) the original feature; (b) the detrended feature with red line as the threshold; (c) hypnogram with red circles indicating the detected Awake epochs. Nearly all the Awake periods have at least one detected epoch.

Table 4.13 shows the performance of each recording. The means of the sensitivity and specificity are 23.29% and 95.14% respectively, which meets the requirement of this process. Only 5% of the epochs in negative class were misclassified as Awake while more than 20% of the Awake epochs were detected.

Class 1: Awake; Class 2: NREM&REM

Feature Set: detrended mean value of the first coefficient of LFCC (mLFCC1)

Method: if the value of one epoch was lower than the median value minus its standard deviation, this epoch was assigned as Awake.

Table 4.13: Performance measures on Awake detection using mLFCC1 with the MITBPD

	TP	TN	FP	FN	Se (%)	Sp (%)
Slp01a	5	225	6	3	62.50	97.40
Slp01b	21	165	14	159	11.67	92.18
Slp02a	19	288	19	33	36.54	93.81
Slp02b	23	157	4	85	21.30	97.52
Slp03	44	560	8	89	33.08	98.59
Slp04	30	536	22	131	18.63	96.06
Slp14	62	379	13	259	19.31	96.68
Slp16	71	366	12	244	22.54	96.83
Slp32	55	239	7	338	13.99	97.15
Slp37	19	581	42	55	25.68	93.26
Slp41	24	521	30	204	10.53	94.56
Slp45	23	624	13	95	19.49	97.96
Slp48	45	535	11	168	21.13	97.99
Slp59	22	289	29	117	15.83	90.88
Slp60	50	369	46	235	17.54	88.92
Slp61	32	564	32	91	26.02	94.63
Slp66	37	236	28	137	21.26	89.39
Slp67x	16	80	1	56	22.22	98.77
Mean					23.29	95.14

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity)

4.3.1.2.Discussions of Relation Between Features and Sleep Stages

The threshold comparison classifier didn't achieve good results, but some interesting problems were revealed. One of them is the relation between HRV parameters and sleep stages in this database. Since REM epochs in the database are too few and no particular pattern has been discovered, the following section only discuss Awake and NREM. Previous research indicated that REM and Awake are characterized by irregular cardiac activity while NREM are dominated by slow and steady cardiac rhythms. This was the reason why HRV features were widely used to determine sleep stages. RMSSD is one of

the HRV parameters, so the value should be higher in the Awake than that in NREM stages. But figure 4.5 shows the opposite. The possible reason for this situation is because subjects in this database suffered apnea. Hence, their physiological features during different sleep stages may be different from healthy people. Figure 4.7 shows the hypnogram of slp02a and its corresponding RMSSD. The occurrences of apnea were also marked in the figure.

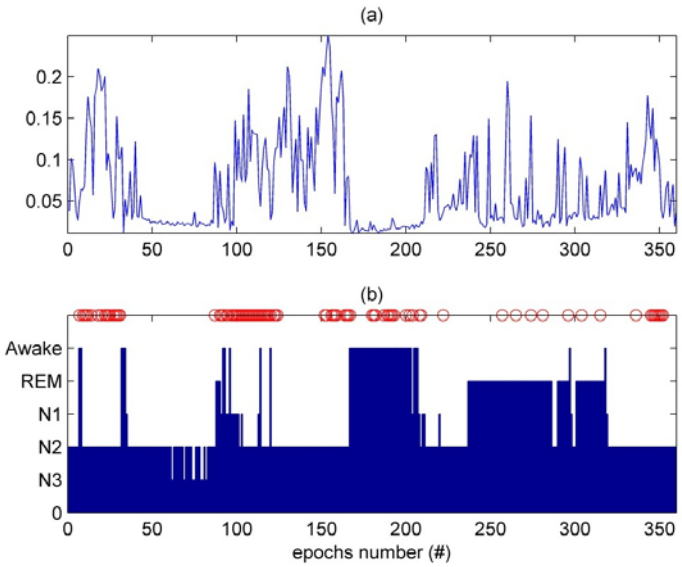

Figure 4.7: RMSSD feature and hypnogram of slp02a marked with apnea occurrences (red circles) Most of the higher values correspond to apnea except the Awake period from 160-200, where the values are as low as the no apnea NREM period from 40-100.(a) RMSSD; (b) hypnogram and apnea marked by red circles.

It is easy to find that the higher variability may be caused by the occurrence of apnea. The values of NREM and no apnea epochs between 50-100 are much lower compared with other apnea epochs. More importantly, their values are similar to those in Awake epochs between 160-200. In addition, REM epochs from 240 to 320 have some peaks but also some lower values. As a result, the classifier with the single RMSSD feature in Exp.1 obtained the outputs shown in Figure 4.8.

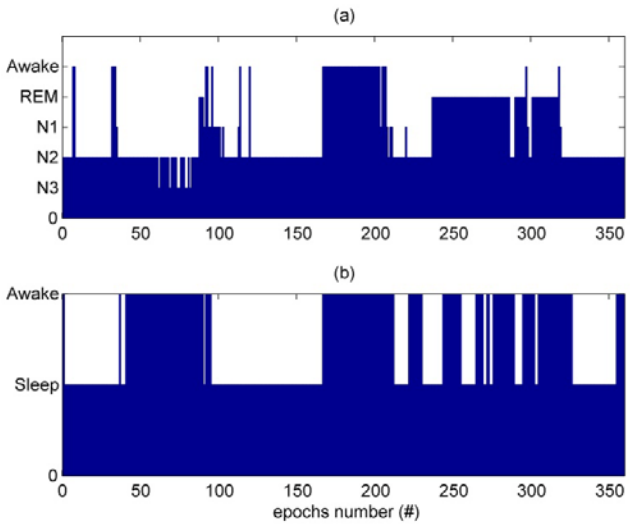

Figure 4.8: Detected Awake stages of slp02a using threshold comparison classifier with RMSSD (a) Ground truth. (b) Outputs of slp02a in Exp. 1. Epochs between 40-100 and 240-320 were misclassified as Awake.

However, although period of 120-150 doesn't show apnea, the RMSSD value still remains high. It could be that the cardiac system hadn't recovered from a previous apnea episode. Besides, the apnea during Awake stages seemed to have little impact on RMSSD for this subject. But this phenomenon was different for different subjects. Figure 4.9 shows the same type of plot for slp16. For this subject, the apnea happened in Awake stages increased the RMSSD value. The figure also shows an up and down trend in the range of 10-300 which demonstrates the relation between apnea and RMSSD. In addition, Figures 4.7 and 4.9 also demonstrate that the RMSSD values of these two subjects have different ranges. Subject slp02a has a wider range from 0 to 0.2 while values of slp16 are only between 0 and 0.12. This had no effect on the threshold comparison classifier because each recording was processed separately. But for the SVM classifier, this will affect the classification results. So features of each recording was normalized to 0 to 1 separately when a SVM

classifier was employed.

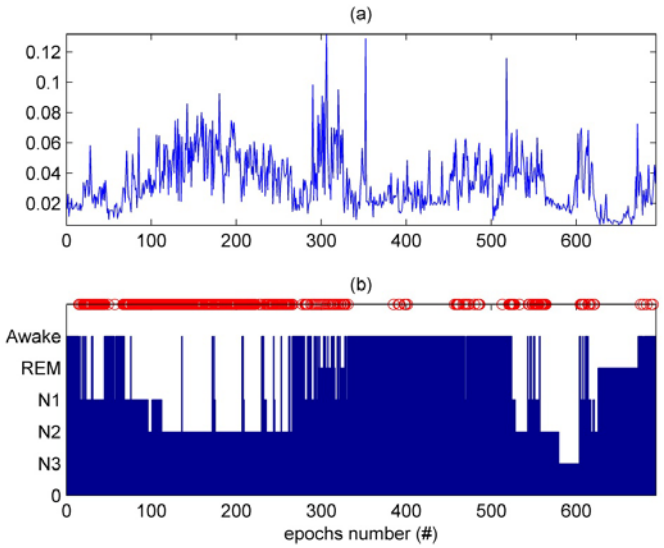

Figure 4.9: RMSSD feature and hypnogram of slp16 marked with apnea occurrences

Epochs between 10-350 show a rising trend from the lower density apnea area around 50 and a downtrend to the lower density apnea area around 280. Apnea in Awake stages 480, 550, 600 also show higher RMSSD values than those in no apnea Awake periods (epochs 0 and 50). (a) RMSSD; (b) hypnogram and apnea marked by red circles.

Besides RMSSD, other HRV parameters also showed they were highly related to apnea. But different from those time domain HRV parameters, although HF also was affected by apnea, it still had a lower value in Awake and higher value in NREM, which was the same as healthy people. Figure 4.10 is the hypnogram of slp02a with HF feature. So for this feature, the apnea caused higher and more fluctuant HF while NREM period without apnea had a relatively low value.

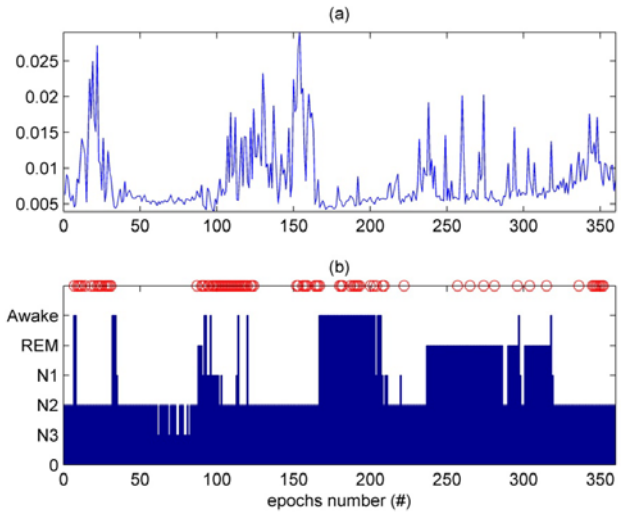

Figure 4.10: HF feature and hypnogram of slp02a marked with apnea occurrences

(a) HF feature. Most of the apnea caused higher HF value while NREM and no apnea period from 50-10 showed a lower value. (b) hypnogram and apnea marked by red circles.

From above discussion, it can be concluded that apnea would affect HRV parameters. Furthermore, these impacts could have different levels, from causing big influences to no effect. This uncertainty brought more challenge to the classification.

However, among all HRV parameters, one parameter is a little different from others. It doesn't reflect the variability directly, instead it relates to the value of heart beat intervals (HBI). This feature is the mean of heart beat intervals (mHBI). It is the most commonly used parameter and will be discussed below. It is inversely proportional to the heart rate (heart beats per minute).

The impact of apnea on this feature was not clear in this database, but the patterns of mHBI values were discovered. It is summarized as follow: 1) Nearly all the sudden short Awake epochs correlated with downward spikes. 2) For those large periods of Awake, besides the first few epochs, some of the middle epochs also showed downward spikes. 3) The values

of other epochs in large periods of Awake were varied. But most of them showed a gradually increasing pattern from Awake to sleep. Examples for these three summarizes can be seen in Figure 4.11 and 4.12.

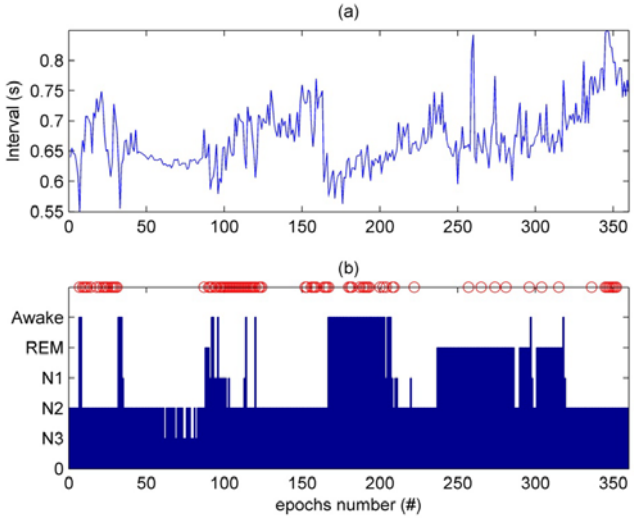

Figure 4.11: mHBI feature and hypnogram of slp02a marked with apnea occurrences
(a) mHBI. Short time Awake epochs around 10,40,100 and 120 show downward spikes. Large Awake period from 160-210 shows a sudden decrease of the value in the beginning and then a gradually increasing pattern.(b) hypnogram and apnea marked by red circles.

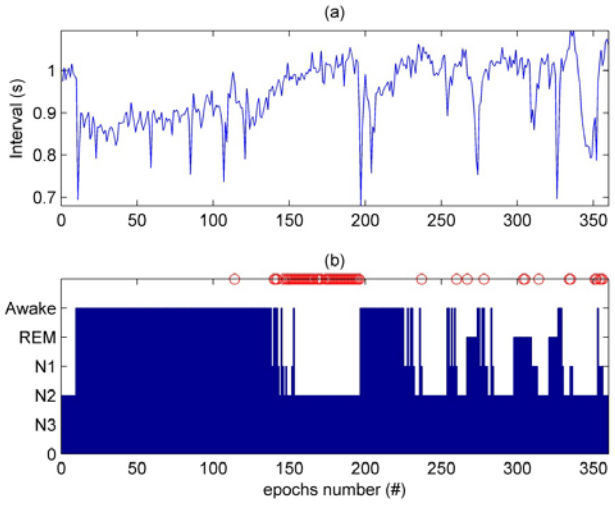

Figure 4.12: mHBI feature and hypnogram of slp01b marked with apnea occurrences
(a) mHBI. Epochs from 10-150 formed a large block of Awake which begins with a big spike. Five other spikes appear within the block. Other epochs show a relatively low value that gradually increases to the NREM level at the end. Epochs around 200-240 have the same pattern but the mHBI value returns to the NREM level more quickly. (b) hypnogram and apnea marked by red circles.

These patterns reflected in the data actually match what we usually experienced. A person woke up suddenly in the middle of the night would feel the increase of the heartbeat and then fall asleep quickly with the heartbeat restored. If the person stayed awake for a long time, it was then the same as the sleep onset process. Some body movements might happen or he/she would just lay there and gradually fall asleep, which resulted in a stable or decrease trend in heartbeat.

So far, the relation between HRV parameters and Awake stage and the impacts of apnea on the HRV parameters was discussed. Besides HRV parameters, the other feature related to sleep was the mean value of the first coefficient of LFCC (mLFCC1). In the third experiment above, mLFCC1 showed that its spikes were highly correlated with the Awake stage. This is an interesting finding because the meaning of LFCC features on the ECG signal haven't been explained. We have discussed above that the downward spikes of mHBI usually happened in the Awake epochs. So, is there any relation between the mLFCC1 and mHBI? These two features of slp01b and slp16 are plotted in Figure 4.13-4.14 with apnea and leg movement marks.

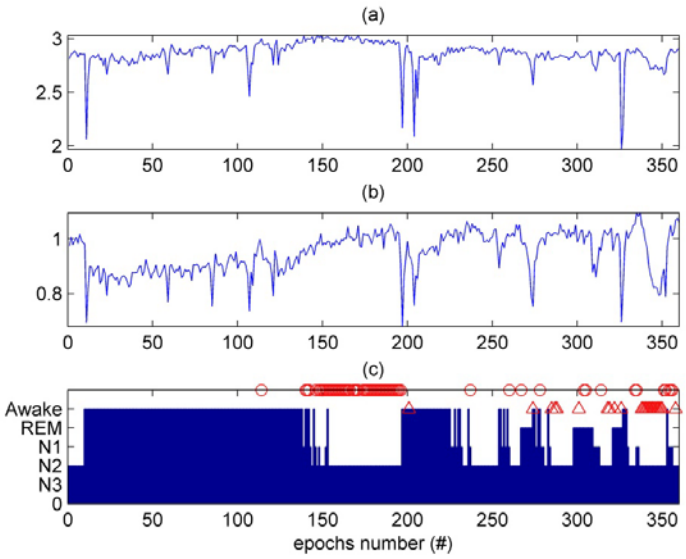

Figure 4.13: mLFCC1 and mHBI features and hypnogram of slp01b

The downward spikes are showed in nearly the same positions in (a) and (b). (a) mLFCC1. (b) mHBI. (c) hypnogram with apnea marked by red circles and leg movements marked by red triangles.

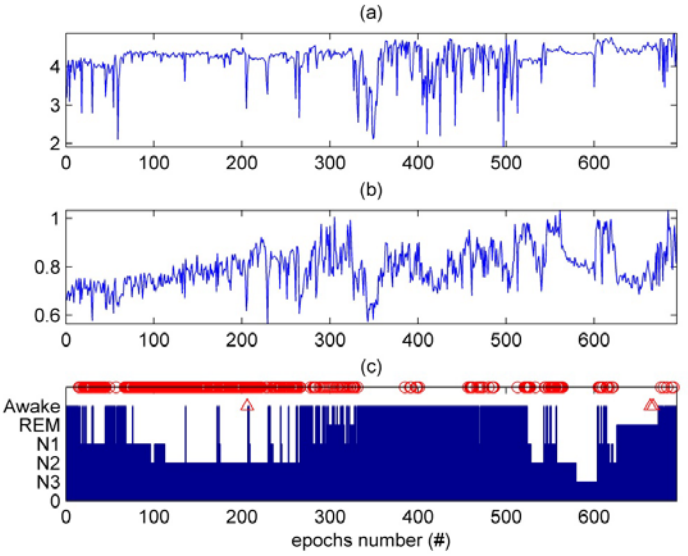

Figure 4.14: mLFCC1and mHBI features and hypnogram of slp16

The downward spikes in (a) are correlated with awake and more clear then those in (b). (a) mLFCC1. (b) mHBI. (c) hypnogram with apnea marked by red circles and leg movements marked by red triangles.

In figure 4.13, the two plots are nearly same, but the mLFCC1 is more like the detrended mHBI. The differences between two features look bigger in figure 4.14. Although the sudden drops of mHBI still can be found in Awake stages, they are not as clear as figure

4.11 and 4.12. The fluctuations caused by apnea may have reduced the contrast between apnea spikes and Awake spikes. In contrast, the spikes that correlated with Awake are very clear and big in the mLFCC1 feature. These are two typical examples. Most of the mLFCC1 show clear spikes in short term Awake, the beginning and some middle portions of large Awake periods. Thus, this feature may have captured big changes in heart rate or other physiological indexes when the subject woke up during the night.

Another possibility is that they are simply just body movements. However, the leg movements annotations (shown with triangles in Figures 4.13(c) and 4.14(c)) provided by the database don't correlate with these spikes. Nevertheless, we still can't rule out the possibility that other types of movements caused such spikes.

Although this feature couldn't separate Awake with other stages completely, Table 4.13 indicates that with a simple threshold comparison classifier, 23.29% of Awake could be detected in average, and only about 5% of sleep epochs were misclassified to Awake. So the proposed method in Exp. 3 has potential to improve the classification results by combining its results with the outputs from SVM classifiers.

4.3.1.3. Classification with an SVM Classifier

When classifying with a threshold comparison, the number of features that can be used is limited. Although most HRV parameters were highly correlated with each other, the combination of these features still could help improve the classification results. In this section, the SVM classifier is employed on the MITBPD with different feature

combinations.

One of the conclusions from the experiments of the bed sensor dataset was that even though the performance was very good with put-all-recordings-together strategy, it could still be terrible with leave-n-nights-out strategy. In addition, the latter strategy is the one that matches the actual application. So only leave-n-nights-out strategy was tried on this database. The MITBPD has 18 recordings, and n was set to 3 (leave-3-nights-out) in order to balance time consumption and the quantity of the training set.

The first three experiments tested effects of different feature sets. In experiment 7, a decision boundary adjustment process was implemented. In experiment 8, the Awake stages detected by the threshold comparison method (Exp.3) were added to the results from the best SVM classifier.

- **Experiment 4: Awake detection using original HRV features**

The original HRV features were used together in this experiment to see if the classification results can be improved.

Classes: Class 1: Awake; Class 2: REM &NREM

Feature Set: HRV

Results:

Table 4.14: Performance measures on Awake detection using original HRV features with the MITBPD

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	AUC	E_{se} (%)
Slp01a	4	115	116	4	50.00	49.78	3.33	49.79	0.00	0.51	46.86
Slp01b	160	77	102	20	88.89	43.02	61.07	66.02	0.32	0.72	22.84
Slp02a	38	277	30	14	73.08	90.23	55.88	87.74	0.56	0.87	4.46
Slp02b	92	105	56	16	85.19	65.22	62.16	73.23	0.48	0.85	14.87
Slp03	83	481	87	50	62.41	84.68	48.82	80.46	0.43	0.81	5.28
Slp04	37	449	109	124	22.98	80.47	25.34	67.59	0.04	0.61	2.09
Slp14	113	324	68	208	35.20	82.65	62.43	61.29	0.19	0.60	19.64
Slp16	230	181	197	85	73.02	47.88	53.86	59.31	0.20	0.64	16.16
Slp32	335	154	92	58	85.24	62.60	78.45	76.53	0.49	0.84	5.32
Slp37	43	513	110	31	58.11	82.34	28.10	79.77	0.28	0.80	11.33
Slp41	115	307	244	113	50.44	55.72	32.03	54.17	0.05	0.56	16.82
Slp45	89	361	276	29	75.42	56.67	24.38	59.60	0.17	0.73	32.72
Slp48	135	432	114	78	63.38	79.12	54.22	74.70	0.40	0.76	4.74
Slp59	66	275	43	73	47.48	86.48	60.55	74.62	0.36	0.69	6.56
Slp60	91	381	34	194	31.93	91.81	72.80	67.43	0.26	0.78	22.86
Slp61	111	208	388	12	90.24	34.90	22.24	44.37	0.11	0.74	52.29
Slp66	140	226	38	34	80.46	85.61	78.65	83.56	0.66	0.88	0.91
Slp67x	6	72	9	66	8.33	88.89	40.00	50.98	-0.03	0.57	37.25
Mean					60.10	70.45	48.02	67.29	0.28	0.72	17.94

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa), AUC(Area under the curve), E_{se} (Absolute error of sleep efficiency)

Table 4.14 indicates the average kappa is 0.28 which is the same as that in Exp. 1 (a threshold comparison classifier with single RMSSD), but the performance of each recording is different. Figure 4.15 displays the output of slp02a from this experiment. When using the single threshold comparison classifier, this recording misclassified many NREM to awake due to apnea (Figure 4.8). In this experiment, most of the epochs between 40-100 now were classified to the correct class. The REM epochs between 250-300 were also in the right class this time.

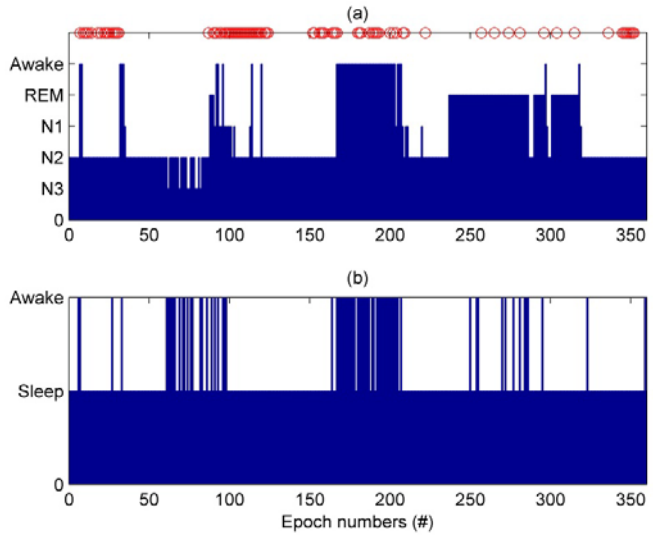

Figure 4.15: Detected Awake stages of slp02a using SVM classifier with 14 HRV features

(a) Ground truth with red circles as apnea. (b) Outputs. Most of the epochs between 40-100 and 250-300 were classified to the correct class.

But since the average performance didn't change, the classifier improved some results, but also reduced others. Figure 4.16 shows the outputs of slp67x from Exp.1 and this experiment (Exp.4) where the kappa value decreased from 0.31 to -0.03.

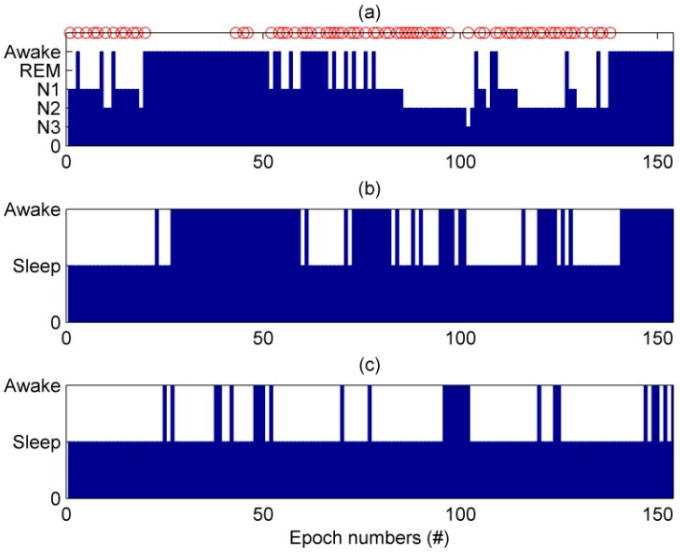

Figure 4.16: Detected Awake stages of slp67x from Exp.1 and this experiment

(a) Ground truth with red circles as apnea (b) Outputs of slp67x from Exp. 1. The rough position of the big block of Awake from 20-50 was detected but with a shift.(b) Outputs of slp67x from this experiment. Only a few Awake epochs were detected and the output around epoch 100 and 125 made the same mistaks as Exp.1.

Comparing figure 4.16 (b) and (c), the latter one was more like a subset of (b). They both misclassified some of the same non-Awake epochs to Awake. This means for some of the recordings, only HRV features may not be sufficient to separate two classes. In Exp. 3 (threshold comparison classifier with mLFCC1), the first coefficient of mean LFCC feature showed its relation with Awake stage. So in the next experiment, all the LFCC features were applied to the SVM classifier.

- **Experiment 5: Awake detection using smoothed LFCC features**

Both original LFCC and smoothed LFCC features were tested with the SVM classifier. The smoothed LFCC feature outperformed the original feature set, so it is described here. The results are listed in Table 4.15.

Classes: Class 1: Awake; Class 2: REM &NREM

Feature Set: Smoothed LFCC features

Table 4.15: Performance measures on awake detection using smoothed LFCC with the MITBPD

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	AUC	Ese (%)
Slp01a	5	158	74	3	62.50	68.10	6.33	67.92	0.06	0.71	29.58
Slp01b	152	137	43	28	84.44	76.11	77.95	80.28	0.61	0.87	4.17
Slp02a	43	116	192	9	82.69	37.66	18.30	44.17	0.08	0.66	50.83
Slp02b	103	113	49	5	95.37	69.75	67.76	80.00	0.61	0.93	16.30
Slp03	60	482	87	73	45.11	84.71	40.82	77.21	0.29	0.66	1.99
Slp04	120	502	56	42	74.07	89.96	68.18	86.39	0.62	0.90	1.94
Slp14	108	291	101	214	33.54	74.23	51.67	55.88	0.08	0.52	15.83
Slp16	236	305	73	80	74.68	80.69	76.38	77.95	0.55	0.85	1.01
Slp32	256	230	16	138	64.97	93.50	94.12	75.94	0.53	0.90	19.06
Slp37	46	473	150	29	61.33	75.92	23.47	74.36	0.22	0.75	17.34
Slp41	109	532	19	120	47.60	96.55	85.16	82.18	0.51	0.82	12.95
Slp45	95	452	185	24	79.83	70.96	33.93	72.35	0.33	0.84	21.30
Slp48	122	301	245	92	57.01	55.13	33.24	55.66	0.10	0.60	20.13
Slp59	100	266	52	40	71.43	83.65	65.79	79.91	0.54	0.77	2.62
Slp60	131	332	83	155	45.80	80.00	61.21	66.05	0.27	0.74	10.27
Slp61	107	116	480	17	86.29	19.46	18.23	30.97	0.02	0.57	64.31
Slp66	136	162	102	39	77.71	61.36	57.14	67.88	0.37	0.75	14.35
Slp67x	26	82	0	46	36.11	100.00	100.00	70.13	0.38	0.79	29.87
Mean					65.58	73.21	54.43	69.18	0.34	0.76	18.55

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa), AUC(Area under the curve), E_{se} (Absolute error of sleep efficiency)

The table displays an improvement of the average kappa value from 0.28 with HRV feature set (Exp.4) to 0.34 with this method. Moreover, the kappa value of slp01b, slp04, slp16,slp41 and slp67x in this experiment were at least 0.29 more than those in Exp.4. Especially slp41, the smoothed LFCC feature set increased the kappa value from 0.05 (Exp.4) and 0.08 (Exp.1 and 2) to 0.51. Figure 4.17 displays the outputs of slp41 from Exp.4 and this experiment.

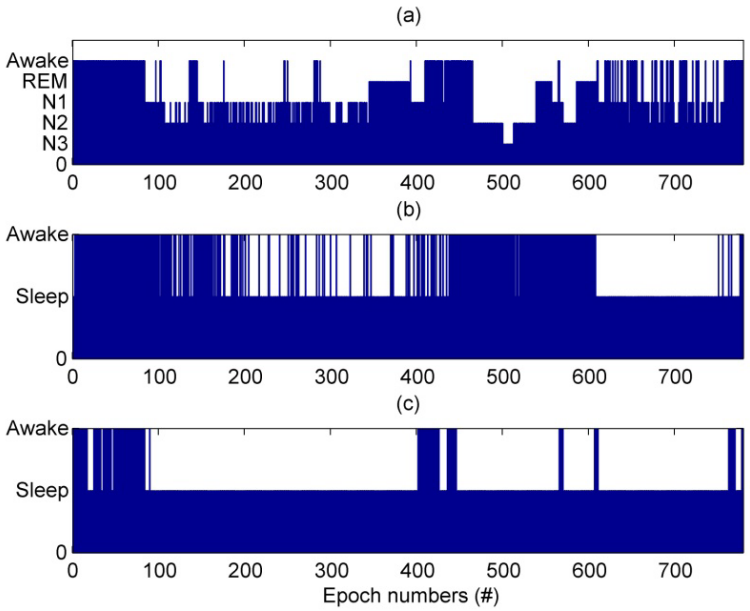

Figure 4.17: Detected Awake stages of slp41 from Exp.4 and 5

(a) Ground truth. This recording didin't provide the apnea annotations. (b) Outputs of slp41x from Exp. 4. Lots of sleep epochs were classified as Awake, especially area from 450 to 600. (c) Outputs of slp41 from this experiment (5). The misclassified awake epochs between 100-400 and 450-600 in Exp.4 were assigned to the correct class this time.

Although the performances of some recordings in Exp.4 also exceeded those in this experiment, the results showed the potential of LFCC features derived from the ECG signal. The differences of the results from two feature sets also implied that LFCC features might have extracted some useful information from the ECG signal that wasn't in the HRV parameters. So a combination of these two types of features should be tested.

- **Experiment 6: Awake detection using smoothed HRV and smoothed LFCC**

Considering both smoothed and unsmoothed versions, there were 10 possible combinations for HRV and LFCC features. The SVM classifier with the smoothed HRV and smoothed LFCC features gave the best results (Table 4.16), the average kappa value increased to 0.41 which joined the range of moderate agreement.

Classes: Class 1: Awake; Class 2: REM &NREM

Feature Set: Smoothed HRV and smoothed LFCC

Table 4.16: Performance measures on Awake detection using smoothed HRV and smoothed LFCC with the MITBPD

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	AUC	Ese (%)
Slp01a	5	174	57	3	62.50	75.32	8.06	74.90	0.09	0.76	22.59
Slp01b	170	131	48	10	94.44	73.18	77.98	83.84	0.68	0.93	10.58
Slp02a	49	124	183	3	94.23	40.39	21.12	48.19	0.14	0.77	50.14
Slp02b	102	127	34	6	94.44	78.88	75.00	85.13	0.70	0.95	10.41
Slp03	88	301	267	45	66.17	52.99	24.79	55.49	0.12	0.68	31.67
Slp04	120	522	36	41	74.53	93.55	76.92	89.29	0.69	0.92	0.70
Slp14	134	357	35	187	41.74	91.07	79.29	68.86	0.34	0.61	21.32
Slp16	199	353	25	116	63.17	93.39	88.84	79.65	0.58	0.84	13.13
Slp32	316	223	23	77	80.41	90.65	93.22	84.35	0.68	0.92	8.45
Slp37	55	534	89	19	74.32	85.71	38.19	84.51	0.42	0.89	10.04
Slp41	109	508	43	119	47.81	92.20	71.71	79.20	0.44	0.77	9.76
Slp45	99	483	154	19	83.90	75.82	39.13	77.09	0.41	0.87	17.88
Slp48	151	435	111	62	70.89	79.67	57.63	77.21	0.47	0.80	6.46
Slp59	85	277	41	54	61.15	87.11	67.46	79.21	0.50	0.79	2.84
Slp60	41	412	3	244	14.39	99.28	93.18	64.71	0.16	0.80	34.43
Slp61	109	195	401	14	88.62	32.72	21.37	42.28	0.09	0.75	53.82
Slp66	138	223	41	36	79.31	84.47	77.09	82.42	0.63	0.88	1.14
Slp67x	19	81	0	53	26.39	100.00	100.00	65.36	0.28	0.90	34.64
Mean					67.69	79.24	61.72	73.43	0.41	0.82	18.89

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa), AUC(Area under the curve), E_{se} (Absolute error of sleep efficiency)

Although most of the recordings showed higher kappa values than those in previous two experiments (only HRV and only smoothed LFCC features), kappa values in slp03 and slp60 decreased after combination. The kappa value of slp03 is only 0.12 but it was 0.43 and 0.29 in Exp.4 and Exp.5, respectively. The kappa value of slp60 is only 0.16 but it was 0.27 and 0.26 in Exp.4 and Exp.5, respectively. Figure 4.18 displays the ground truth and

three classification results from Exp.4, 5 and 6 of subject slp60.

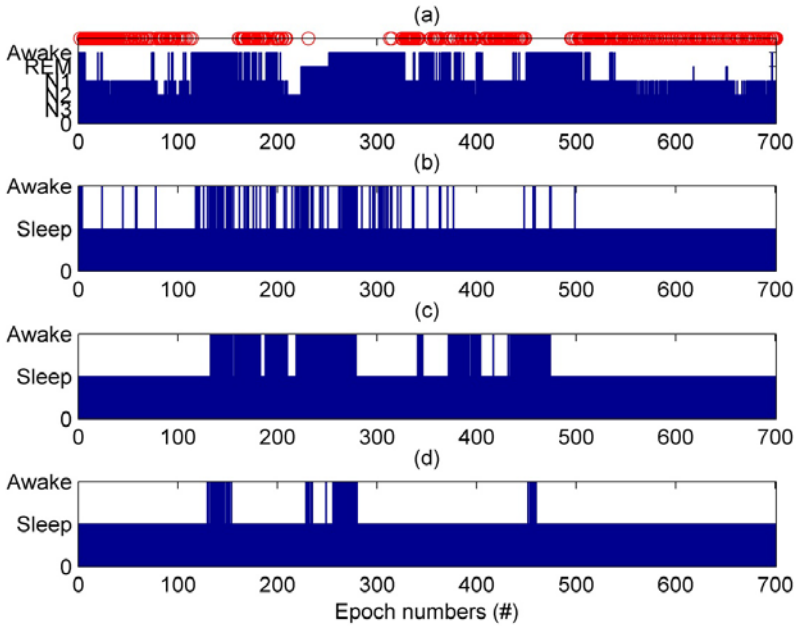

Figure 4.18: Ground truth and detected Awake stages of slp60 from Exp.4, 5 and 6
(a) Ground truth with red circles as apnea. There are four long-time Awake periods (100-200, 250-320, 350-500 and 450-500) in this recording. (b) Outputs of slp60 from Exp. 4. (c) Outputs of slp60 from Exp.5. (d) Outputs of slp03 from this experiment. Compared with (b) and (c), the method in this experiment only assigned a few epochs as Awake. However, most of the detected Awake stages were really Awake.

So the combination of features may not always be good. However, comparing the area under the curve (AUC) values of slp60 among these experiments, it was found that the AUC in this experiment was the best (0.8). The recording also showed a great imbalance between sensitivity and specificity. This is confirmed by figure 4.18(d). The sensitivity for slp60 was 14.39%, but the specificity reached 99.28%. Such imbalance also appeared in many other recordings such as slp02a, slp61 and slp67x. The AUC of slp67x was already 0.9, but the kappa value and the sensitivity were only 0.28 and 26.39%, respectively. This suggested that the decision boundary for the SVM classifier needed to be adjusted. In the next experiment, this process is discussed in detail.

- **Experiment 7: Awake detection using smoothed HRV and smoothed LFCC with a decision boundary adjustment process.**

The purpose of the decision boundary adjustment was to find a boundary that could maximize the average performance of all recordings. In this experiment, Kappa value was selected as the measure. The process was applied on the outputs of experiment 6 where smoothed HRV and smoothed LFCC features were used. By default, the SVM classifier implemented by "libsvm" library predicts a label f(x) by:

$$\mathrm{Dec}(x) = \sum_{i=1}^{l} a_i K(x, x_i) + b \quad , i = 1,2, \dots, l$$

$$f(x) = \mathrm{sign}(\mathrm{Dec}(x))$$

where l is the number of training samples, a_i is the Lagrange multiplier, K(•) is a kernel function, b is the bias, x_i is the training point and x is the unlabeled testing sample. The Dec(x) in the equation is called decision value. In this equation, each test record is assigned as positive class if the decision value is larger than 0, so the decision boundary here is 0. After adjusting the decision boundary, the classifier predicts a label f(x) by:

$$f(x) = \begin{cases} 1, & \text{if } \mathrm{Dec}(x) \geq B \\ -1, & \text{if } \mathrm{Dec}(x) < B \end{cases}$$

where B is the new decision boundary.

First, for each output in Exp.6, the receiver operating characteristic (ROC) curves were plotted. Figure 4.19 shows the ROC curve for slp60. Then the best boundary was defined as the closest point on the ROC to the perfect classification point (FPR=0, TPR=1).

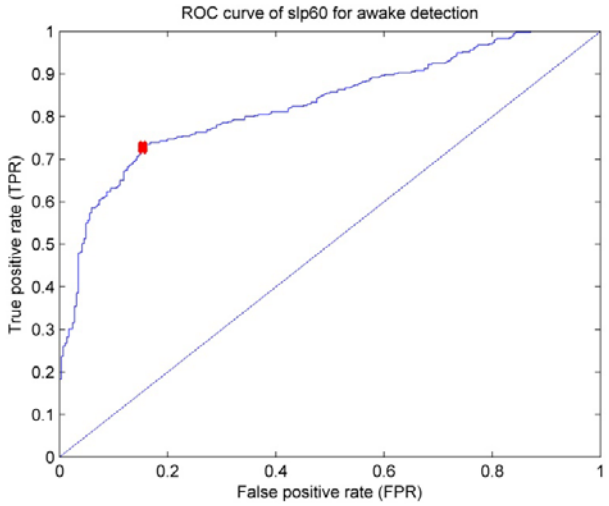

Figure 4.19: ROC curve of slp60 from Exp.6

The left top vertex (0,1) is defined as the perfect classification point where all samples are classified to the correct class. The red point indicated the best decision boundary which is the closest point to (0,1).

The figure also illustrates that if the best decision boundary is selected, the recording will obtain a 70% sensitivity and 80% specificity which is better than the current performance in Table 4.16. Figure 4.20 displays the histogram of the best decision boundaries for all recordings in Exp.6.

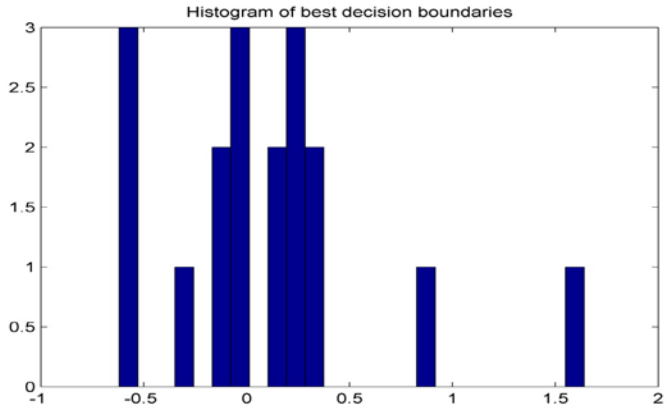

Figure 4.20: Histogram of best decision boundaries for each recording in Exp.6.

The figure shows a big range of values for the best decision boundaries and some of them are far from 0. It explained why some recordings have good AUC but bad kappa values.

In order to find the decision boundary that could maximize the average kappa value, a

range of {-0.7, -0.75,-0.8, ...,1.7} was chosen for searching the best boundary. Then the potential boundaries in this range were applied to each recording. For each potential boundary, the average kappa value of all recordings was calculated. The decision boundary of -0.15 was then selected according to the maximum average kappa value. Table 4.17 reveals the results after this process.

Classes: Class 1: Awake; Class 2: REM &NREM

Feature Set: Smoothed HRV, smoothed LFCC

Table 4.17: Performance measures on Awake detection using smoothed HRV and smoothed LFCC with the MITBPD. Decision boundary was adjusted in order to obtain the best average kappa value.

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	AUC	E_{se} (%)
Slp01a	4	196	36	4	50.00	84.48	10.00	83.33	0.12	0.74	13.33
Slp01b	160	159	21	20	88.89	88.33	88.40	88.61	0.77	0.94	0.28
Slp02a	48	149	159	4	92.31	48.38	23.19	54.72	0.18	0.78	43.06
Slp02b	102	131	31	6	94.44	80.86	76.69	86.30	0.73	0.95	9.26
Slp03	84	330	239	49	63.16	58.00	26.01	58.97	0.14	0.67	27.07
Slp04	118	533	25	44	72.84	95.52	82.52	90.42	0.71	0.92	2.64
Slp14	102	371	21	220	31.68	94.64	82.93	66.25	0.28	0.61	27.87
Slp16	176	367	11	140	55.70	97.09	94.12	78.24	0.55	0.85	18.59
Slp32	279	237	9	115	70.81	96.34	96.88	80.63	0.62	0.93	16.56
Slp37	55	570	53	20	73.33	91.49	50.93	89.54	0.54	0.89	4.73
Slp41	90	520	31	139	39.30	94.37	74.38	78.21	0.39	0.75	13.85
Slp45	95	535	102	24	79.83	83.99	48.22	83.33	0.50	0.89	10.32
Slp48	138	457	89	76	64.49	83.70	60.79	78.29	0.47	0.78	1.71
Slp59	80	304	14	60	57.14	95.60	85.11	83.84	0.58	0.79	10.04
Slp60	35	415	0	251	12.24	100.00	100.00	64.19	0.14	0.83	35.81
Slp61	102	226	370	22	82.26	37.92	21.61	45.56	0.10	0.74	48.33
Slp66	134	233	31	41	76.57	88.26	81.21	83.60	0.65	0.88	2.28
Slp67x	19	82	0	53	26.39	100.00	100.00	65.58	0.28	0.90	34.42
Mean					62.85	84.39	66.83	75.53	0.43	0.82	17.79

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa), AUC(Area under the curve), E_{se} (Absolute error of sleep efficiency)

After adjusting the decision boundary, the average kappa value was increased only a little, from 0.41 (Exp.6) to 0.43. However, this process didn't solve the unbalance of sensitivity and specificity. Figure 4.20 explains the reason. Because the best decision boundary for each recording had big differences, it was hard to pick a boundary that could meet all requirements. The result also demonstrates that in order to maximize the kappa value, the specificity was adjusted to be higher than sensitivity. Since the NREM&REM class is the majority class in this experiment, it means the imbalance of data still had an impact on the kappa value, though it was smaller than some other measures.

- **Experiment 8: Combining outputs from the threshold comparison classifier (Exp.3) with the outputs from the SVM classifier (Exp.7)**

Awake stages detected by the threshold comparison classifier using mLFCC1 (Exp.3) were added to the results of experiment 7. The outputs were combined with the OR operation. So as long as one of the classifiers assigned an epoch as awake, this epoch was considered as awake epoch. Table 4.18 listed the results.

Classes: Class 1: Awake; Class 2: REM &NREM

Feature Set: Smoothed HRV, smoothed LFCC

Method: Combined outputs from the threshold comparison classifier (Exp.3) and the SVM classifier (Exp. 7).

Table 4.18: Performance measures on Awake detection by combining outputs of two classifiers with the MITBPD

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	AUC	E_{se} (%)
Slp01a	8	194	38	0	100.00	83.62	17.39	84.17	0.25	0.74	15.83
Slp01b	162	153	27	18	90.00	85.00	85.71	87.50	0.75	0.94	2.50
Slp02a	51	140	168	1	98.08	45.45	23.29	53.06	0.19	0.78	46.39
Slp02b	104	128	34	4	96.30	79.01	75.36	85.93	0.72	0.95	11.11
Slp03	88	327	242	45	66.17	57.47	26.67	59.12	0.15	0.67	28.06
Slp04	121	522	36	41	74.69	93.55	77.07	89.31	0.69	0.92	0.69
Slp14	126	365	27	196	39.13	93.11	82.35	68.77	0.34	0.61	23.67
Slp16	197	357	21	119	62.34	94.44	90.37	79.83	0.58	0.85	14.12
Slp32	288	232	14	106	73.10	94.31	95.36	81.25	0.63	0.93	14.38
Slp37	56	544	79	19	74.67	87.32	41.48	85.96	0.46	0.89	8.60
Slp41	101	493	58	128	44.10	89.47	63.52	76.15	0.37	0.75	8.97
Slp45	98	528	109	21	82.35	82.89	47.34	82.80	0.50	0.89	11.64
Slp48	147	451	95	67	68.69	82.60	60.74	78.68	0.49	0.78	3.68
Slp59	84	282	36	56	60.00	88.68	70.00	79.91	0.51	0.79	4.37
Slp60	72	369	46	214	25.17	88.92	61.02	62.91	0.16	0.83	23.97
Slp61	105	206	390	19	84.68	34.56	21.21	43.19	0.09	0.74	51.53
Slp66	142	213	51	33	81.14	80.68	73.58	80.87	0.61	0.88	4.10
Slp67x	26	81	1	46	36.11	98.78	96.30	69.48	0.36	0.90	29.22
Mean					69.82	81.10	61.60	74.94	0.44	0.82	16.82

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa), AUC(Area under the curve), E_{se} (Absolute error of sleep efficiency)

After combining these two classifiers, the average sensitivity increased about 7% while the specificity decreased slightly, about 3%. The mean kappa value also improved to 0.44. Although the overall performance for awake detection with subject-independent scheme was not as good as expected. It was better than a previous study with the same database (Table 4.19).

Table 4.19: Comparison of results of Awake detection using subject-independent scheme with previous research

Author/year	Features	ACC (%)
Werteni, H. et al., 2014 [45]	HRV+DFA	70.78
Proposed method	HRV+LFCC	74.94

Figure 4.21 and 4.22 display the detected awake stages of slp01b and slp61 from this experiment. They have the best (0.75) and worst (0.09) kappa values, respectively. The performances on slp61 was not good in all the experiments, the best kappa value of 0.17 was obtained with the threshold comparison classifier (Figure 4.23).

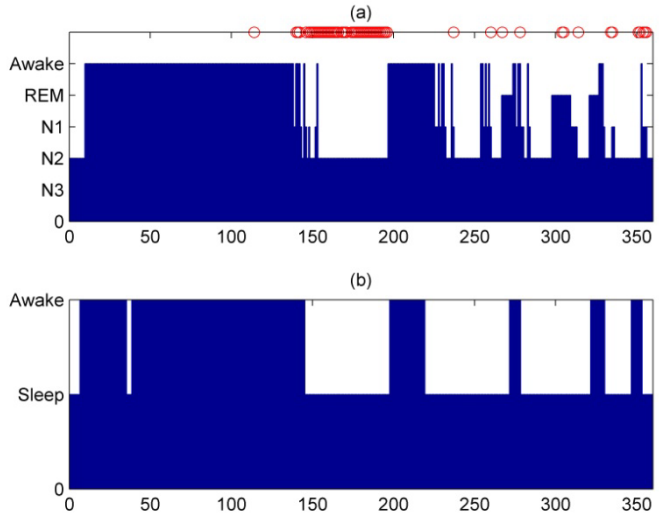

Figure 4.21: Detected awake stages of slp01b in this experiment
(a) Ground truth with red circles as apnea. (b) Outputs of slp01b with the kappa value as 0.75.

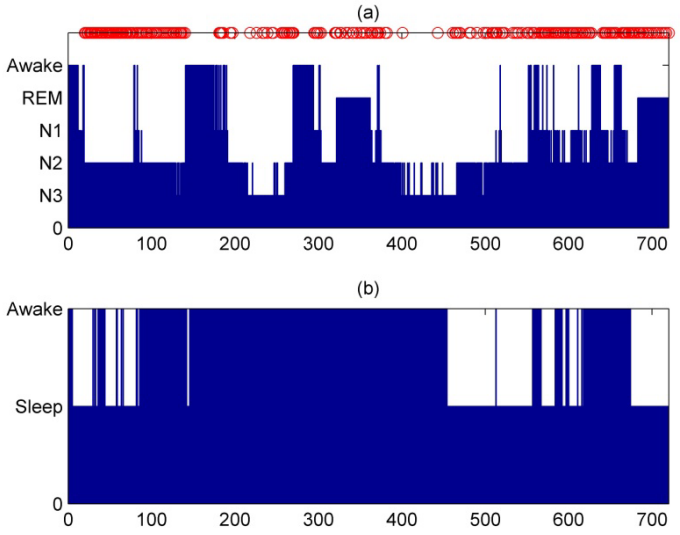

Figure 4.22: Detected awake stages of slp61 in this experiment
(a) Ground truth with red circles as apnea. (b) Outputs of slp61 with the kappa value as 0.09.

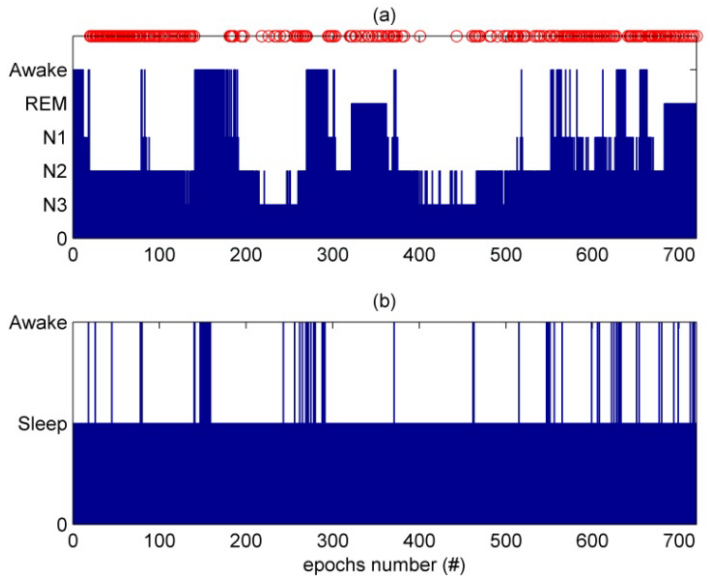

Figure 4.23: Detected awake stages of slp61 in Exp.1 (threshold comparision classifier with RMSSD) (a) Ground truth with red circles as apnea. (b) Outputs of slp61 in Exp. 1 with the kappa value as 0.17.

These variations in performances of different recordings reflected the fact that each recording was so different from the others. From the proportion of each sleep stage to the differences of individuals' physiological indexes and the influences of apnea, how to find some common points and eliminate these differences are essential to improve the classification results.

In addition to the measurements which are used to evaluate the performance of classification itself, sleep quality measures are also important. In the above experiments, relative errors of sleep efficiency (E_{se}) were listed in all tables. But the average performance of E_{se} was not good. The smallest average error obtained was 16.82% in this experiment.

4.3.2. Subject-specific Experiment

The subject-specific scenario was to train on data from one subject and test on data from

the same subject. So, for each recording, a portion of epochs were selected for training and the remaining was used to test.

Only the 25 means of the LFCC features (mLFCC) were used in this experiment. In order to compare the results with [31], the same settings as described in their paper were implemented here. They selected 20% of epochs randomly with appropriate proportion of sleep and awake for training with the remaining 80% as the test set. The classification was repeated 5 times for every record and the results are the average of the 5 measures. The exact same steps were implemented with mLFCC features and the average measurements are reported in Table 4.20. Table 4.21 shows the comparison of the results with the previous works.

Classes: Class 1: REM &NREM Class 2: Awake;

In this experiment, REM&NREM stages were assigned to the positive class in order to keep the same settings as in [31].

Feature Set: Twenty-five mLFCC features

Results:

Table 4.20: Performance measures on Awake detection using mLFCC with the subject-specific scenario

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k	AUC	E_{se} (%)
Slp01a	164.6	2.2	0.8	16.4	90.94	73.33	99.51	90.65	0.20	0.96	8.48
Slp01b	126.4	121.8	21.2	10.6	92.26	85.17	85.76	88.64	0.77	0.95	3.79
Slp02a	199.8	36.8	5.2	38.2	83.95	87.62	97.58	84.50	0.55	0.95	11.79
Slp02b	113.0	71.8	8.2	16.0	87.60	89.75	93.33	88.42	0.76	0.94	3.73
Slp03	407.4	53.2	34.8	46.6	89.74	60.45	92.17	84.98	0.48	0.86	2.18
Slp04	384.8	97.4	24.6	61.2	86.28	79.84	94.09	84.89	0.60	0.94	6.44
Slp14	260.0	211.6	45.4	51.0	83.60	82.33	85.62	83.03	0.66	0.92	0.99
Slp16	249.6	199.8	52.2	52.4	82.65	79.29	82.79	81.12	0.62	0.89	0.04
Slp32	183.0	286.0	22.0	14.0	92.89	92.86	89.30	92.87	0.85	0.97	1.58
Slp37	445.8	46.2	12.8	47.2	90.43	78.31	97.22	89.13	0.55	0.91	6.23
Slp41	368.4	124.0	51.0	72.6	83.54	70.86	88.11	79.94	0.53	0.84	3.51
Slp45	523.6	3.8	1.2	71.4	88.00	76.00	99.77	87.90	0.08	0.99	11.70
Slp48	402.6	135.8	34.2	27.4	93.63	79.88	92.19	89.73	0.74	0.95	1.13
Slp59	211.2	84.2	26.8	42.8	83.15	75.86	88.93	80.93	0.57	0.86	4.38
Slp60	247.8	192.8	30.2	74.2	76.96	86.46	89.11	80.84	0.62	0.91	8.07
Slp61	404.2	70.6	27.4	72.8	84.74	72.04	93.65	82.57	0.48	0.90	7.90
Slp66	178.4	112.4	21.6	32.6	84.55	83.88	89.40	84.29	0.67	0.91	3.19
Slp67x	49.8	37.8	13.2	15.2	76.62	74.12	79.55	75.52	0.51	0.83	1.72
Mean					86.19	79.34	91.00	85.00	0.57	0.92	4.80

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy), k (Kappa), AUC(Area under the curve), E_{se} (Absolute error of sleep efficiency)

Table 4.21: Comparison of results of Awake detection using subject-specific scheme with previous research

Author/year	Features	ACC (%)	Sp(%)	Pr(%)	k	*RE_{se}(%)
Adnane, M. et al. 2011 [31]	HRV+DFA+WDFA	79.99	85.27	82.99	0.43	4.64
Werteni, H. et al., 2014 [45]	HRV+DFA	78.33	/	/	/	/
Proposed classifier	LFCC	85.00	86.19	91.00	0.57	6.45

* REse here is the relative absolute error which computed by: REse= |SE-SE'|/SE

Except the relative error of SE, other measures all outperformed previous methods. In

addition, the LFCC used in this work could be calculated with simple steps compared with HRV and DFA and no correction process was needed to adjust the detected heart beat. This experiment again showed the potential of LFCC features applied to sleep stage recognition problems.

4.4. Sleep-EDF Database (Expanded)

The threshold comparison method was applied to the sleep-EDF database. A two-layer system was implemented. The first layer was the Awake&REM detection. In this layer, REM and Awake were treated as the same class. The second layer was to separate REM and Awake based on the outputs from layer one. Two features were selected: the respiratory rate (RR) and the maximum absolute differences of intervals (MADI). The theory behind the algorithm is that breathing would more likely to be irregular during Awake and REM stage.

Three experiments are detailed described in this section. The first two experiments implemented two steps of the first layer of the algorithm. The last experiment implemented the second layer of the algorithm.

- **Experiment 1: Awake&REM detection using detrended respiratory rate**

In the first layer, only the respiratory rate was used to detect REM and Awake. For each night, feature detrending was applied to remove the nonlinear trends from the respiratory rates. Figure 4.24 shows the box plots of the detrended RR for all recordings. On each box,

the central mark is the median, the edges of the box are the 25th(Q_1) and 75th (Q_3) percentiles. The red crosses indicate the outliers. Points are drawn as outliers if they are larger than Q_3+1.5(Q_3-Q_1) or smaller than Q_1-1.5(Q_3-Q_1).

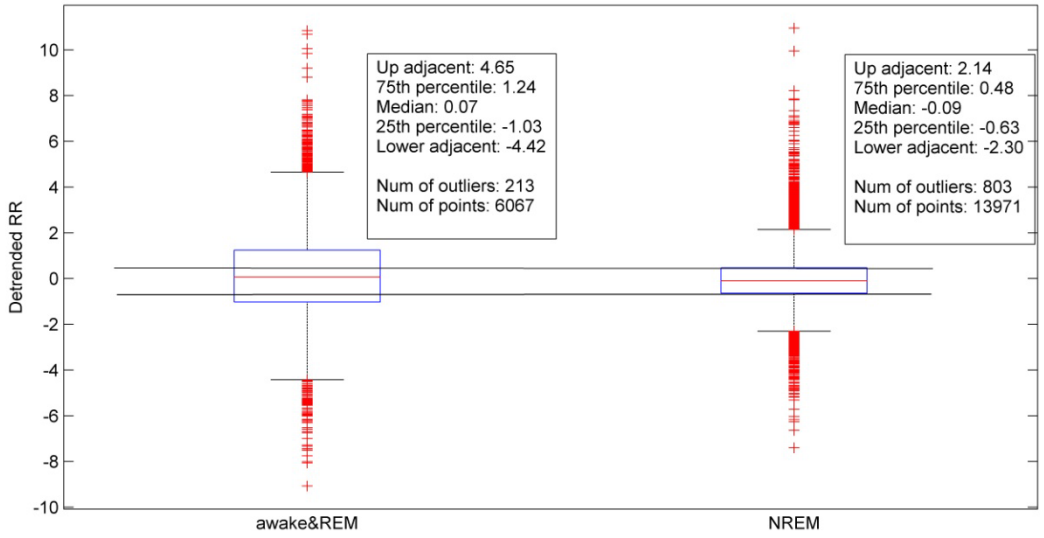

Figure 4.24: The box plots of the detrended respiratory rate in two classes.
Red crosses indicate the outliers. Text boxs diaplay the information of two box plots. Awake&REM class on the left and NREM class on the right. The NREM class has a much narrower interquartile range than the Awake&REM class. Two black lines compare the interquartile range of detrended RR in NREM to that in Awake&REM class.

Although according to the box plots, two classes can't be simply separated, the values of the NREM epochs are gathered in a smaller range than those of the Awake&REM epochs. So the idea was to separate part of these two classes by setting two thresholds like the black lines in the figure. For the detrended RR in each recording, thresholds were set as the 75th percentile value plus the standard deviation and the 25th percentile value minus the standard deviation. So all epochs had the detrended RR value out of range of these two thresholds were estimated as Awake&REM class. Figure 4.25 displays the original respiratory rate, the detrended feature with thresholds, and the hypnogram with detected

epochs.

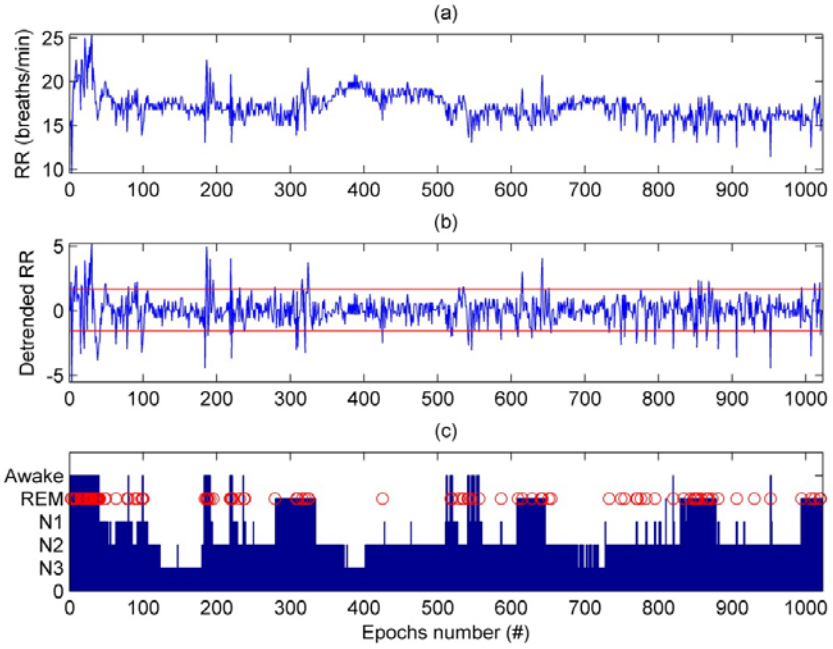

Figure 4.25: Process of Awake&REM detection.

(a) The orginal respiratory rate. (b) The detrended respiratory rate. Two red lines indicate two thresholds. (c) The corresponding hyponogram with the red circles as the detected epochs.

Figure 4.25(c) shows that for one block of Awake or REM, the detected epochs were scattered across that block. This result is consistent with the box plot in figure 4.24 in which there were many epochs having the values in the same range of NREM. The results for all recordings are described in Table 4.22.

Classes: Class 1: Awake &REM; Class 2: NREM

Feature Set: Detrended respiratory rate

Method: The first threshold was the 75th percentile value plus the standard deviation; the second threshold was the 25th percentile value minus the standard deviation. If the value of one epoch was bigger than the first threshold or smaller than the second threshold, this epoch was classified to Awake&REM class.

Results:

Table 4.22: Performance measures on Awake&REM detection using detrended RR with the sleep-EDF

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k
SC4001	68	498	30	164	29.31	94.32	69.39	74.47	0.28
SC4002	84	689	40	233	26.50	94.51	67.74	73.90	0.25
SC4011	70	743	33	176	28.46	95.75	67.96	79.55	0.30
SC4012	71	819	29	185	27.73	96.58	71.00	80.62	0.31
SC4031	68	557	44	200	25.37	92.68	60.71	71.92	0.22
SC4041	90	783	56	225	28.57	93.33	61.64	75.65	0.26
SC4042	87	688	57	287	23.26	92.35	60.42	69.26	0.18
SC4061	41	558	41	122	25.15	93.16	50.00	78.61	0.22
SC4062	61	603	33	238	20.40	94.81	64.89	71.02	0.19
SC4071	62	620	34	178	25.83	94.80	64.58	76.29	0.25
SC4101	74	687	55	207	26.33	92.59	57.36	74.39	0.23
SC4102	65	703	46	197	24.81	93.86	58.56	75.96	0.23
SC4121	70	589	29	277	20.17	95.31	70.71	68.29	0.18
SC4122	81	472	16	327	19.85	96.72	83.51	61.72	0.18
SC4131	72	660	41	173	29.39	94.15	63.72	77.38	0.29
SC4141	77	560	24	262	22.71	95.89	76.24	69.01	0.22
SC4142	62	530	28	251	19.81	94.98	68.89	67.97	0.18
SC4151	75	553	19	224	25.08	96.68	79.79	72.10	0.26
SC4161	83	623	45	312	21.01	93.26	64.84	66.42	0.17
SC4162	56	565	64	237	19.11	89.83	46.67	67.35	0.11
SC4181	43	634	73	132	24.57	89.67	37.07	76.76	0.16
Total					24.45	94.06	64.08	72.79	0.22

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy) ,k (Kappa)

The result shows only 24.45% Awake&REM epochs were successfully detected. It was not a high value, but coupled with the specificity, 94.06%, the method mixed only 6% NREM with the other class. In addition, from figure 4.25, we can find that the rough positions of Awake&REM were detected. The only problem was that the detected epochs were dispersed. So, a post-processing step was designed and described in the next experiment.

- **Experiment 2: Awake&REM detection using detrended respiratory rate with a post-processing step**

Two rules were designed based on the observed results of Exp.1. They are:

1) For all detected Awake&REM epochs in the first layer, if two successive epochs were less than or equal to 15 epochs apart, then all epochs between these two epochs were assigned as the target class.
2) Because the sleep always starts with Awake, if the first detected Awake&REM epoch was not the first epoch in the recording, the epochs from the first epoch to the position of the first detected epoch were all assigned as the target class.

The role of the first rule was to connect these scattered detected epochs. Not only does figure 4.25 suggest this process, characteristics of REM sleep also indicate such a reassignment (REM periods usually last for a long time (30 minutes) [18]). During a long period of REM, respiration activity would not always be irregular, so this connection scheme will link these periods to irregular periods to form a complete block. Figure 4.26 gives the comparison of the original detected results and the post processed results. The results demonstrate that by connecting the scattered epochs, most of the REM and Awake stages were detected. Although the connection may also increase the number of false detections, a balance can be found (15 epochs in this system) to make both sides look good.

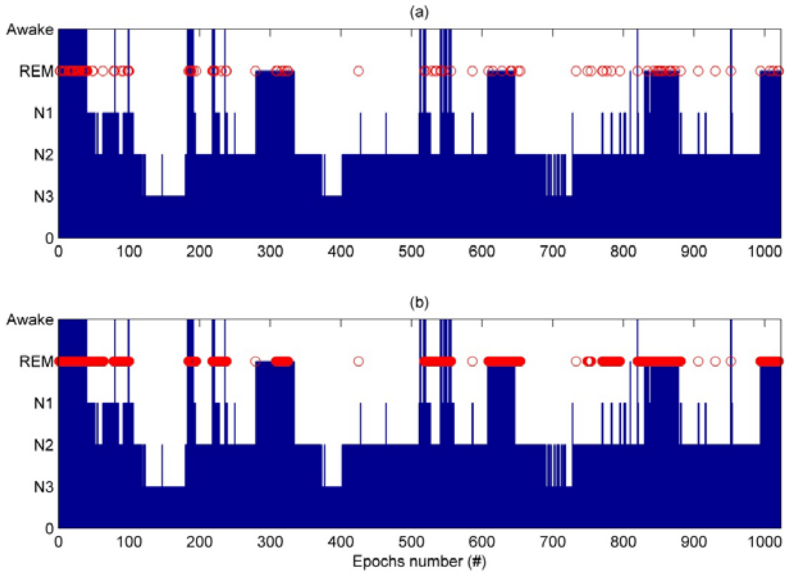

Figure 4.26: An example of the original detected results and the post processed results (a) hypnogram with red circles as target epcohs without post-processing; (b) hypnogram with red circles as target epochs after post-processing.

Table 4.23 displays the results of all recordings after post-processing.

Classes: Class 1: Awake &REM; Class 2: NREM

Feature Set: Detrended respiratory rate

Method: A post-processing step with two rules was applied to the outputs of Exp.1.

Table 4.23: Performance measures on Awake&REM detection using detrended RR with the sleep-EDF after a post-processing step

	TP	TN	FP	FN	Se (%)	Sp (%)	Pr (%)	Acc (%)	k
SC4001	217	421	107	15	93.53	79.73	66.98	83.95	0.66
SC4002	269	591	138	48	84.86	81.07	66.09	82.22	0.61
SC4011	199	625	151	47	80.89	80.54	56.86	80.63	0.54
SC4012	171	769	79	85	66.80	90.68	68.40	85.14	0.58
SC4031	247	484	117	21	92.16	80.53	67.86	84.12	0.66
SC4041	247	636	203	68	78.41	75.80	54.89	76.52	0.48
SC4042	288	507	238	86	77.01	68.05	54.75	71.05	0.41
SC4061	149	441	158	14	91.41	73.62	48.53	77.43	0.49
SC4062	203	518	118	96	67.89	81.45	63.24	77.11	0.48
SC4071	195	539	115	45	81.25	82.42	62.90	82.10	0.58
SC4101	199	532	210	82	70.82	71.70	48.66	71.46	0.37
SC4102	185	565	184	77	70.61	75.43	50.14	74.18	0.41
SC4121	227	481	137	120	65.42	77.83	62.36	73.37	0.43
SC4122	272	431	57	136	66.67	88.32	82.67	78.46	0.56
SC4131	202	543	158	43	82.45	77.46	56.11	78.75	0.52
SC4141	237	494	90	102	69.91	84.59	72.48	79.20	0.55
SC4142	190	428	130	123	60.70	76.70	59.38	70.95	0.37
SC4151	203	497	75	96	67.89	86.89	73.02	80.37	0.56
SC4161	279	500	168	116	70.63	74.85	62.42	73.28	0.44
SC4162	206	388	241	87	70.31	61.69	46.09	64.43	0.28
SC4181	154	493	214	21	88.00	69.73	41.85	73.36	0.41
Total					76.08	78.05	60.27	77.05	0.49

* TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative), Se(Sensitivity), Sp(Specificity), Pr(Precision), Acc(Accuracy) ,k (Kappa)

The sensitivity increased about 50% while only losing 16% of the specificity. The kappa value of 0.49 was also impressive since only the respiratory rate was used as a feature. Among 21 recordings, only 3 has a kappa value below 0.41, which means that most of the recordings were in the moderate or substantial agreement range. The variation of the results were also smaller than that in the MITBPD. The connection process was thus shown to be efficient.

However, one of the shortcomings of this method was that the threshold for the connection rule was handpicked in order to balance the sensitivity and specificity for this particular database. So the constant threshold does not work well with all recordings. Figure 4.27 displays the outputs of subject SC4162 where the specificity is only 60%.

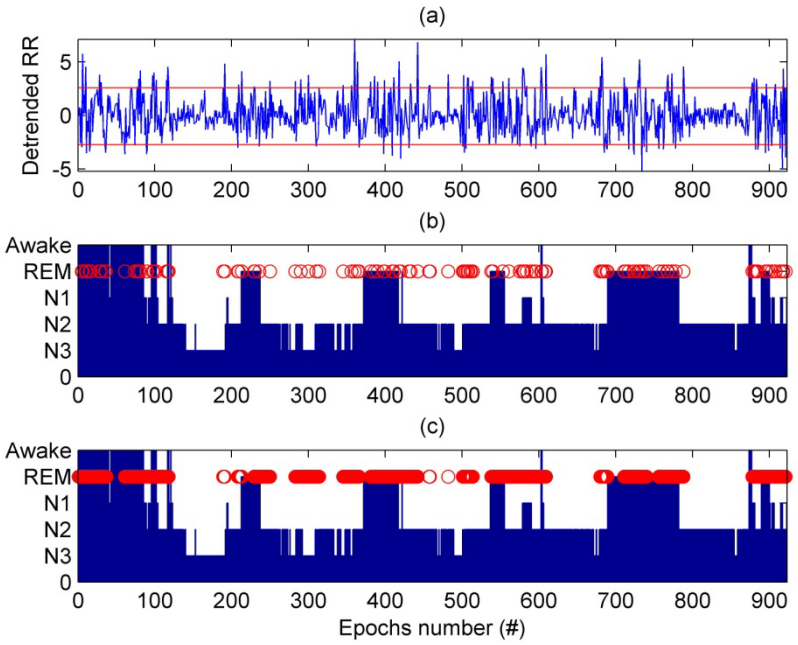

Figure 4.27: Original detected results and the connection processed results of SC4162

(a) The processed respiratory rate. Two red lines indicate two thresholds.;(b) hypnogram with red circles as target epochs before connection. Some NREM epochs were detected as Awake &REM such as epochs around 300 and 500;(b) hypnogram with red circle as target epochs after connection. The scattered misdetected epochs were also connected to large blocks.

It's clear that the variability of respiratory rate was high in some of the NREM periods. So these NREM epochs were misclassified to Awake&REM in Exp.1 and further formed larger blocks of errors after post-processing. Hence, for recordings that had less variability between NREM and other stages, a smaller connection threshold might be more reasonable. Hence, a further improvement can be done for this method by finding a way to pick a dynamic threshold for the connection process.

- **Experiment 3: separating REM and Awake using MADI**

The second layer aimed at separating REM and Awake stage. It was hard to separate these two stages because they had similar physiological indexes, such as irregular respiration. One assumption was that if a person woke up in the middle of the night suddenly, there might be a more severe fluctuation in these indexes, for example, a suddenly shortened breath interval or a deeper breath. This was the reason why the maximum absolute differences of intervals (MADI) was selected as the feature for this layer. This feature measured the biggest change of the breath intervals in each epoch. Because the breath-to-breath intervals were integers bigger than 2 seconds, MADI were also integers. Figure 4.28 displays the box plots of this feature for all recordings.

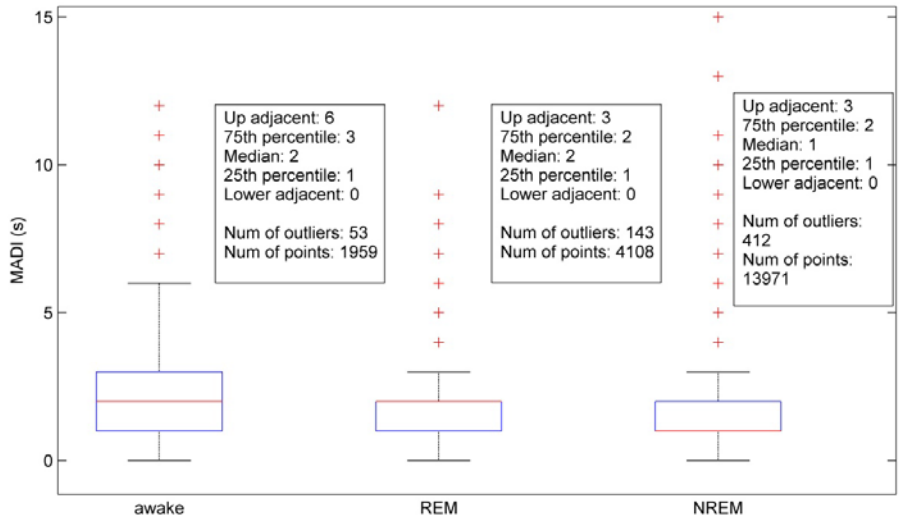

Figure 4.28: Box plots of MADI in three classes

The plots have the same settings as previous box plots. Red crosses indicate the outliers. Text boxes diaplay the information of three box plots. Awake class on the left, REM class on the middle and NREM class on the right. The 75th percentile of Awake equals to the the up adjacents of REM and NREM.

Although the feature couldn't completely separate the Awake from REM, some of the Awake epochs can be detected using this feature. Additionally, two other rules were

applied on the outputs of the first layer. So there were totally three rules:

1) If the detected Awake&REM epochs in the first layer occur in the first 60 minutes (120 epochs), these epochs were assigned as Awake.
2) If the duration of a block of continuous detected Awake&REM epochs in the first layer was less than 5 minutes (10 epochs), these epochs were assigned as Awake.
3) If the MADI was larger than 4, the epoch was assigned as Awake.

The first two rules were set according to sleep theory: 1) the first REM period usually occurs about 70 minutes after sleep onset; 2) the first REM period is short, but the duration of REM period after that is approximately 30 minutes. Hence, a five minute period can't be REM. The threshold for the third rule was set a little higher than the 75th percentile value (which was 3) for the Awake class.

The three class confusion matrix of all 21 recordings is displayed in Table 4.24. Table 4.25 lists the relative errors of the three sleep quality measures: sleep efficiency (SE), sleep onset (SL) and percentage of REM stages (%SR). It was interesting to find that the average absolute error of the sleep efficiency was only 3.61% while the error of %SR was 9.75%. The E_{sl} had a large range from the smallest error of 0.5 min (SC4001 and SC4061) to the largest one of 24 min (SC4162).

Classes: Class 1: NREM; Class 2: REM; Class 3: awake

Feature Set: Detrended respiratory rate and MADI

Method: Two layers threshold comparison classifier

Results:

Table 4.24: Confusion matrix of overall results of three stage classification with the sleep-EDF

Actual / Output	NREM	REM	Awake
NREM	10826	1065	408
REM	2405	2960	521
Awake	740	83	1030

ACC =74.00%±5.30%, Kappa=0.49±0.08

Table 4.25: Sleep quality performance of three stage classification with the sleep-EDF

	E_{sl} (Min)	E_{se} (%)	E_{re} (%)
SC4001	0.5	5.92	18.82
SC4002	5.5	4.78	13.34
SC4011	12.0	3.23	8.54
SC4012	2.0	0.91	0.42
SC4031	4.0	4.72	9.70
SC4041	9.5	0.78	13.32
SC4042	8.5	1.16	16.40
SC4061	0.5	6.69	15.91
SC4062	15.0	5.03	7.39
SC4071	6.0	4.70	5.40
SC4101	7.0	1.08	13.37
SC4102	2.5	1.68	10.91
SC4121	7.0	0.73	2.54
SC4122	3.0	16.63	3.81
SC4131	4.5	1.16	12.78
SC4141	3.0	2.28	1.43
SC4142	17.5	0.46	1.79
SC4151	14.0	1.03	1.57
SC4161	7.0	4.80	9.37
SC4162	24.0	1.52	18.06
SC4181	4.5	6.46	19.81
Mean	7.5	3.61	9.75

The confusion matrix shows that only half of the Awake stages were detected, but the kappa value was high at 0.49, better than the previous studies with this data [11],[31]. The

result also similar to some other work [46],[47], but with a more simple feature set and method. Figure 4.29 and 4.30 display the outputs of the recordings with the worst (kappa=0.36) and best results (kappa=0.67).

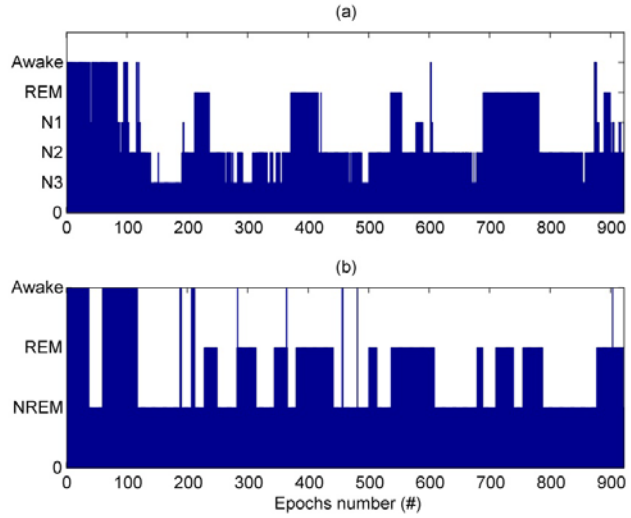

Figure 4.29: The outputs of the recording had the worst results (SC4162)
(a) Ground truth. (b) output of Exp.3, kappa=0.36

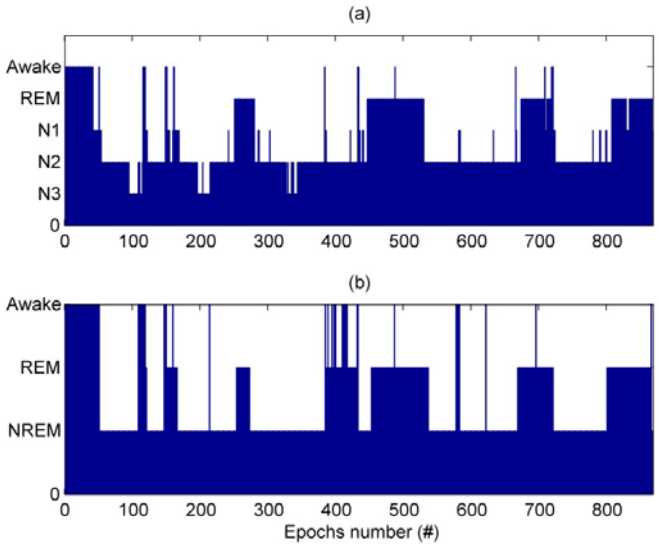

Figure 4.30: The outputs of the recording had the best results (SC4031)
(a) Ground truth. (b) output of Exp.3, kappa=0.67

Although the kappa value for the first recording was only about 0.35, the rough position of

REM&awake stages were detected. This subject also had the biggest error of SL which was 24 min. Through figure 4.30(b), it can be found that most of the awake epochs before sleep onset (around epoch 100) were detected. But there was a gap around epoch 50 which cut off the continuous Awake period and shorten the estimated sleep onset. If you look at figure 4.30 (a) carefully, there actually were one or two epochs in the N1 stage around epoch 50. But the actual gap was less than 3 minutes so these epochs were not defined as sleep onset. This could imply that the short transit from awake to N1 was captured by the proposed method, but the transit from N1 to awake wasn't captured immediately. It could be the delay of the body or the transit didn't make any big impact on respiration.

The second figure shows the output of the best results. Except the epochs around 400, most of other epochs were assigned to the right class with minor errors. Some epochs around 600 were misclassified to awake from N1 stage. This kind of error also happened in other recordings very often. N1 is the lightest sleep stage compared with other stages. It usually happens in the beginning of the night or between awake periods, which can be understood as a transition stage from awake to sleep. So it has similar physiological indexes to awake. This could be a reason that the epochs in N1 were classified as awake or REM very often.

5. Discussion and Conclusions

5.1. Discussion About Sleep Stage Recognition Problem

The three databases studied in this work have covered most types of the signals used to solve the sleep stages recognition problem. The BCG signal provides a noninvasive characteristic and the ability to extract heartbeat, respiration and body movement information; the ECG signal gives a more reliable HBI and thus more accurate HRV parameters; respiration is easy to process and to detect peaks and troughs. But they also had some limits. The BCG signal was sensitive to motion and the signal itself had some uncertainties such as the shape of the waveform. Thus, the detected heart beats and breath cycles may not be as accurate as with the ECG. The ECG and respiration signals had more accuracy in this regard, but the sensors need to be connected to the body.

The results of three databases all showed both potential and limitations of automatic sleep stage recognition. Although the bed sensor dataset with BCG signals had the worst performance in this study, in my opinion, it is actually the most promising sensor for sleep stage recognition. Its ability of tracking both cardiac and respiratory activity is very attractive. The studies of the MITBPD and the sleep-EDF have indicated HRV and RV features were useful in separating sleep stages. Whether the HRV parameters of healthy people are also useful cannot be verified in this study, although plenty of previous studies

have demonstrated it. The combination of HRV and RV parameters can get both cardiac and respiratory information and improve performance. Moreover, body movements calculated from the BCG would also be useful to detect different sleep stages, especially REM and awake. These two stages have highly similar patterns in both cardiac and respiratory activity, the only difference is REM sleep has a low muscle tone [4]. So, less body movements should be observed in this stage. In addition, it is possible that LFCC features can extract some useful information from BCG and ECG signals. Thus, the BCG is a very rich signal which contains different types of the information.

However, to obtain accurate HRV parameters from BCG signals might be difficult, especially for very short 30 second time periods. A very accurate peak detection algorithm is needed. Reference [48] indicated that the outliers due to missed or false beat detection can lead to greater than 1000% error for frequency domain measures. Most of the time domain measures also can cause error of greater than 100%. These conclusions were based on a 24-hour data set. Thus, for a 30 seconds time period, the error would be more serious. Another challenge for the sleep stages recognition problem is the imbalance of the data. It is more than the fact that the three stages have different proportions; it is that each test recording has different proportions of the three stages., resulting in a difficulty of evaluating and comparing results. A method developed based on one set of recordings may not be suitable for another set that has different proportions of sleep stages. For example, the REM class in MITBPD was put together with NREM and it seemed not to lead a big

problem for Awake detection. But the REM and Awake couldn't be separated well in the sleep-EDF database. One of the reasons could be that apnea affected the original characteristic of REM. But of most importance was that REM only accounted for 7% in this dataset. So even if REM were misclassified to Awake, it would not have a big impact on the overall results.

But the most challenging part of this problem is the problem itself. Sleep stages are defined primarily by EEG and assisted by EMG and EOG, but now we want to classify these stages by cardiac, respiratory and body movement information. This is like wanting to separate females and males by their weights and heights. It may be hard to find a perfect solution. However, there is still much room for improvement of current methods, such as how to eliminate the differences between individuals and how to effectively combine different types of features.

In addition, if the heart rate and body movement information can be added to the sleep-EDF database, the REM and awake stages should be better separated. The mean respiration rate and mean heart rate are also more reliable from the bed sensor. If a reliable ground truth can be obtained, acceptable results should be achieved with the bed sensor. One possible approach based on the method used in sleep-EDF database is describe as follows.

1) Use the threshold comparison classifier with mean respiratory rate as in sleep-EDF database to detect possible REM and awake periods without any connection processing.

2) All epochs that have more than 15s body movements are assigned as Awake.

3) Detect spikes in mHBI or LFCC features (if the relation can be verified in BCG) and assign these epochs as Awake.

4) Assign outputs of 1) to awake if the epoch has body movement.

5) The remaining outputs of 1) are considered as REM. Connect these epochs if there are no detected awake epochs within some range.

Although it is a very straightforward process without any complex classifiers, it may give a good performance based on the results of sleep-EDF.

5.2. Conclusions

This work studied sleep stage recognition problems on three datasets. The bed sensor dataset in which the BCG signal was used, the MIT-BIH Polysomnographic Database in which a single-lead ECG signal was studied, and the Sleep-EDF Database in which the respiration signal was employed. All three datasets were pre-processed by certain rules built for the different situations. Heart beat intervals (HBI) were calculated in the bed sensor dataset and MITBPD. Breath intervals were calculated in bed sensor dataset and Sleep-EDF. Then HRV and RV were calculated based on heart beat intervals and breath intervals, respectively. The LFCC features were calculated from the BCG_{hr} and filtered ECG signals. Two types of processes: smoothing and detrending were applied on some of the features. An SVM classifier was used as the main classifier and grid search method was

implemented for model selection. In addition, a threshold comparison method was also tested on the MITBPD and Sleep-EDF database. Results of experiments on these three datasets were reported.

A large differences of the performance on the bed sensor dataset was observed with different training strategies (put-all-recordings-together and leave-one-night-out). For REM detection, the best accuracy of two strategies was 93.16% (k=0.78) and 62.75% (k=0.09), respectively. The accuracy of the three stages classification was 81.69% (k=0.71) with put-all-recordings-together. The reason of the big differences of two strategies and the bad performances with the leave-one-night-out strategy were discussed. The major reason points at unreliable ground truth.

Two types of experiments were implemented for the MITBPD. They were subject-independent and subject-specific schemes. The studies were focused on awake detection. For the subject-independent scheme, the best accuracy was 74.94% (k=0.44) with smoothed HRV and smoothed LFCC features. A decision boundary adjustment process was applied, and additional awake stages detected by a threshold comparison method were also added to the results. The obtained result was better than that reported in [31]. In the threshold comparison classifier, HF and RMSSD were tested with hard thresholds. The experiment demonstrated a higher value of RMSSD in NREM than that in Awake stages which was reversed from what was observed in previous research. Further analyses showed that apnea raises the variability of heart rate. Also, a three stage

classification experiment was conducted and the obtained accuracy was 65.97% (k=0.37). For the subject-specific scheme, accuracy of 85.00% (k=0.57) was obtained which was also better than results reported in [31] and [45]. For this dataset, the relation between mHBI and Awake stages were also analyzed. In addition, using LFCC features improved the results. The mLFCC1 feature also displayed a potential relation with the Awake stage. Finally, the Sleep-EDF database was studied. A two-layer classification system was developed. The first layer was used to separate NREM from other stages. Only respiratory rate was used with the threshold comparison classifier and a connection step as post-processing. The connection process improved the results from 72.79% (k=0.22) to 77.05 (k=0.49). The second layer aimed to separate REM and awake. The MADI feature and other two rules were set for this mission. The method achieved the accuracy as 74% (k=0.49) which was better or similar to some of the previous work [11],[46],[47]. Three sleep quality measures were also computed. The relative error of sleep efficiency, percentage of REM stages and sleep onset were 3.61%, 9.75% and 7.5 min, respectively.

5.3. Future works

1) Since classifying sleep stages with a bed sensor was the original goal, the most important thing that needs to be done is to collect a dataset with the bed sensor and a reliable ground truth. Then all the proposed methods can be tested on the BCG signal and applied to the actual application.

2) There are still places for improvement. Some experiments that can be run are:

- For the MITBPD, instead of putting all types of features in one classifier, use multiple classifiers with different feature sets and combine them.
- For the sleep-EDF database, instead of the thresholding with a single value, some adaptive thresholds can be designed.
- Other types of classifiers such as HMM can be applied.
- Try to minimize the influence caused by physiological differences among different people.

References

[1] Garcia-Molina, G., Abtahi, F., & Lagares-Lemos, M. (2012, August). Automated NREM sleep staging using the Electro-oculogram: A pilot study. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (pp. 2255-2258).

[2] Hansen, I. H., Marcussen, M., Christensen, J., Jennum, P., & Sorensen, H. B. D. (2013, July). Detection of a sleep disorder predicting Parkinson's disease. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (pp. 5793-5796).

[3] Alzheimer's Disease and Sleep, sleepfoundation.org, retrieved from https://sleepfoundation.org/sleep-disorders-problems/alzheimers-disease-and-sleep.

[4] Iber, C; Ancoli-Israel, S; Chesson, A; Quan, SF for the American Academy of Sleep Medicine (2007). The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Westchester: American Academy of Sleep Medicine.

[5] Yilmaz, B., Asyali, M. H., Arikan, E., Yetkin, S., & Özgen, F. (2010). Sleep stage and obstructive apneaic epoch classification using single-lead ECG. *Biomedical Engineering Online*, 9(1), 39-39.

[6] Heise, D., & Skubic, M. (2010, August). Monitoring pulse and respiration with a non-invasive hydraulic bed sensor. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE* (pp. 2119-2123).

[7] Huang, C. S., Lin, C. L., Ko, L. W., Liu, S. Y., Sua, T. P., & Lin, C. T. (2013, April). A hierarchical classification system for sleep stage scoring via forehead EEG signals. In *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2013 IEEE Symposium on* (pp. 1-5).

[8] Rosales, L., Skubic, M., Heise, D., Devaney, M. J., & Schaumburg, M. (2012, August). Heartbeat detection from a hydraulic bed sensor using a clustering approach. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (pp. 2383-2387).

[9] Kortelainen, J. M., Mendez, M. O., Bianchi, A. M., Matteucci, M., & Cerutti, S. (2010). Sleep staging based on signals acquired through bed sensor. *IEEE Transactions on Information Technology in Biomedicine*, 14(3), 776-785.

[10] Park, K. S., Hwang, S. H., Yoon, H. N., & Lee, W. K. (2014, August). Ballistocardiography for nonintrusive sleep structure estimation. In*Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE* (pp. 5184-5187).

[11] Mendez, M. O., Matteucci, M., Cerutti, S., Bianchi, A. M., & Kortelainen, J. M. (2009, September). Automatic detection of sleep macrostructure based on bed sensors. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE* (pp. 5555-5558).

[12] Tataraidze, A., Anishchenko, L., Korostovtseva, L., Kooij, B. J., Bochkarev, M., & Sviryaev, Y. (2015, August). Sleep stage classification based on respiratory signal. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (pp. 358-361).

[13] Ichimaru, Y., & Moody, G. B. (1999). Development of the polysomnographic database on CD‐ROM. *Psychiatry and Clinical Neurosciences, 53*(2), 175-177.

[14] Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A., & Oberyé, J. J. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *Biomedical Engineering, IEEE Transactions on, 47*(9), 1185-1194.

[15] Rechtschaffen, A., & Kales, A. (1968). A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Bethesda, Md: U.S. Dept. of Health, Education, and Welfare.

[16] Silber, M. H., Ancoli-Israel, S., Bonnet, M. H., Chokroverty, S., Grigg-Damberger, M. M., Hirshkowitz, M.. . Iber, C. (2007). The visual scoring of sleep in adults. *Journal of Clinical Sleep Medicine*, 3(2), 121-131.

[17] Khemiri, S., Aloui, K., & Naceur, M. S. (2011, March). Automatic detection of slow-wave sleep and REM-sleep stages using polysomnographic ECG signals. In *Systems, Signals and Devices (SSD), 2011 8th International Multi-Conference on* (pp. 1-4).

[18] McCarley, R. W. (2007). Neurobiology of REM and NREM sleep. *Sleep Medicine*, 8(4), 302-330.

[19] Hori, T., Sugita, Y., Koga, E., Shirakawa, S., Inoue, K., Uchida, S.. . Sleep Computing Committee of the Japanese Society of Sleep Research Society. (2001).

Proposed supplements and amendments to ‘A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects’, the rechtschaffen & kales (1968) standard. *Psychiatry and Clinical Neurosciences*, 55(3), 305-310.

[20] Yao, Y., Bruser, C., Pietrzyk, U., Leonhardt, S., van Waasen, S., & Schiek, M. (2014). Model-based verification of a non-linear separation scheme for ballistocardiography. *IEEE Journal of Biomedical and Health Informatics,* 18(1), 174-182.

[21] Pinheiro, E., Postolache, O., & Girão, P. (2010). Theory and developments in an unobtrusive cardiovascular system representation: Ballistocardiography. *Open Biomedical Engineering Journal*,4(2), 201-216.

[22] Starr, I., Rawson, A. J., Schroeder, H. A., & Joseph, N. R. (1939). Studies on the estimation of cardiac ouptut in man, and of abnormalities in cardiac function, from the heart's recoil and the blood's impacts; the ballistocardiogram. *American Journal of Physiology--Legacy Content*, *127*(1), 1-28.

[23] BILCHICK, K. C., & BERGER, R. D. (2006). Heart rate variability. *Journal of Cardiovascular Electrophysiology*, 17(6), 691-694.

[24] Carnethon, M. R., & Craft, L. L. (2008). Autonomic regulation of the association between exercise and diabetes. *Exercise and Sport Sciences Reviews*, 36(1), 12-18.

[25] Vaughn, B. V., Quint, S. R., Messenheimer, J. A., & Robertson, K. R. (1995). Heart period variability in sleep. *Electroencephalography and Clinical Neurophysiology*, 94(3), 155-162.

[26] Malik, M. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5), 1043-1065.

[27] Bonnet, M. H., & Arand, D. L. (1997). Heart rate variability: Sleep stage, time of night, and arousal influences. *Electroencephalography and Clinical Neurophysiology*, 102(5), 390-396.

[28] Akselrod, S., Gordon, D., Ubel, F. A., Shannon, D. C., Barger, A. C., & Cohen, R. J. (1981). Power spectrum analysis of heart rate fluctuation: A quantitative probe of beat-to-beat cardiovascular control. *Science*, 213(4504), 220-222.

[29] Barett, K. E., Barman, S. M., Boitano, S., & Brooks, H. L. (2010). "Pulmonary Function", *Ganong’s review of medical physiology.*

[30] Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5), 360-363.

[31] Adnane, M., Jiang, Z., & Yan, Z. (2012;2011;). Sleep-wake stages classification and sleep efficiency estimation using single-lead electrocardiogram. *Expert Systems with Applications*, 39(1), 1401-1413.

[32] Mindo, National Chiao Tung University Brain Research Center, Taiwan. See more at: http://mindo.com.tw/en/index.php

[33] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215-e220.

[34] Mourtazaev, M. S., Kemp, B., Zwinderman, A. H., & Kamphuisen, H. A. C. (1995). Age and gender affect different characteristics of slow waves in the sleep EEG. *SLEEP-NEW YORK-*, *18*, 557-557.

[35] Kemp, B. (1987). *Model-based monitoring of human sleep stages*. Universiteit Twente.

[36] Lydon, K., Su, B. Y., Rosales, L., Enayati, M., Ho, K. C., Rantz, M., & Skubic, M. (2015, August). Robust heartbeat detection from in-home ballistocardiogram signals of older adults using a bed sensor. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (pp. 7175-7179).

[37] Venkatachalam, K. L., Herbr, J. E., Herbrandson, J. E., son, & Asirvatham, S. J (2011). Signals and signal processing for the electrophysiologist: part I: electrogram acquisition Circulation. *Arrhythmia And Electrophysiology*, 4(6), 965-73. - See more at: http://www.ems12lead.com/2014/03/10/understanding-ecg-filtering/#sthash.5aybjOCT.dpuf.

[38] De Boor, C. (1978). A practical guide to splines. *Mathematics of Computation*.

[39] James Lyons. Mel Frequency Cepstral Coefficient (MFCC) tutorial. Retrieved from http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

[40] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368), 829-836.

[41]Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*,*2*(3), 27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[42] Chapelle, O., & Zien, A. (2005, January). Semi-supervised classification by low density separation. In *Proceedings of the tenth international workshop on artificial intelligence and statistics* (Vol. 1, pp. 57-64).

[43] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

[44] Penzel, T., Kantelhardt, J. W., Grote, L., Peter, J. H., & Bunde, A. (2003). Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *IEEE Transactions on Biomedical Engineering*, 50(10), 1143-1151.

[45] Werteni, H., Yacoub, S., & Ellouze, N. (2014). An automatic sleep-wake classifier using ECG signals. *International Journal of Computer Science Issues (IJCSI)*, 11(4), 84.

[46] Long, X., Yang, J., Weysen, T., Haakma, R., Foussier, J., Fonseca, P., & Aarts, R. M. (2014). Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging. *Physiological measurement*,*35*(12), 2529.

[47] Kurihara, Y., & Watanabe, K. (2012). Sleep-stage decision algorithm by using heartbeat and body-movement signals. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *42*(6), 1450-1459.

[48] Joseph E. Mietus (2006). Time domain measures: from variance to pNNx [PDF document]. Retrieved from https://physionet.org/events/hrv-2006/mietus-1.pdf.