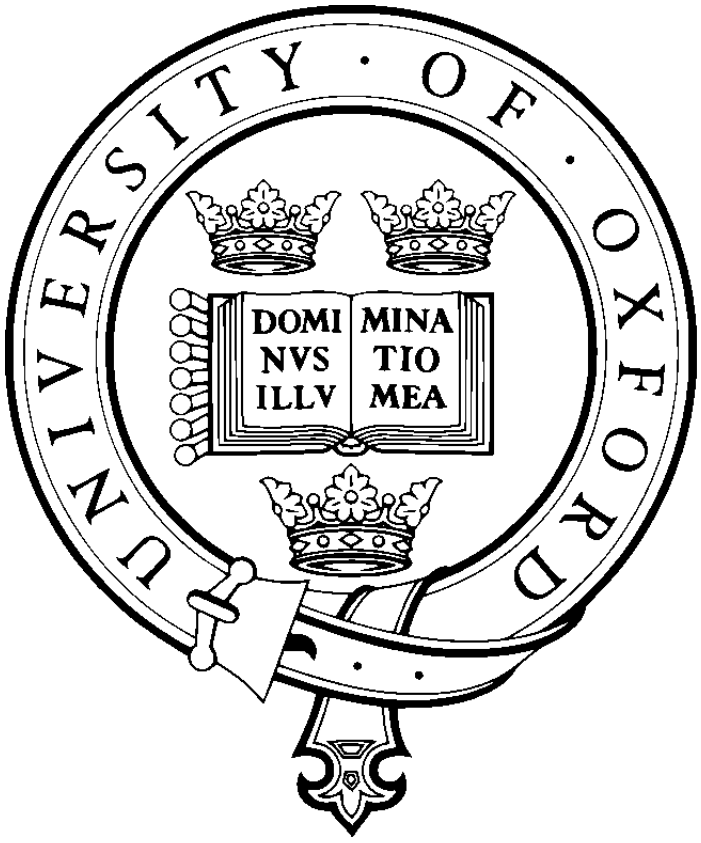
Comparative analysis of polysomnographic signals for classifying Obstructive Sleep Apnoea



Aoife Roebuck

Kellogg College

University of Oxford

Supervised by
Dr Gari Clifford

Submitted: Trinity Term,

April 14, 2015

Abstract

Comparative analysis of polysomnographic signals for classifying Obstructive Sleep Apnoea

Aoife Roebuck
Kellogg College
Department of Engineering Science

Doctor of Philosophy (D.Phil)
University of Oxford
Michaelmas Term, 2013

Obstructive sleep apnoea (OSA) is a common disorder involving repeated cessations of breathing due to airway collapse, causing disruption of sleep cycles. The condition is under-diagnosed and the side effects are many and varied. Currently, the 'gold standard' diagnostic tool for OSA is a polysomnogram (PSG) which is carried out overnight in a hospital using multiple sensors. A PSG is expensive to set-up, run and analyse, and some subjects experience different sleep patterns due to the artificial conditions of the sleep laboratory.

The aim of this thesis was to find a parsimonious and easy-to-collect set of signals (from the superset of signals recorded in sleep clinics) and other related information (such as demographics), and a set of automated methods that reliably determine which subjects are suitable for standard treatments, *i.e.* classify subjects requiring treatment (moderate OSA, severe OSA) from those not requiring treatment (normal, snorer, mild OSA), using a smartphone. Data were collected from 1354 subjects in the home using the Grey Flash polysomnographic recording device (Stowood Scientific Instruments, Oxford, UK). Analysis of the audio signal was initially performed using standard speech processing methods, where individual events were annotated and classified. The results achieved (accuracy (Ac) = 69.6%) using this approach were lower than those required for clinical acceptance. In all subsequent work in the thesis, subjects were classified from entire recordings rather than events. Multiscale entropy (MSE) was used to identify non-linear correlations in the audio data and quantify the irregularity of the data over many time scales. The inter-snore interval (ISI) was developed, motivated by clinical intuition. MSE and ISI were then applied to both actigraphy and photoplethysomgraphy (PPG) data, and different combinations of features were analysed. The features which displayed the highest predictive accuracy were derived from the PPG signal (Ac = 89.2%). This work demonstrated that, although audio- and actigraphy-based OSA screening is possible, to achieve clinically acceptable performance PPG remains an important key factor in diagnosis.

Acknowledgements

This thesis would not be possible without the help of numerous people. In particular I want to thank the EPSRC, Prendergast Bequest, Department of Engineering Science and Kellogg College for funding this research.

Many thanks go to my supervisor, Dr Gari Clifford, for his support, guidance, encouragement and advice throughout the three years of my D.Phil. Thanks also go to Prof John Stradling for his insights into the physiology of OSA and patience when answering my questions, and Dr Lyn Davies for support and equipment.

A big thank you to the Intelligent Patient Monitoring group, past and present, who have been great to work with. A special thank you goes to Louis, Alistair and Nic for their advice and help with random forests, and Joachim and the SleepAp group for pushing forward with the project.

Last, but by no means least, I thank my family. Catherine and Trish, who got me through the hardest times, Craig, Sadhbh, Mam and Dad for the constant support and belief. I couldn't have done it without you. I dedicate this thesis to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

The International Classification of Sleep Disorders (ICSD) has identified over 80 different sleep disorders, all of which have associated treatments [1]. It is thought that the effects of sleep disorders are extensive, impacting sufferers physically, psychologically and financially [2]. There are numerous health effects of sleep disorders, from the apparently simple daytime sleepiness to an increased risk of cardiovascular disease and stroke [3]. It should be noted that daytime sleepiness is the cause of hundreds of road traffic accidents, and has even been linked to disasters such as Chernobyl [2].

This chapter contains an overview of the purposes of this thesis, a brief description of the physiology of sleep, along with the current classification of sleep disorders by the ICSD. Sleep apnoea is discussed in detail in terms of physiology, diagnosis and treatment.

1.1 Overview

The aim of this thesis was to design a smartphone application (app), using the available sensors and appropriate software algorithms, for the home screening of individuals thought to be at risk of OSA. This was done by using data collected in the homes of subjects referred to a sleep clinic. The signals analysed included those that could be easily recorded by a smartphone, namely audio and actigraphy, with the addition of pulse oximetry, which can be recorded by a smartphone using a Bluetooth pulse oximeter. Although the data were not recorded by a smartphone, the recordings occurred in the subjects' homes, which is the same environment that a smartphone would record the data.

1.2 Sleep Physiology

Loomis [4, 5] provided the earliest detailed description of various stages of sleep, based on electroencephalography (EEG), in the mid-1930s. In 1953 Aserinsky & Kleitman [6] identified rapid eye movement (REM) sleep, which is related to dreaming. Sleep is traditionally divided into two broad types: non rapid eye movement (NREM) and REM sleep. The sleep staging criteria were standardised in 1968 by Rechtschaffen & Kales (or R&K rules) [7], based on EEG changes, dividing NREM sleep into four further stages (stage I, stage II, stage III, stage IV). (It should be noted that some dreaming has been observed during NREM sleep.) The staging was updated in 2004 by the American Academy of Sleep Medicine (AASM) with the most significant change being the combining of stages III and IV into stage N3.

NREM and REM sleep occur in alternating cycles, each lasting approximately 90-110 minutes (*min*) in adults, with approximately 4-6 cycles during the course of a normal 6-8 hour (*h*) sleep period. However, these timings change depending on the length of time asleep, age, medication, physical health and mental health. Furthermore, brief micro-arousals can occur, lasting (by definition) from 1.5-3 seconds (*s*) and short awakenings (defined to be longer than 15*s*) [8].

There are many changes that occur in physiology during sleep, which can be different in REM and NREM sleep. Colten *et al.* [9] compiled data and found that during NREM sleep, brain activity decreases from wakefulness while during REM sleep there is an increase in the motor and sensory areas. Heart rate slows during NREM but increases and varies in REM. Blood pressure and sympathetic nerve activity both decrease in NREM sleep and increase in REM. Autonomic nervous system[1] activity primarily determines changes in blood pressure and heart rate. There is also an increased risk of myocardial infarction in the morning due to sharp increases in heart rate and blood pressure that accompany awakening. Muscle tone is similar to that in wakefulness during NREM and is absent in REM. Respiration decreases from that in wakefulness in NREM but increases and varies in REM and may show brief stoppages with coughing suppressed. Airway resistance increases in both REM and NREM but varies only in REM. Ventilation and respiratory flow become increasingly faster and more erratic, particularly during REM sleep. Hypoventilation has been indicated in both REM and NREM

[1] A control system that acts largely unconsciously and regulates the function of internal organs

sleep caused, it is currently believed, by reduced pharyngeal muscle tone. During REM sleep rib cage movements are reduced and upper airway resistance increases due to the loss of tone in the intercostal and upper airway muscles. Generally, ventilation and respiratory flow show less effective adaptive responses during sleep. During NREM, body temperature is regulated at a lower set point than in wakefulness while it is not regulated at all in REM sleep.

Respiration during sleep is different to respiration while awake. There is a reduction in ventilation, particularly during REM sleep, accompanied by a reduction in mean inspiratory flow rate in REM sleep. Snoring is an obvious respiratory disorder that occurs during sleep. It is a common ailment, affecting approximately 20-40% of the general population. It is caused by the vibration of anatomical structures in the pharyngeal airway. During sleep, the upper airway (UA) dilator muscles become relaxed causing the airway to narrow, thereby increasing resistance. Airflow becomes turbulent and the pharyngeal tissues vibrate as the air passes through. The snoring sound is subject to many influences: the route of breathing, the predominant sites of UA narrowing, sleep stage, body position, naturally occurring versus induced sleep, the presence of sleep-disordered breathing (SDB) [10]. It is now known that snoring is an audible sign of increased UA resistance and is a clinical hallmark of obstructive sleep apnoea (OSA) [1]. Pevernagie *et al.* [10] postulate that acoustic analysis of snoring will enable discrimination between 'simple snorers' and patients with OSA.

Snoring and speech are both generated in the vocal tract. Fundamental frequencies and harmonics can be observed in snoring, similar to formant frequencies in speech. However, the differences between speech and snoring must be taken into account during analysis. Snoring is caused by vibratory activity of pharyngeal structures, not by the vocal cords. The soft palate flutters during snoring, while other structures may also vibrate. There is no articulation of sound during snoring, which occurs mainly on inspiration [10]. According to Hill *et al.* [11] the majority of snores contain a broad spectrum of sound frequencies, but palatal vibration produces marked peaks and troughs, or impulses of sound loudness at low frequency, usually below $50Hz$.

Cheyne-Stokes respiration, or the apnoea-respiration cycle, occurs when breathing is characterised by rhythmic waxing and waning of the depth of respiration; the patient breathes deeply for a short time and then breathes very slightly or stops breathing altogether. The

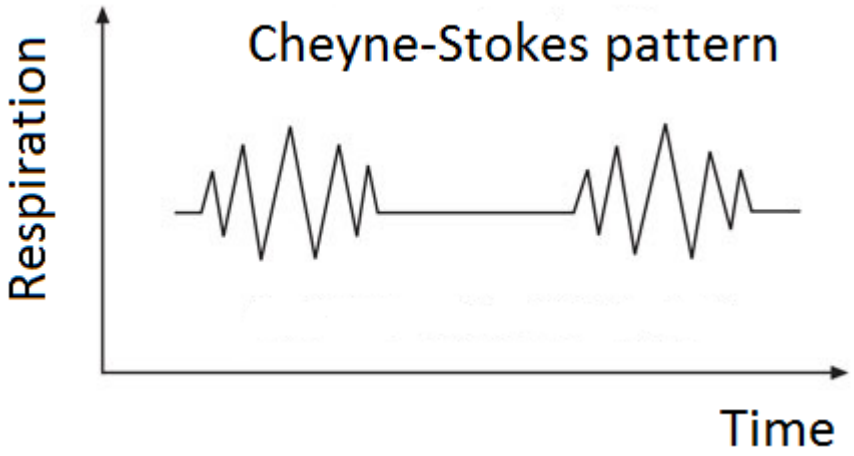


Figure 1.1: *Cheyne-Stokes respiration.*

pattern occurs over and over, every 45s to 3min [12], as in Figure 1.1.

1.3 Sleep Disorders

The ICSD divides sleep disorders into eight categories [13]:

1. Insomnias: difficulty falling asleep, difficulty staying asleep, early awakening or poor sleep quality.
2. Sleep related breathing disorders.
3. Hypersomnias of central origin not due to a circadian rhythm sleep disorder, sleep related breathing disorder or other cause of disturbed nocturnal sleep.
4. Circadian rhythm sleep disorders.
5. Parasomnias: disorders that intrude into the sleep process and are manifestations of central nervous system activation.
6. Sleep related movement disorders.
7. Isolated symptoms, apparent normal variants and unresolved issues.
8. Other sleep disorders.

Sleep apnoea falls under the category of sleep related breathing disorders. Although awareness of this disorder has been increasing, it is estimated that 93% of females and 82% of males

with the condition are undiagnosed and untreated [14]. The rest of this chapter details the physiology of the condition, its prevalence, current diagnostics and treatments.

1.4 Sleep Apnoea

Between 1960 and 1980 sleep apnoea syndrome (SAS) was identified and classified [15]. A detailed paper written in 1976 by Guilleminault *et al.* [16] defined an apnoea as the cessation of airflow at the nose and mouth lasting at least $10s$ and SAS is diagnosed when at least 30 apnoeic episodes are observed in both REM and NREM sleep over a $7h$ period. A hypopnoea is defined as reduced airflow for at least $10s$ and a fall in oxygen saturation (SpO_2) of at least 4%. There are two metrics commonly used to determine the severity of the condition: the AHI and the ODI/RDI. The AHI (or apnoea-hypopnoea index) is the average number of apnoeas and hypopnoeas per hour of sleep while the ODI/RDI (oxygen desaturation index or the respiratory disturbance index) is the average number of oxygen saturation dips per hour of sleep. The ICSD currently defines OSA as the combination of an AHI of at least five as well as excessive daytime sleepiness [10].

There are two forms of SAS: central sleep apnoea (CSA) and OSA, with the latter being more common [17], although the two forms may occur at different times in the same patient. According to Thalhofer & Darrow [17], CSA is characterised by repeated apnoeas during sleep resulting from loss of respiratory effort. OSA occurs when there is a physical obstruction in the airway, as shown in Figure 1.2. In patients with OSA, apnoeas and hypopnoeas are caused by the complete or partial collapse of the upper airway. With no air flowing into the lungs, the arterial oxygen levels drop and carbon dioxide levels rise. There are also increasingly negative pressure swings in the thorax. Blood pressure initially drops and then drifts upwards during the episode. The patient eventually awakens with a surge of sympathetic nervous system activity, leading to a spike in heart rate and blood pressure, and the resumption of breathing. These repeated arousals cause sleep fragmentation which leads to daytime sleepiness [18].

There are a number of factors that predispose a subject to OSA including, being male, increasing age, smoking and alcohol consumption. OSA has been shown to increase the risk of motor vehicle accidents, hypertension, stroke, heart disease and diabetes [18, 19] and is prevalent around the world (Table 1.1). Ben-Bassey *et al.* [20] studied teenagers in Nigeria.

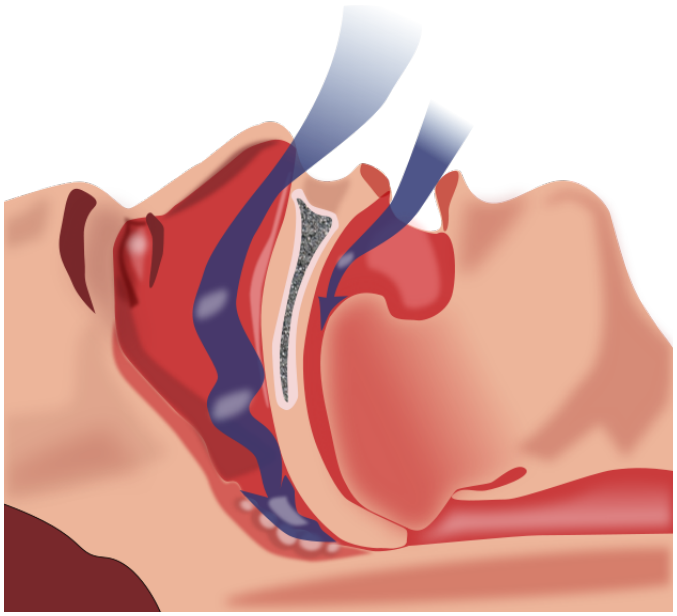


Figure 1.2: *View of the upper airway during an apnoeic event. Taken from* `http://en.wikipedia.org/wiki/Obstructive_sleep_apnea`.

The results showed that being overweight is an evolving problem for this population, which is attributed to an increasingly Westernised lifestyle. This Westernised lifestyle may lead to an increase in the incidence of OSA in populations that are currently unaffected. Taj *et al.* [21] carried out a cross-sectional survey in a hospital in Pakistan in order to determine whether patients were at high- or low-risk for OSA. The authors found that 10% of the population were at high risk for sleep apnoea. Pakistan has a high prevalence of factors which predispose an individual to OSA; however, no studies have been carried out to determine the prevalence of OSA there.

1.5 Diagnosing OSA

Currently, the 'gold standard' in terms of sleep disorder diagnosis (for all sleep disorders) is a sleep study, or an overnight polysomnogram (PSG). When a subject undergoes a PSG, a large number of signals are recorded including the EEG, the electrocardiogram (ECG), the electrooculogram (EOG), the electromyogram (EMG), air flow, thoracic and abdominal movements, and oximetry. Other parameters that may be monitored include body position, video, and audio surveillance. Specialised equipment and a trained technician are required to record these signals correctly. In addition, there are controversies surrounding the efficacy of sleep

Table 1.1: *Prevalence of OSA around the world, m=male, f=female, N/A=not available.*

Study	**Location**	**Ethnicity**	**Gender**	**Age** (*yrs*)	**OSA rate** (%)
Bearpark *et al.* [22]	Australia	Caucasian	m	40-65	3.0
Ip *et al.* [23]	Hong Kong	Chinese	m	30-60	4.1
Ip *et al.* [24]	Hong Kong	Chinese	f	30-60	2.1
Kim *et al.* [25]	Korea	Korean	m, f	40-69	4.5 (m) 3.2 (f)
Lam *et al.* [26]	Asia	Asian	m, f	middle aged	4.1-7.5 (m) 2.1-3.2 (f)
Sharma *et al.* [27]	India	Indian	m, f	N/A	4.9 (m) 2.1 (f)
Udwadia *et al.* [28]	India	Indian	m	25-65	7.5
Young *et al.* [29]	USA	Caucasian	m, f	30-60	4.0 (m) 2.0 (f)

labs in that some subjects do not sleep as well in the lab as they do at home. However, this claim has been refuted by Portier *et al.* [30], who provided evidence that sleep architecture and evaluation of sleep quality were no different between the home or laboratory setting.

PSGs are expensive, limited by the number of beds available in the hospital and the number of trained sleep specialists in the area. In order to overcome some of these issues, there has been a move to home diagnostics. There are a wide range of home diagnostic kits available: ApneaLink (ResMed, San Diego, California) a single channel device; Apneoscreen Pro (VIASYS Healthcare, Causeway Bay, Hong Kong) a 17 channel device; AURA PSG (Grass Technologies, West Warwick, Rhode Island) a 16 channel device; and SleepMinder (Biancamed, Dublin, Ireland) a non-contact sleep and breathing monitoring device, these are just some of the commercially available kits. In addition, a number of questionnaires have been devised as screening tools for SAS. See section 3.1 for more details.

Flemons *et al.* [31] focused on determining the wait time for diagnosis and treatment in five different countries (Table 1.2). The authors postulated that the difference in wait times resulted from the limited beds available for sleep studies in each country, as well as a lack of sleep specialists to score the data. Therefore, a home diagnostic device that is readily available, reliable, and aids in scoring the data would be beneficial, and ideally would considerably reduce the wait time.

Table 1.2: *Wait time for diagnosis and treatment in five different countries [31].*

Country	**Wait time** (*months*)
United Kingdom	7 - 60
Belgium	2
Australia	3 - 16
United States	2 - 10
Canada	4 - 36

1.6 Treatment for OSA

Guilleminault & Abad [32] categorised the various treatments for OSA as follows:

1. Diet and Lifestyle: A number of lifestyle changes, such as losing weight, avoiding tobacco, alcohol and sleeping tablets, and modifying the usual sleeping body position, can all aid in reducing the number of apnoeas and hypopnoeas that occur throughout the night.

2. Pharmacological Treatments: Avoiding benzodiazepines and barbiturates in particular, and minimising the use of narcotics in general, will help as they worsen apnoeas, hypopnoeas and UA functionality. Some research has been carried out with limited success on drug treatments which stimulate the neurotransmitters which contract the UA dilator muscles in an effort to maintain UA patency [33–35].

3. Therapeutic Devices: These are oral appliances that physically modify the UA whilst being worn. They are usually mandibular advancement devices (MAD) or tongue trusses which hold the lower jaw and tongue forward. The efficacy of oral appliances (OAs) in the treatment of OSA is questionable as, on average, only 52% of patients treated with OAs had some success in controlling OSA. Effects on sleepiness and quality of life were demonstrated but improvement in other neurocognitive outcomes were not consistent [36]. Tongue retaining devices (TRDs) are another possibility which were originally designed to combat snoring. They are mouthpieces which are worn while asleep, fitting over both upper and lower dental arches with a compartment to hold the tongue in a forward position by suction. TRDs can improve nocturnal respiration for a wide range of apnoea severities, provided that the disorder is more severe in the

supine position and that the body weight is not greater than 50% above the ideal [37]. Although these devices have been shown to be effective, patient tolerance of the device has appeared to be lower than for MAD [38]. This might explain why they are prescribed so infrequently [39].

4. Surgery: There are a number of options for surgery on the UA. The area operated on depends on where the obstruction occurs in the individual patient. Some of the surgical treatments available include: nasal reconstruction - to improve normal respiration; tonsillectomy and adenoidectomy - usually used for children with OSA in order to enlarge the nasal inferior turbinates; mandibular osteotomy with genioglossus advancement - to enlarge the retrolingual (posterior to the tongue) airway.

5. Assistive Devices: Positive airway pressure devices are the most commonly used therapy for OSA and include continuous positive airway pressure (CPAP), bilevel positive airway pressure (BiPAP) and autopositive airway pressure (APAP). A device like an oxygen mask is worn over the mouth and/or nose and pressurised air is forced down the airway thereby keeping it open. They are extremely effective when used correctly; however, approximately 30-35% of patients are intolerant or non-compliant due to the side effects of use, which include skin abrasions, bruising, chaffing from the mask, nasal congestion or dryness, abdominal cramping [32].

6. Electrical stimulation: Electrical stimulation of the lingual musculature is another form of treatment. Fine wire electrodes are implanted into either the genioglossus or the hypoglossal nerve. By stimulating the nerves, UA patency is improved and it is possible to maintain airflow without arousing patients from sleep [40–42], as shown in Figure 1.3.

This list comprises the typical treatments available to sufferers of OSA in the developed world. Although the same treatments can also be used in developing countries, cost considerations and supply infrastructure limitations severely restrict their availability. Lam *et al.* [26] conclude that while CPAP is available in many parts of Asia it may not be a financially viable option. They also suggest that OAs may be a more suitable treatment for Asian patients. This is due to the belief that there are more modifiable factors in the craniofacial structure of Asian

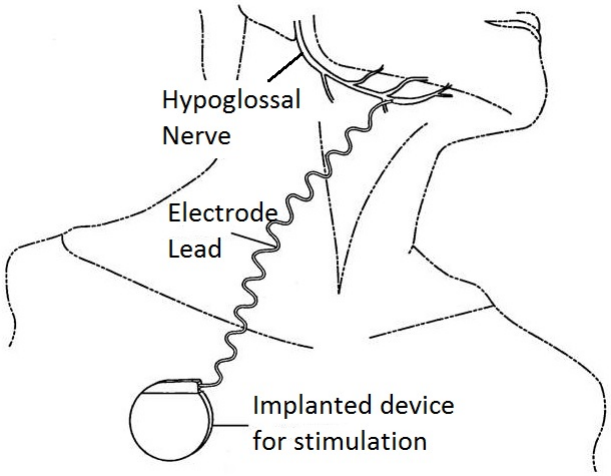


Figure 1.3: *A device is implanted into the chest with an electrode used to stimulate the hypoglossal nerve when breathing is reduced. This helps to maintain UA patency thereby reducing the occurrences of apnoeas and hypopnoeas. Adapted from [43].*

patients.

1.7 Summary and Thesis Overview

The current method used to diagnose OSA and the increasing awareness of the disorder (which will only increase the wait times for diagnosis and treatment) both indicate that a low-cost system using as few sensors as possible would be beneficial in screening for and monitoring the effects of the treatment of OSA. The aim of this thesis was to find a parsimonious and easy to collect set of signals (from the superset of signals recorded in sleep clinics) and other related information (such as demographics), that can be used to reliably determine which subjects are suitable for standard treatments, *i.e.* classify subjects requiring treatment (moderate OSA, severe OSA) from those not requiring treatment (normal, snorer, mild OSA), using a smartphone. In order to develop an app to screen for OSA in the home, the signals analysed were those that could be collected by a smartphone, namely audio and actigraphy, with the addition of pulse oximetry that can be recorded by a Bluetooth pulse oximeter. Such an app would be of particular benefit in developing countries, where the prevalence of OSA is similar to that in developed countries (see Table 1.1) but there are far fewer sleep laboratories and sleep clinicians available to screen for OSA.

The above therefore motivates the focus of this thesis, which is to identify a parsimonious set of features from the superset of PSG signals, which can provide an acceptable accuracy for

diagnosing OSA using signals recorded by a smartphone. A large clinical database of 1354 overnight PSGs of varying diagnoses was collected over a three year period. A subset of signals which are easy to record in the home using a mobile phone or similar low cost and small form-factor Holter device were analysed both univariately, then in a multivariate manner using novel signal processing approaches and machine learning. The main contents of the thesis are described below. Chapter 1 outlines and motivates the research in this thesis, and provides some clinical background. Chapter 2 details the commonly used metrics and techniques to quantify model performance. Chapter 3 reviews previous work in the field on audio, actigraphy and pulse oximetry analysis as well as the utility of questionnaires. These three chapters form the basis of an extensive review article now published: Roebuck *et al.* [44]. Chapter 4 examines the use of the audio signal for identifying sleep apnoea using a traditional signal processing approach to provide a baseline. Chapter 5 presents a novel approach to analysing audio recordings during sleep (based on multiscale entropy) which provides an improvement on the baseline method. Using a random forest classifier, the accuracy in identifying OSA is elevated from 71% to 79% when using multiscale entropy alone or 89% when fused with demographics and ODI. Chapter 6 adds another traditional signal, the actigraph (recording movement from an accelerometer). A similar approach is applied as with the audio signal, and combines features derived from both signals together with information on demographics. The findings indicate that the actigraph adds no additional information to the classification accuracy. Chapter 7 adds one final signal to the analysis, the photoplethysmogram. After conducting baseline studies to determine the accuracy of the PPG on its own, a machine learning approach is taken to combining features from audio and the PPG. An accuracy of 88%, which is superior to the PPG alone (73%) was found. Chapter 8 discusses the importance of this work and presents a smartphone app which has been designed to translate this work to the home and demonstrate the feasibility of using just the signals (and analysis techniques) described in this thesis. The work in this thesis has been ported to a smartphone and ethical clearance has been given to start trials in a local hospital to prospectively validate the system. Parts of Chapter 8 appear in Behar *et al.* [45] and Roebuck *et al.* [44] and were presented at the Computing in Cardiology conference 2013. See Appendix B for more details regarding these journal articles.

Chapter 2

Metrics used to predict OSA

2.1 Performance Metrics

Classifier/model performance can be evaluated in terms of its ability to distinguish between positive and negative outcomes. For models that output a probability estimate, a threshold is selected and observations that are above the threshold are classified as the positive outcome. This threshold is known as the operating point and transforms a continuous prediction into a binary output. A variety of performance metrics can be calculated once the continuous prediction has been transformed into a binary output. Depending on the threshold or operating point chosen, every observation can be classified as a positive or a negative outcome. These predictions can be seen as true positives, false positive, true negatives or false negatives. Table 2.1 describes these metrics.

Operating point statistics are measures of a classifier's performance at a single operating point, where an operating point can be thought of as the separation boundary between the predicted classes. The following statistics represent the efficacy of a classifier at a given operating point.

Accuracy (Ac) represents the rate at which the classifier correctly labels all data. Though

Table 2.1: *Description of the measurements which form the basis of many model evaluation statistics.*

Metric	Description	Actual	Prediction
True Positive (TP)	Model correctly predicts a positive outcome	Yes	Yes
False Positive (FP)	Model incorrectly predicts a positive outcome	No	Yes
True Negative (TN)	Model correctly predicts a negative outcome	No	No
False Negative (FN)	Model incorrectly predicts a negative outcome	Yes	No

useful for evaluating the overall accuracy of a model in a single metric, it can be misleading in binary cases where one class occurs much more frequently.

$$Ac = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity (Se) measures how accurately a model discerns whether an instance is truly a member of the positive class [46]. This is a very effective measure for determining the discrimination of a model, especially in cases with low probability of a positive outcome.

$$Se = \frac{TP}{TP + FN}$$

Specificity (Sp) evaluates how accurately a model discerns whether an instance is truly a member of the negative class [46]. A low frequency of positive outcome may result in an exaggerated specificity due to a low number of true positives.

$$Sp = \frac{TN}{TN + FP}$$

The positive predictive value (PPV) is the ratio of subjects with a positive outcome that were correctly diagnosed [47]. It is useful for determining how effectively a classifier correctly assigns a positive outcome, as well as providing a credibility measure in this positive classification. High positive predictive values infer a higher confidence in predicted positive outcomes.

$$PPV = \frac{TP}{TP + FP}$$

The negative predictive value (NPV) is the ratio of subjects with a negative outcome that were correctly diagnosed [47]. High negative predictive values infer a higher confidence in predicted negative outcomes.

$$NPV = \frac{TN}{TN + FN}$$

The receiver operating characteristic curve (ROC curve) represents a model's discrimination in a simple and easy-to-interpret 2D plot [48]. The true positive rate (Se) of the model is plotted against the false positive rate (1 - Sp). By calculating the TPs, TNs, FPs and FNs for every possible threshold in the predicted outcome, where values above the threshold are

classified as positive and values below the threshold are classified as negative, the sensitivity and specificity can then be calculated. Ideally, there would exist a point on the plotted curve equivalent to 100% specificity and 100% sensitivity. As a model improves, the ROC curve will approach $(0, 1)$, which is the top left corner of the plot. This curve is useful for assessing the trade off between sensitivity and specificity and for selecting the operating point for the model being evaluated.

The area under the ROC curve (AUC) ranges between zero and one, where higher values indicate that their respective model is more effectively discriminating between the two outcomes. A value of 0.5 indicates that the model does not predict any better than chance. Though the AUC summarises the discrimination and performance of the model well, it does not provide the information regarding the trade off between sensitivity and specificity that the ROC curve does.

2.2 Model Comparison

The common metrics used to describe a particular model's discriminatory power have been described above. However, it is also useful to be able to compare one model to another and to find out whether the difference is statistically significant or not.

It is possible to compare two AUCs calculated by two different models on the same observations, although this is not straightforward. Hanley & McNeill [49] showed that the standard error (se) of an AUC can be estimated as follows:

$$se = \sqrt{\frac{W(1-W) + (n_{\ominus} - 1)(Q_1 - W^2) + (n_{\oplus} - 1)(Q_2 - W^2)}{n_{\ominus} \times n_{\oplus}}} \tag{2.1}$$

$$Q_1 = \frac{W}{2 - W} \tag{2.2}$$

$$Q_2 = \frac{2W^2}{1 + W} \tag{2.3}$$

where $n_{\ominus}$ and $n_{\oplus}$ denote the number of subjects in the negative and positive classes respectively, W is the Wilcoxon statistic[1], Q_1 is the probability that two randomly chosen positive observations are predicted with a greater probability than any observation and Q_2 is the proba-

[1] A non-parametric test that compares two paired groups

bility that one randomly chosen positive observation is predicted with greater probability than two randomly chosen negative cases [49, 50]. Hanley & McNeill [51] then went on to define the difference between two AUCs as:

$$se(W) = \frac{W_1 - W_2}{\sqrt{se_1^2 + se_2^2 - 2 \cdot R \cdot se_1 \cdot se_2}} \tag{2.4}$$

and is assumed to follow a standard normal distribution from which statistical significance can be obtained. The correlation coefficient R between two models that are derived from the same observations is estimated from a lookup table provided by Hanley & McNeill [51]. DeLong *et al.* [52] provide an implementation of this problem that does not require the use of a lookup table.

The net reclassification index (NRI) developed by Pencina *et al.* [53] is another method for comparing two models. The notion of upward movement (*up*) is introduced as a change from a lower category of risk with the old model to a higher category with the new model. The downward movement (*down*) is the opposite. In terms of this classification problem, a patient previously classified as requiring treatment and predicted as not requiring treatment by the new model would be considered as a *downward* move. The NRI is defined as:

$$NRI = (P(up|y = 1) - P(down|y = 1)) - (P(up|y = 0) - P(down|y = 0)) \tag{2.5}$$

$$z = \frac{NRI}{\sqrt{\frac{P(up|y=1)-P(down|y=1)}{\#events} + \frac{P(up|y=0)-P(down|y=0)}{\#nonevents}}} \tag{2.6}$$

Pencina *et al.* showed that a simple asymptotic test for the null hypothesis of $NRI = 0$ can be used as in 2.6.

This method evaluates performance at a specific operating point and the authors propose another metric which accounts for all possible classification thresholds. The integrated discriminative improvement (IDI) is defined as follows:

$$IDI = (IS_{new} - IS_{old}) - (IP_{new} - IP_{old}) \tag{2.7}$$

where IS denotes the integration of sensitivity over all the cut-off values and IP denotes the integration of one minus specificity over all cut-off values. As for the NRI, a simple asymptotic

test can be used to test the null hypothesis that $IDI = 0$ as follows:

$$z = \frac{IDI}{\sqrt{se^2_{events} + se^2_{nonevents}}} \tag{2.8}$$

2.3 Choosing Metrics

The metrics described above reflect how well the model fits the data. The metric chosen to represent the model will depend on the expected use of the model. Accuracy can be used to identify how well the model fits the data overall. AUC provides an aggregated measure of discrimination, but in reality a single operating point needs to be chosen. Depending on the risks of the model predicting FP or FN, different metrics can be used to represent the model. Sensitivity is useful in cases where it is important to identify all positive cases, *i.e.* a screening tool for OSA would need to identify all cases of OSA, while specificity can be used where it is more important to identify all negative cases, *i.e.* deciding which subjects to treat for OSA would require that all negative cases are excluded. Positive predictivity is useful when deciding whether to continue treatment or not, while negative predictivity can be used for deciding which subjects to remove from treatment.

2.4 Validating Model Performance

Ideally, models would be validated by collecting new data prospectively and computing the performance metrics on this unseen data. The collection of prospective data is not always possible and so a different approach is usually taken: prospective data is 'collected' by setting the most recent observations aside while the model is designed, then these untouched observation are used to validate the model. In some circumstances this is not possible as the data associating each subject with the date when the recording occurred is lost, usually as part of the de-identification process. In these cases, cross-validation is usually used for model validation [54] which involves splitting the data set into independent and complementary datasets:

design set (X_{design}), composed of training and test data;

validation set (X_{val}), independent of X_{design}.

The proportion, v, used to divide the data into design and validation may influence the re-

sults achieved and performance will vary depending on which data is drawn for the design and validation sets. In addition, missing data in the design set may impair parameter estimation, and different training/test distributions may affect results. To quantify this variability, cross-validation is carried out using re-sampling:

- K-Fold cross-validation defines K independent and complementary groups of equal size. The model is fitted to the data K times, leaving one K^{th} of the data aside as X_{val}. When complete, each observation will have been used K times for the design and once for validation, guaranteeing an equal contribution of all observations in the computation of the performance metrics.

- Bootstrapping draws, with replacement, B random design sets consisting of $v\%$ of the entire data set. Performance metrics and their variability can be estimated, although there are two limitations to this approach: first, the results are not reproducible and second, the respective contribution of each observation is uncontrolled.

- Jack-knifing considers all possible combinations of data splits for a given v. This means that the performance metrics are reproducible however, the computation time can be high for this method.

- Leave-one-out is a special case of jack-knifing where v is chosen so that a single observation is left out as X_{val}. In this case, N training datasets of $N-1$ size are used as X_{design} meaning that N independent predictions are used for model validation. It is not possible to estimate the variance of the model using this method however, it is often used where there is limited data.

2.5 Portable Monitors for OSA

The efficacy of portable monitors (PMs) and what performance statistics are required for them to be considered reliable for a clinical diagnosis in an unattended home setting is still controversial. A review by Collop *et al.* [55] of PMs divided them into four types:

1. Type 1: full **attended** PSG (≥7 channels) in a laboratory setting.

2. Type 2: full **unattended** PSG (≥7 channels), which can be considered as comprehensive portable PSG.

3. Type 3: limited channel devices (4–7 channels).

4. Type 4: 1–2 channels using oximetry as one of the channels.

Many of the PMs that the authors evaluated were attempting to replicate AHI values from PSG, to mimic manual/visual scoring or were interested in reproducibility. Table 2.2 details the relevant studies carried out, the type of device used and the associated results of those reviewed by Collop *et al.* As the work in this thesis is not based on full PSG and all recordings have taken place in the home, only PMs of types 3 and 4 in an unattended home setting have been considered.

Table 2.2: *Efficacy of portable monitors. All studies compared the results of a home PM to full in-laboratory PSG. Se = sensitivity, Sp = specificity, PPV = positive predictive value, NPV = negative predictive value, Ac = accuracy.*

Study	Type	Gold Standard	Results (%)
Álvarez [56]	3	PSG	Se = 90.1, Sp = 82.9, Ac = 87.2
Ancoli-Israel [57]	3	PSG	Se = 100, Sp = 66
Ayappa [58]	3	PSG	Se = 88, Sp = 92
Bachour [59]	3	PSG	Se = 64, Sp = 78
Dingli [60]	3	PSG	failure rate of 18%
Nakano [61]	4	PSG	Se/Sp = 54/100, 83/97, 98/78 normal-weight, overweight and obese
Schäfer [62]	3	PSG	Se = 94, Sp = 41
Westbrook [63]	3	PSG	Se = 91.5, Sp = 85.7, PPV = 91.5, NPV = 85.7
Whitelaw [64]	4	PSG	correct prediction rate = 0.64
Yin [65]	3	PSG	AHI specificity became higher with respect to OSA severity. Record time had to be more than 390*mins* to get enough data. Sleep position was significantly different between PM and PSG.

Collop [66] recently updated this review online including literature up to September 2013. The latest review concludes that PM is an acceptable approach for diagnosing OSA in those subjects with a high pre-test probability of having moderate to severe OSA and no comorbid medical or sleep disorders. The author notes that type 3 and type 4 devices have highly variable diagnostic performance and most do not include a conventional measure of sleep, which has many drawbacks. In addition, the authors suggest that pulse oximetry should not be used alone to diagnose suspected OSA, as it tends to have high specificity but low sensitivity when quantitative criteria are used. When qualitative criteria are used, pulse oximetry tends to have high sensitivity but low specificity. As such, pulse oximetry can be either sensitive or specific

but not both; therefore, false-positive or false-negative tests will be common depending on the criteria chosen to define a positive test.

Flemons *et al.* [67] carried out a large review of PMs and addressed the question of whether a single PM could be used to both reduce and increase the probability that a subject has an abnormal AHI, among others. In order to both reduce (give a negative result) and increase (give a positive result) the probability that a subject has OSA, a PM has to have both a high sensitivity/low likelihood ratio (LR) for a negative result and a high specificity/high LR for a positive result. If a PM is able to achieve this at a single threshold, then subjects will have either positive or negative results, and if the test has excellent operating characteristics there will be few false positive or false negative results. For type 3 devices in an unattended home setting, the authors found two of four home studies (with evidence level II) provided both high positive and low negative LRs. Both studies used flow thermistry and an optional oxygen desaturation criterion for detecting events. These studies also used different thresholds to produce the low and high LRs, given 22% of subjects in one study and 37% in the other had intermediate results and were not classified as having, or not having, sleep apnoea. The misclassification rates were 16% and 5% respectively. The authors note that, due to the limited data from the home setting, additional evaluation is required. For type 4 devices, only one study (using SpO_2) reported data that allowed the monitor to classify some patients as either having or not having sleep apnoea. There was a high percentage of subjects not classified (50%) but there was a low misclassification rate (4%). The authors note that overall, most of the studies reporting data that produced both low and high LRs were rated as good quality (evidence level I/II) and the monitor type or primary signal measured did not appear to influence the results. An interesting point that the authors raise is that the standard approach used to validate PMs is to compare a PM with a reference standard. However, this approach is limited in that it assumes that lab-based PSG is the optimal approach for diagnosing OSA. This is not necessarily true. As stated previously, subjects often do not sleep as well in a lab-setting as they do at home, and they are likely to spend more time on average sleeping supine. In addition, the AHI correlates poorly with outcomes that are important to subjects, such as quality of life and daytime sleepiness, and does not predict very well those subjects who will ultimately use and benefit from therapy. The authors conclude that a more appropriate validation study would compare

the impact of PM and PSG on a physician's decision-making ability and outcomes important to subjects. To date, no studies have been published using this approach.

Epstein *et al.* [68] state that PM for diagnosing OSA should only be performed in conjunction with a comprehensive sleep evaluation, and sleep evaluations using PM must be supervised by a practitioner with board certification in sleep medicine, or equivalent. The authors recommend that PMs should record airflow, respiratory effort and blood oxygenation at a minimum, with an experienced sleep technician, sleep technologist, or appropriately trained healthcare practitioner either educating the subject on the correct placement of the sensors or applying them personally. PMs may be used in a home setting as an alternative to PSG for subjects with a high pretest probability of moderate to severe OSA, with no comorbid sleep disorders or major comorbid medical disorders. As the diagnosis of OSA is confirmed and severity determined using the same criteria as for PSG, scoring criteria should be consistent with the current published AASM standards from scoring apnoea and hypopnoeas. PMs are known to have a high false negative rate, and so in-laboratory PSG should be performed in cases where PM is technically inadequate or fails to establish a diagnosis of OSA in subjects with a high pretest probability.

2.6 Summary

This chapter comprises the statistical methods used to quantify and compare model performance. In this thesis, the parameter maximised is Ac. The population studied has an approximately 60:40 split of treatment vs. non-treatment subjects; Ac should not be misleading as one class does not occur more frequently than the other. Both NRI and IDI are used to compare model performance; NRI for classifiers with binary outputs and IDI for classifiers with probabilistic outputs. The models are trained and validated using five-fold cross-validation.

Chapter 3

Review of existing data analysis techniques for assessing OSA

The literature is replete with studies which report on a variety of signals and algorithms to detect OSA. The work presented in this thesis is based upon some of the easiest and least costly signals to collect: audio, actigraphy, pulse oximetry and demographic information (collected from questionnaires). Therefore the literature review presented here focuses on these data types only.

3.1 Questionnaires and Demographics

A number of questionnaires have been devised for the pre-screening of OSA. The most commonly used are: the Epworth Sleepiness Scale (ESS) [69], the STOP BANG questionnaire [70], Flemon's questionnaire [71] and the Berlin Questionnaire (BQ) [72]. The individual questionnaires and how they are scored can be found in Appendix C, while an overview of what the different scores mean can be found in Table 3.1.

The ESS [69] asks a subject to rate how likely they are to fall asleep under a range of conditions such as driving, while sitting down, etc. The scale ranges from 0 to 24 where ESS$<$ 11, 11 $\leq$ESS$\leq$ 14, 15 $\leq$ESS$\leq$ 18 and ESS$>$ 18 are defined as normal, mild subjective daytime sleepiness, moderate subjective daytime sleepiness and severe subjective daytime sleepiness respectively [73]. However, there is a relatively weak correlation between ESS and OSA severity [74, 75].

The STOP BANG questionnaire was developed for OSA screening in surgical patients (*i.e.* those about to undergo any surgical operation) [70]. Undiagnosed OSA in surgical patients can have a serious impact on postoperative outcomes. Identifying patients with a high risk of OSA can help to prevent adverse health events and perioperative outcomes. The name derives from the eight questions it is comprised of: snoring, tiredness during the day, observed apnoeas, high blood pressure, body mass index (BMI)[1] (cut-off of 30), age (cut-off of 50), neck circumference (cut-off of 40cm) and gender (male). Any three of the eight indicates a higher probability of OSA. The STOP BANG was completed by 2,974 patients in the preoperative clinics of Toronto Western Hospital and Mount Sinai Hospital, Toronto, Ontario, Canada. 211 patients underwent PSG; 34 for the pilot study and 177 for validation. For an AHI of 5, 15 and 30 Se of 83.6% (Sp = 56.4%), 92.9% (Sp = 43%) and 100% (Sp = 37%) were found. It is clear that the STOP BANG is highly sensitive but not specific.

Flemons' questionnaire, or the Calgary Sleep Apnoea Quality of Life Index (CSAQLI), is a non-clinical questionnaire that evaluates health-related quality of life in patients with sleep apnoea, both before and after treatment. This is a very detailed questionnaire including questions regarding daily activity, symptoms and treatment for sleep apnoea. More information can be found in Appendix C. Flemons & Reimer [76] found that there was a weak correlation between the CSAQLI and the ESS ($r = -0.26, p = 0.02$). A change in CSAQLI score with treatment was correlated with changes in the RDI ($r = -0.46, p < 0.0001$). The CSAQLI had a high reliability coefficient of 0.92 on testing and retesting at two weeks.

The BQ identifies subjects at risk of SAS. The BQ was assessed on 130 sleep clinic patients and, for an RDI > 10, Se was 62% (Sp = 43%) [77]. The authors concluded that the BQ was not appropriate for identifying patients with sleep apnoea in a sleep clinic population.

A variety of demographics have also been used to screen/predict OSA, including age, gender, height and weight. Stradling & Crosby [78] found that neck size ($r^2 = 7.9\%, p < 0.0001$) and alcohol consumption ($r^2 = 3.7\%, p < 0.0001$) correlated best with OSA, and less well with age ($r^2 = 1\%, p = 0.009$) and general obesity ($r^2 = 1\%, p = 0.01$). Chung *et al.* [70] developed the STOP BANG questionnaire in two stages: firstly looking at STOP questions (snoring, tiredness during the day, observed apnoeas and blood pressure) and then seeing the improvement that could be obtained by including demographic information from

[1]BMI is a proxy for body fat percentage based on an individual's height and weight.

Table 3.1: *Meaning of the questionnaire scores.*

Questionnaire	Meaning
ESS	ESS < 11: normal 11 ≤ ESS ≤ 14: mild subjective daytime sleepiness 15 ≤ ESS ≤ 18: moderate subjective daytime sleepiness ESS > 18: severe subjective daytime sleepiness
STOP BANG	STOP BANG < 3: low risk of OSA STOP BANG ≥ 3: high risk of OSA
CSAQLI	used as an evaluative instrument to measure within-subject change in response to a therapeutic intervention. Some of the adverse consequences of currently available therapies for sleep apnoea are captured
BQ	≤ 1 categories with a positive score: low risk of OSA ≥ 2 categories with a positive score: high risk of OSA

the BANG questions (BMI, age, neck circumference, gender). The authors found that Se (Sp) increased from 65.6% (60.0%) to 83.6% (56.4%) when demographics were included for an AHI > 5, indicating that demographics may be useful. It is unclear whether demographics improve OSA diagnosis which may be because subjects are asked to fill in the information themselves, and could therefore be reporting inaccurate figures.

3.2 Audio

Due to the physiological similarities between how snoring and speech are produced (detailed in section 1.2), not to mention the availability of common methods for digital processing and analysis, the audio analysis of snoring has been approached from the perspective of speech analysis [10]. The following are some basic metrics used to quantify speech and snoring:

- Sound Pressure Level (SPL): The logarithm of the ratio of a given pressure level p to a reference pressure p_0. $SPL = 20log_{10}(p/p_0)$.

- Signal to Noise Ratio (SNR): The ratio of signal to noise that is received, where S is an estimate of the underlying signal of interest, and N is an estimate of the noise; $S + N$ is the complete observation. $SNR = 20log_{10}(S/N)$.

- Weighted sound intensity measurement: Used to mimic the perception of loudness, and reduce the influence of certain frequencies in an audio signal. There are three commonly used weightings: A, B and C which refer to different sensitivity scales for noise mea-

surement. Each weighting is a filter with different poles and zeros depending on which frequencies are to be reduced/removed. The A-weighting curve follows the frequency sensitivity of the human ear at low levels, *i.e.* much of of the low-frequency noise is filtered out. The B-weighting curve follows the frequency sensitivity of the human ear at moderate levels while the C-weighting curve follows the frequency sensitivity of the human ear at very high noise levels. This scale is quite flat and includes more of the low-frequency range than the A and B scales.

- Root Mean Square (RMS) value: The square root of the time average value of the square of a signal. $V_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} V_i^2}$.
- Crest Factor: The highest absolute value of a signal divided by the RMS value. This method has been used to distinguish between palatal and non-palatal snoring [11,79,80] with the crest factor being higher for palatal snoring.
- Equivalent noise level: The squared ratio of the signal (p_A) and a reference signal (p_0), averaged over a time T, then converted to a logarithmic value. $L_{eq} = 10\log\left[\frac{1}{T}\int \frac{p_A^2}{p_0^2}\,dt\right]$.

More advanced methods exist that both analyse and model snoring sounds and these are explored in the following sections.

3.2.1 Linear Predictive Coding

Developed in the late 60s by Atal & Hanauer [81], linear predictive coding (LPC) attempts to model each new speech sample as a linear combination of previous samples. LPC is a model of an all pole filter; the vocal tract can be approximated by LPC due to its resonant chamber, except for nasal sounds which introduce zeros (see section 1.2).

Solá-Soler *et al.* [82] used LPC to analyse the snoring signal's spectral envelope. They used 16 snorers: 8 simple snorers (6 males (m), 2 females (f); age = 46.0±8.15yrs; BMI = 27.93±3.01kg/m^2; AHI = 8.78±2.64$events/h$) and 8 OSA patients (8 m; age = 50.75±8.01yrs; BMI = 28.96±2.32kg/m^2; AHI = 34.04±25.1$events/h$). Significant differences in the formant frequency variability between the two groups were found: the standard deviation of some snoring formant frequencies was significantly lower ($p < 0.005$) in simple snorers than in OSA patients, even including non post-apnoeic snorers in this latter group.

Ng *et al.* [83] used LPC to model 30 apnoeic snorers (24 m, 6 f; AHI = 46.9±25.7$events/h$) and 10 benign snorers (6 m, 4 f; AHI = 4.6±3.4$events/h$). LPC was used to calculate the first three formant frequencies[2] (F1, F2 and F3) and quantitative differences were found between the two groups with higher formant frequencies appearing in apnoeic snorers. Apnoeic snorers were differentiated from benign snorers with Se = 88% and Sp = 82% using a threshold value of F1 = 470Hz.

Yadollahi & Moussavi [84] modelled 15 snorers (12 m, 3 f; age = 52.3±15.2yrs; BMI = 35.1±4.6kg/m^2; AHI = 33.9±42.3$events/h$) using LPC. A total of 1636 snore segments and 3059 breath segments at different sleeping positions were selected from all subjects and the authors found that F1 and F3 were significantly different between breath and snore segments ($p = 0.003$ and $p = 0.0244$ respectively).

3.2.2 Frequency Analysis

Fiz *et al.* [85] studied 17 snorers: 10 with OSA (10 m; BMI = 32.9±7.6kg/m^2; AHI = 26.2$events/h$) and 7 simple snorers (7 m; BMI = 29.7±7.2kg/m^2; AHI = 3.8$events/h$). Spectral analysis showed two distinct patterns. The first pattern had a fundamental frequency and harmonics while the second pattern had a low frequency peak with the sound energy scattered on a narrower band of frequencies, but without clearly identifiable harmonics (with a lower peak frequency). The simple snorers and two of the OSA patients displayed the first pattern while the other eight OSA patients displayed the second pattern. Significant negative correlations between AHI and peak and mean frequencies of the snoring power spectrum were found ($p < 0.0016$ and $p < 0.0089$, respectively).

McCombe *et al.* [86] developed an acoustic index, *Hawke Index*, on 9 OSA patients (8 m, 1 f; BMI = 28.7±4.1kg/m^2; AHI>15$events/h$) and 18 simple snorers (16 m, 2 f; BMI = 28.6±3.9kg/m^2; AHI<15$events/h$). The index is the ratio between the overall A-weighted sound level in decibels (dB(A)) and linear sound level in decibels (dB(SPL)) for the recorded snoring sound of each subject for the maximum snoring sounds for each subject (L_{max}) [HI = dB(A)/dB(SPL) for L_{max}]. A large low-frequency peak in SPL at approximately 80Hz was observed, while the OSA group had a larger high frequency sound component. The *Hawke*

[2]These appear where there is a concentration of energy around a particular frequency in the acoustic wave.

Index had Se = 67%, Sp = 100%, PPV = 100% and NPV = 86%.

Meslier *et al.* [87] found that the fundamental frequency of snoring was between 40 and $75Hz$. Perez-Padilla *et al.* [88] observed similar results when they analysed 10 heavy snorers and 9 OSA patients using the fast Fourier transform (FFT). They found that most of the snoring noise power occurred below $2kHz$ with a peak power less than $500Hz$. The OSA subjects displayed a sequence of snores with spectral characteristics that varied markedly through an apnoea-respiration cycle (see section 1.2). OSA patients exhibited residual energy at $1kHz$ while heavy snorers did not. Fiz *et al.* [85] postulated that the differences between their results and the results from Perez-Padilla *et al.* [88] are due to the differences in experimental methods and recording equipment. Specifically, Fiz *et al.* placed the microphone above the larynx while Perex-Padilla *et al.* placed the microphone on the manubrium sterni.

Hara *et al.* [89] analysed 12 simple snorers (8 m, mean BMI = $24.5kg/m^2$; 4 f, mean BMI = $24.8kg/m^2$; AHI≤$5events/h$) and 46 OSA patients (40 m, mean BMI = 25.8 kg/m^2; 6 f, mean BMI = $26.1kg/m^2$; AHI≥$20events/h$). The parameters used were peak frequency, soft phonation index (SPI), noise to harmonics ratio (NHR), and power ratio. SPI is the average ratio of lower frequency harmonic energy in the 70-$1600Hz$ range to higher frequency harmonic energy in the 1600-$4500Hz$ range. NHR is defined as the average ratio of the inharmonic spectral energy in the 1500-$4500Hz$ range to the harmonic spectral energy in the 70-$4500Hz$ range. The power ratio is the ratio of the power spectrum below $800Hz$ to the power spectrum above $800Hz$. The authors found that simple snorers had a high SPI value, while OSA snorers had a high NHR and a low power ratio.

Herzog *et al.* [90] studied the peak intensity of the power spectrum using 60 subjects (60 m; mean age = $50yrs$; mean BMI = $29.6kg/m^2$; 18 subjects had an AHI≥$10events/h$). Simple snoring had a low frequency pattern (peak intensities between 100-$300Hz$) and rhythmic periodicity, while apnoea events had a high frequency pattern (peak intensities above $1000Hz$) and non-rhythmic periodicity. A raised AHI correlated significantly with an increase in peak intensity of the FFT curve ($p < 0.001$). A number of acoustic properties have been used to try to classify OSA including noise to harmonics ratio [89], peak intensity [90], formant frequencies [83, 91] and phase coupling relations [91–93].

It should be noted that the above techniques assume that the signals are stationary/quasi-

stationary; an assumption which can be false depending on the length of the speech/snoring signal. However, the signal can be assumed to be quasi-stationary when windowed at an appropriate length. Wavelet analysis is a suitable method for analysing signals which are non-stationary. Ng *et al.* [94] developed a noise reduction preprocessing system for snore signal enhancement and snore activity detection in a laboratory setting. They used the snoring sounds of 30 snorers with OSA (24 m, 6 f; age = 44±13yrs; BMI = 29.3±6.9kg/m^2; AHI = 46.9±25.7$events/h$) and 10 snorers without OSA (6 m, 4 f; age = 41±12yrs; BMI = 26.9±5.6kg/m^2; AHI = 4.6±3.4$events/h$). The average SNR of the de-noised snoring signal was improved. When the power spectrum of snore signals was generated by the FFT and a peak frequency of 232Hz was used to distinguish between apnoeic snorers and benign snorers, the results obtained were Se = 45% and Sp = 86%. When the power spectrum was generated via the Welch's method, the peak frequency was 245Hz resulting in Se = 86% and Sp = 70%.

Matsiki *et al.* [95] used the continuous wavelet transform to analyse the snoring signals of seven OSA patients (6 m, 1 f; age = 56±11.4yrs; BMI = 33.4±5.0kg/m^2; AHI = 35.84±17.15$events/h$). They found that for each patient there was a frequency region which was almost always active and could be used for diagnosis.

3.2.3 Hidden Markov Models

Hidden Markov Models (HMMs) model the system as a number of states with unknown parameters. The challenge is to determine the hidden parameters from the observable data. Features are derived and combined from the input signal via a statistical framework that leads to a decision or the labelling of the input signal.

Duckitt *et al.* [96] recorded the sounds of six self-acknowledged snorers (4 m, 2 f; age range = 43-75yrs) sleeping in their own homes. The data were parametrised with mel-frequency cepstral coefficients (MFCCs)[3] [97], with 12 MFCCs calculated every 10ms using a 30ms window of data. HMMs were used to model the different types of sounds. The audio data were manually and automatically segmented (by experts and the HMM approach) into periods of 'snoring', 'silence', 'breathing', 'noise from bed clothing movement' and 'other

[3]MFCCs make up a mel-frequency cepstrum which is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. The mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another.

noise'. The authors found that the Se was 89% when the HMMs were trained and validated on the data from all six subjects; when the HMMs were trained on three subjects and validated on the other three subjects the system had 82% Se when compared to the expert manual annotations. The latter approach is more representative of how the system is likely to perform on unseen patients, although the population is still very small and the results must therefore be treated with suspicion.

3.2.4 Energy Distribution

Cavusoglu *et al.* [98] analysed the energy distribution in order to distinguish between snoring and non-snoring events. The study used 30 subjects: 12 OSA patients (12 m; age = 53.26yrs (range 44.87-61.65); BMI = 32.76kg/m^2 (range 27.47-38.05); AHI = 39.21$events/h$ (range 22.17-56.25)) and 18 simple snorers (16 m, 2 f; age = 46.92yrs (range 40.21-53.63); BMI = 27.66kg/m^2 (range 23.41-31.91); AHI = 4.29$events/h$ (range 3.03-5.55)) to study the 500Hz sub-band energy distribution to identify snoring and non-snoring events. The 0-7500Hz frequency range was divided equally into non-overlapping 500Hz sub-bands and the average normalised energy for each sub-band was calculated for each sound activity episode, where an episode was classified as snore or non-snore. The data were divided equally into train and validation sets and the classification boundary was found by linear regression. Snoring episodes exhibited a regular pattern in the spectrogram and could be easily distinguished from other sounds. For simple snorers, the algorithm had a Se of 90.2% and a PPV of 98.7% while for OSA patients Se was 86.8% and PPV was 93.8%.

Jones *et al.* [99–101] studied a number of acoustic parameters: snore duration, snore loudness, snore periodicity and sub-band energy distribution. There were 20 subjects involved in this study (18 m, 2 f; age = 46yrs (range 33-65); BMI = 31.6kg/m^2 (range 26.9-44.1)). The total energy in the 0-200Hz, 0-250Hz and 0-400Hz frequency ranges was estimated and expressed as a ratio of the total energy level of the recording. The results were used to determine whether palatal surgery had been successful. The Pringle & Croft grading (using objective methods) had Sp = 62.5%, Se = 50%, PPV = 66.6% and NPV = 45.5%; the Camilleri *et al.* grading (using objective methods) had Sp = 37.5%, Se = 91.7%, PPV = 68.7% and NPV = 75%. The different gradings can be found in Table 3.2, where the Camilleri *et al.* grading

is a simplified version of the Pringle & Croft grading. The simplification was based on the observation that uvulopalatopharyngoplasty alone will not cure patients with grade three, or greater, sleep nasendoscopy results.

Table 3.2: *Sleep nasendoscopy grading systems proposed by Pringle & Croft, later simplified by Camilleri et al..*

Pringle & Croft [102]	
Grade 1	simple palatal snoring
Grade 2	single level palatal obstruction
Grade 3	multi-segment involvement
Grade 4	sustained multi-segment collapse
Grade 5	tongue base level obstruction
Camilleri *et al.* [103]	
Grade 1	palatal snoring
Grade 2	mixed snoring
Grade 3	non-palatal (tongue base) snoring

3.2.5 Pitch

Pitch is associated with the vibration frequency of the vocal cords and is the psycho-acoustic equivalent of the fundamental frequency. Speech, and therefore pitch, change over time and so it is common to track pitch over time [10].

Abeyratne *et al.* [104] divided snore-related sounds into pure breathing, silence, voiced snores and unvoiced snores. Voiced components, $s_{sv}(n)$, were separated from unvoiced ones, $s_{suv}(n)$, using pitch information particular to s_{sv}. A number of parameters in the pitch information could be changed, and when applied to a clinical database of 16 patients (8 m, 8 f; age = $52 yrs$ (range 36 - 71); AHI = $30.3 events/h$ (range 3.3 - 85.7)), the Se ranged from 86% to 100%, with Sp remaining between 50% and 80%. Abeyratne *et al.* [105] divided snore-related sounds into three main classes: benign snoring (BS), apnoeic snoring (AS) and speech. The authors analysed snoring sounds from 14 patients, of whom half were classified as AS and the other half as BS and found that the data could be separated into AS class with 92% Ac and into BS with 90% Ac, while the separation of speech from the rest of the data was 100% accurate.

3.2.6 Higher Order Statistics

Higher order statistics, or cumulants, and their associated Fourier transforms, or polyspectra, reveal both amplitude and phase information about a process [106].

Ng *et al.* [93] studied nine OSA patients (age = 47±18yrs; BMI = 29±7kg/m^2; AHI = 41±19$events/h$), and seven simple snorers (age = 38±11yrs; BMI = 27±6kg/m^2; AHI = 4±3$events/h$). The raw snore signals were denoised using a modified level-wavelet-dependent thresholding scheme under an undecimated wavelet environment. Non-linear properties in the noise-suppressed snore signals were extracted to discriminate between apnoeic and simple snorers. The authors found that apnoeic snores exhibited a higher degree of phase coupling phenomena than simple snores (77% of benign snores indicated the presence of self-coupling, compared to only 49% of apnoeic snores).

3.2.7 Other Methods

Shepard *et al.* [107] noted that the diagnosis of OSA may be more accurate if any structural and/or functional abnormalities of the UA are known. The SNAP testing system (SNAP Laboratories, Glenview, Illinois) is a reflective acoustic device that may be used for screening and analysis to locate the source of snoring and detect sleep apnoea conditions. A number of studies have been carried out comparing the SNAP testing system to conventional PSG. Liesching *et al.* [108] found that SNAP did not assess the severity of OSA correctly compared to conventional PSG. Michaelson *et al.* [109] found that for an AHI≥ 15 SNAP had Se = 100%, Sp = 88.5%, PPV = 57% and NPV = 100%. Su *et al.* [110] found that 20% of patients were classified incorrectly using the SNAP system. Galer *et al.* [111] focused on the audio channel and found that analysing snoring has limited use in the evaluation of patients with sleep apnoea.

An effort has been made recently to use snoring to estimate the AHI. Solá-Soler *et al.* [112] analysed the sounds of 36 subjects (25 m, 11 f; age range = 23-69yrs; AHI range = 0-90.8$events/h$). Snoring sounds were automatically identified and both time and frequency domain features were computed. The authors found that they could classify into AHI< 5, 5≤AHI<30 and AHI≥30 with 83.3% Ac using these features. Ben-Israel *et al.* [113] automatically identified the snoring sound from 90 subjects (57 m, 33 f; age = 53±13yrs; BMI = 31±4kg/m^2) and calculated a variety of features (MFCCs, pitch density, etc.). These features

correlated well with the AHI calculated from the PSG ($r^2 = 0.81, p < 0.001$) and achieved an AUC of 0.85 and 0.92 for thresholds of 10 and 20$events/h$, respectively, for OSA detection. Fiz *et al.* [114] automatically identified the snores of 37 snoring subjects (25 m, 12 f; age range = 40-65yrs; BMI = 29.65±4.7kg/m^2). The number of snores, average intensity and power spectral density parameters were calculated for each subject, who were then classified with AHI = 5 and AHI = 15 thresholds giving Se (Sp) of 87% (71%) and 80% (90%), respectively.

3.3 Actigraphy

Actigraphy is often used in assessing sleep, usually for determining whether the subject is asleep or awake in a given epoch. However, there is much written on whether actigraphy is actually capable of reliably detecting awakenings, particularly in specific sub-populations.

Ancoli-Israel *et al.* [115] identified the role that actigraphy plays in sleep and circadian rhythms. The authors found that actigraphy is used in four main areas: studying the validity of actigraphy; using actigraphy in populations with sleep disorders; using actigraphy to study circadian rhythms; using actigraphy as a treatment outcome measure. The authors identified a number of studies that attempt to detect OSA using actigraphy. In all cases, this is based on the fact that apnoeics have more fragmented sleep compared to normal subjects. This fragmentation is manifested in body movements that are detected by the actigraph. The authors concluded that actigraphy, due to its greater sensitivity, is better than sleep logs at detecting brief arousals from sleep. Actigraphy is capable of distinguishing moderate to severe sleep apnoea patients from normal controls.

3.3.1 Sleep-wake Identification

Much of the research performed on sleep-wake identification and algorithm development is based on the work of Webster *et al.* [116]. The authors found that an algorithm that summed changes in activity level over a 2s interval was the most sensitive. EEG and activity sleep-wake scores agreed 94% of the time.

Cole *et al.* [117] developed an automatic scoring method to distinguish sleep from wake-

fulness based on wrist activity. 41 subjects (32 m, 9 f; age = 50.2±14.7yrs), 18 were normal and 23 had sleep or psychiatric disorders, wore a wrist actigraph during PSG. The algorithm computed a weighted sum of the activity in the current minute, the preceding four minutes and the following two minutes as follows:

$$S = 0.0033(1.06A_{-4} + 0.54A_{-3} + 0.58A_{-2} + 0.76A_{-1} + 2.3A_0 + 0.74A_{+1} + 0.67A_{+2})$$

where $A_{-4}, A_{-3}, A_{-2}, A_{-1}$ are the activity counts from the previous four minutes respectively, A_0 is the activity count from the current minute, and A_{+1} and A_{+2} are the activity counts of the following two minutes. The current minute is scored as sleep when $S < 1$. The algorithm correctly distinguished sleep from wakefulness 88% of the time in the validation data set.

Sadeh *et al.* [118] analysed actigraph data from 20 adults (9 m, 11 f; age = 22.6±1.7yrs) and 16 adolescents (8 m, 8 f; age = 13.8±1.9yrs) who wore actigraphs on both wrists. The authors found that, although activity levels differed between the dominant and non-dominant wrists during sleep and wake, the sleep-wake scoring algorithm was equally explanatory (R^2 = 0.64, $p < 0.0001$) and agreement between wrist actigraphy and PSG ranged between 91% and 93% for both wrists. The sleep-wake scoring algorithm the authors developed was:

$$P_S = 7.601 - 0.065M_5 - 1.08N - 0.056\sigma_6 - 0.073 \ln A_c$$

where P_S is the probability of sleep, M_5 is the average number of activity counts during the scored epoch and a window of five epochs preceding and following it, N is the number of epochs with activity level equal to or higher than 50 but lower than 100 activity counts in a window of 11$mins$, including the scored epoch and the five preceding and following it, σ_6 is the standard deviation of the activity counts during the scored epoch and the five epochs preceding it, $\ln A_c$ is the natural logarithm of the number of activity counts during the scored epoch+1. If $P_S \geq 0$ the epoch is scored as sleep; otherwise it is scored as wake.

DeSouza *et al.* [119] compared Cole's [117] and Sadeh's [118] algorithms with PSG. 21 healthy volunteers (7 m, 14 f) underwent two nights of simultaneous PSG and actigraphic measurement. Only the second night was studied. Each 30s epoch of PSG data was automatically classified according to the R&K rules. These 30s epochs were then pooled into

one minute intervals and classified as wake or sleep. Wake was taken to be two consecutive wake epochs or two epochs of sleep-wake or wake-sleep; sleep was taken to be two consecutive sleep epochs. The actigraphy data were collected using the zero-crossing method in one minute epochs. Both algorithms correctly classified 91% of the activity epochs with the same state (sleep or wake) as the PSG epochs, with Se of 99% (Sp = 34%) and 97% (44%) for Cole's and Sadeh's algorithms respectively. Actigraphy consistently overestimated sleep latency, total sleep time (TST) and sleep efficiency (SE) while underestimating intermittent awakenings. The authors concluded that in assessing sleep, actigraphy has its uses; however, it is limited in identifying wake epochs during sleep.

Paquet *et al.* [120] evaluated the ability of actigraphy to detect wakefulness in healthy subjects with different amounts of wakefulness. 15 healthy subjects (7 m, 8 f; age range = 20-60yrs) underwent non-dominant wrist actigraphy with simultaneous full PSG. Four sleep-wake scoring algorithms were used: two were threshold-based algorithms provided by the manufacturers of the actigraphs used in the analysis and the remaining two were derived from Lötjönen *et al.*'s [121] regression analysis based method (which are different versions of Sadeh's algorithm). When actigraphy and PSG were compared epoch-by-epoch, there was a significant decrease in actigraphy accuracy with increased wakefulness in sleep conditions. The authors postulated that the reason for this was the low specificity of actigraphy. TST and SE were overestimated more in conditions involving more wakefulness. The authors concluded that the use of actigraphy in detecting wakefulness in populations with fragmented sleep or when the sleep-wake cycle is disrupted may not be valid.

Kushida *et al.* [122] compared PSG derived sleep parameters (TST, SE and number of awakenings) with those derived from actigraphy and subjective questionnaires. 100 SDB subjects (69 m, 31 f; age = 49±14.7yrs) were studied. An epoch-by-epoch comparison was carried out between PSG data and actigraphic data. The subjects were asked to fill out a subjective questionnaire the following morning. The study found that TST and SE did not differ significantly between PSG and actigraphy+subjective questionnaire. The authors also found that when a low wake-sensitivity algorithm was used, the number of actigraphic awakenings detected was not significantly different from those detected by PSG. The algorithm used was

provided by the analysis software and was calculated as follows:

$$A = 0.04E_{-4} + 0.04E_{-3} + 0.20E_{-2} + 0.20E_{-1} + 2E + 0.20E_{+1} + 0.20E_{+2} + 0.04E_{+3} + 0.04E_{+4}$$

where A is the sum of activity counts for $30s$ scored epoch and surrounding epochs, E is the activity counts recorded during scored epoch, E_n is the activity counts recorded during each previous or following epochs. If the summed activity counts (A) is greater than a defined threshold (T), the epoch is scored as wake, otherwise the epoch is scored as sleep. A variety of thresholds were used and the results were reported for a high threshold (low wake-sensitivity).

Drinnan *et al.* [123] determined whether the placement of the actigraph affected its accuracy in identifying arousals associated with SDB in 36 subjects. Although the left and right tibia, left ankle and left wrist were studied, no statistically significant correlation with EEG arousals were found. Ancoli-Israel *et al.* [115] noted that there was relatively low severity of sleep apnoea in this study which may limit the ability to distinguish between groups.

The fundamental assumption of sleep identification as the absence of movement introduces a significant problem in the detection of quiet wakefulness by actigraphy. A wakefulness detection specificity of 35-50% is often reported, especially with increased subject wakefulness [120, 124] and this affects all derived sleep characteristics. Special care needs to be taken when using actigraphy for sleep analysis in subjects with limited mobility and serious sleep disturbances. It should be noted that the results presented here have been obtained using older generation uni-directional actigraphs and newer devices may allow for the development of more sensitive algorithms.

3.3.2 OSA Detection

Middlekoop *et al.* [125] analysed wrist actigraphy of 116 subjects suspected of OSA. Simultaneous ambulatory oronasal flow thermistry, non-dominant wrist actigraphy and a sleep log were recorded. The subjects were divided by their apnoea index (AI): AI<1 44%; AI 1-5 39%; AI≥5 17%. Higher AI values corresponded with self reported disturbed sleep initiation and more fragmentation. Across subjects, the duration of immobility periods was the only predictor of the AI. The subjects were classified into those with an AI<1 and ≥5 giving Se (Sp) of 75% (43%) and 5% (100%) respectively.

Elbaz *et al.* [126] studied whether adding actigraphy to simplified polygraphy would improve AHI evaluation when compared to simplified polygraphy alone. 20 adults (15 m, 5 f; age = 52±15yrs; BMI = 28±5kg/m^2) with suspected OSA underwent simultaneous full PSG as well as wearing a wrist actigraph. Three different AHI's were calculated: AHI-pg (from PSG), AHI-tib (from simplified polygraphy) and AHI-act (from actigraphy). The actigraph data were assessed using an algorithm provided by Sleepwatch sleep analysis software for Windows (Cambridge Neurotechnology Ltd.) where the Actiwatch arousal threshold was set at an integrated activity count of 40 movements within a 1min epoch. AHI-pg indicated that 12 subjects had OSA (AHI>10), with eight subjects having severe OSA (AHI≥ 30). AHI-act was more closely correlated with AHI-pg (r = 0.976) than with AHI-tib (r = 0.940). When using AHI-act and AHI-tib to diagnose severe OSA, Se was 88% (NPV = 92.5%) and 50% (NPV = 75%) respectively. The authors concluded that a system combining simplified polygraphy and actigraphy could assist in the diagnosis of OSA.

Ayas *et al.* [127] assessed the accuracy of wrist actigraphy (using a Watch_PAT100) to diagnose OSA. 30 subjects with and without suspected OSA (age = 47.0±14.8yrs; BMI = 31.0±7.6kg/m^2) underwent simultaneous in-lab PSG and actigraphy. The Watch_PAT100 recorded four signals: peripheral arterial tonometry (PAT), SpO_2, heart rate, and actigraphy. The data were analysed using an automated computerised algorithm that calculated the frequency of respiratory events per hour of actigraphy-determined sleep. For an AHI of 10, 15, 20 and 30 the Watch_PAT100 AHI gave a Se (Sp) of 82.6% (71.4%), 93.3% (73.3%), 90.0% (84.2%) and 83.3% (91.7%) respectively. The authors concluded that the Watch_PAT was able to detect OSA with reasonable accuracy.

Hedner *et al.* [128] developed a novel automatic algorithm to detect sleep-wake in subjects with OSA. 228 subjects (163 m, 65 f; age = 48.8±14.0yrs; BMI = 29.4±6.3kg/m^2; RDI = 30.3±23.1$events/h$) underwent simultaneous PSG and wrist actigraphy. Overall, the algorithm identified sleep with Se = 89% and Sp = 69%; agreement was 86%, 86%, 84% and 80% in normal, mild, moderate and severe cases respectively. The authors found that there was a tight agreement between actigraphy and PSG in determining SE (78.4±9.9% vs. 78.8±13.4%), TST (690 ±152 epochs vs. 690±154 epochs), and sleep latency (56.8±31.4 epochs vs. 43.3±45.4 epochs). The authors concluded that the algorithm could be useful in

assessing TST for identifying OSA in the home.

Kim *et al.* [129] compared three algorithms used to detect sleep/wake status on subjects with OSA. The algorithms evaluated were the Cole, Sadeh and University of California San Diego (UCSD) algorithms. 101 OSA patients (84 m, 17 f; mean age = 49.4yrs (range 19-73)) with an AHI≥5 were used in this study. Full PSG and non-dominant wrist actigraphy data were recorded simultaneously. The authors found that mean TST was not significantly different from PSG when the UCSD algorithm was used ($p = 0.798$). The correlation levels with PSG data were: UCSD, $r = 0.498, p < 0.001$; Cole, $r = 0.389, p < 0.01$; Sadeh, $r = 0.272, p = 0.057$. Actigraphy mean TST underestimated sleep in patients with an AHI≥ 30 but was overestimated in patients with 5 ≤AHI< 15 or 15 ≤AHI< 30. The authors concluded that none of the algorithms were reliable enough to estimate sleep time in subjects with SDB or severe OSA.

Littner *et al.* [130] were tasked with formulating a guide to the appropriate use of actigraphy based on the available research. The authors concluded that, although actigraphy may be a useful extra for portable sleep apnoea testing, the use of actigraphy for SAS detection on its own had not been established. Sadeh [124, 131] reviewed the role of actigraphy in sleep medicine, both as a sleep assessment and a diagnostic tool. Overall, Sadeh concluded that actigraphy is valid and reliable in normal subjects with good sleep patterns. However, in subjects with sleep disorders or in specific sub-populations, the use of actigraphy is more questionable. The author raises the important issue of the low specificity of actigraphy in detecting wakefulness within sleep periods as reported by some devices. In terms of SDB and SAS, the literature suggests that actigraphy is in fact, not reliable in this subject population, although recent research suggests that actigraphy may assist in diagnosing SDB.

3.3.3 Other Uses

Chin *et al.* [132] analysed the relationship between severe OSA, metabolic syndrome (Mets) and short sleep duration. 275 Japanese males (age = 44±8yrs; BMI = 23.9±3.1kg/m^2) were asked to fill in a questionnaire, wear an actigraph for seven days and use a type 3 portable monitor (Somté, Compumedics, Victoria, Australia) for two nights at home. There was a significant relationship between OSA severity and the prevalence of Mets ($p < 0.001$); however,

the association was not significant after adjustments were made for age and BMI. Subjects with severe OSA have significantly shorter sleep duration ($p < 0.05$); and sleep duration in Mets subjects was also significantly shorter than in those without ($p < 0.05$). The conclusion was that sleep duration should be taken into consideration as an important factor in studies investigating the prevalence of severe OSA and Mets.

Morgenthaler *et al.* [133] recommended that actigraphy could be used to estimate TST in patients with OSA; but only where PSG is not available.

3.4 Pulse Oximetry

Together with ECG, pulse oximetry, derived from the photoplethysmogram (PPG), is the most widely used technique for at-home sleep monitoring and in simplified PSG systems. The main use of PPG in sleep studies is the measurement of SpO_2, either for sleep apnoea alarm systems or for OSA detection. Allen [134] identified the applications for PPG in clinical physiological measurement. The PPG is comprised of a pulsatile wave, which is attributed to cardiac synchronous changes in the blood volume with each heart beat, with a slowly varying baseline superimposed on top where various lower frequency components are attributed to respiration, sympathetic nervous system and thermoregulation. The PPG has been used in recent years in a wide range of medical devices for measuring SpO_2, blood pressure and cardiac output to mention a few uses.

There is extensive research on the derivation of heart rate (HR) from the PPG. Existing methods usually compute HR by upsampling the PPG signal and detecting peaks or zero crossings, sometimes including artefact-rejection algorithms [134]. An open-ended question is whether the variability of the PPG derived HR (PHR) accurately reflects the heart rate variability (HRV) as measured using the ECG in sleep studies. Recent research indicates that PHR variability and HRV indices could be significantly different during OSA [135]. The authors recorded ECG and PPG simultaneously from 29 healthy subjects and 22 OSA patients. The HR and PHR were significantly correlated ($r > 0.95, p < 0.01$). Comparing $2min$ epochs demonstrated significant differences ($p < 0.01$) between normal and OSA events using PHR variability and HRV measures.

A number of signal processing algorithms have been proposed to estimate the respira-

tion rate from the PPG. This is possible because respiration causes variation in the peripheral circulation, which is reflected in the PPG as a low-frequency component [134]. Flemming & Tarassenko [136] devised a novel way of determining breathing rate using the PPG. Seven records, comprising of 14×5min sections, of PPG and a synchronous respiratory waveform were used. An autoregressive model was developed that achieved a mean error of 0.04$breaths/min$. However, most existing methods have been validated in healthy population, which may preclude their use in SDB patients [134].

Mendez *et al.* [137] analysed the relationship between PPG, HRV and apnoea. Cardiac oscillations correlated to decrements in the amplitude of the PPG signal (DAP) were analysed during normal and apnoeic sleep. 268 ECG excerpts from 12 patients were divided into five groups depending on SpO_2 and respiratory behaviour during DAP events. The results showed that there is an increase in sympathetic activity during DAP events; and the increase was greater when DAP events were associated with respiratory or SpO_2 variations.

Scully *et al.* [138] showed that it is possible to record physiological parameters using optical recording from a mobile phone. They analysed the varying colour signals of a fingertip placed over an optical sensor. Participants were asked to place the palmar side of the left index finger over the camera lens while the flash was on. Videos of the fingertip were recorded. The videos were then analysed to determine HR, respiration rate and SpO_2. The data recorded with the phone were compared with standard methods. The mean$\pm\sigma$ for the HR derived from ECG was 92.2±5.3bmp and 92.3±5.9bmp for the HR derived from the phone. Breathing rates from the respiration trace and the phone were estimated at three metronome rates (0.2, 0.3 and 0.4Hz) as 0.18 and 0.16, 0.30 and 0.32, and 0.40 and 0.38Hz respectively. The SpO_2 was computed using the red and blue phone channels and compared to a commercial pulse-oximeter. The phone SpO_2 decreases appeared to correlate with the commercial pulse-oximeter SpO_2 decreases.

Bennett & Kinnear [139] identified the role of oximetry in the diagnosis of OSA. The authors noted that oximetry has been used to diagnose OSA because it is readily available, relatively inexpensive and can be used in the home thus allowing the subject to experience a typical night's sleep. However, it is not possible to distinguish between saturation occurring secondary to obstructive apnoea, central apnoeas, primary pulmonary disease and cardiac

disease using oximetry. The authors suggested that using a cut-off of ODI> 5 allows for reasonable positive screening when taken with other simple measures. The overall conclusion was that oximetry is insufficient on its own to diagnose OSA.

3.4.1 ODI Based Methods

One of the most common quantitative indices derived from PPG is the ODI. In order to mirror the definition of an abnormal AHI, similar cut-off points for an abnormal ODI have been proposed (either 5, 10, 15, 20 or 30 desaturations per hour). However, there is little evidence of one definition having greater validity than the others [140]. Chiner *et al.* [141] used nocturnal oximetry for the diagnosis of OSA. 275 subjects with suspected OSA underwent simultaneous PSG and nocturnal oximetry. OSA was identified in 216 subjects (194 m, 22 f). Subjects with abnormal lung function were removed and for three different ODI levels (≤5, ≤10 and ≤15) the results were Se = 80%, Sp = 89%, PPV = 97%, NPV = 48%, Ac = 81%; Se = 71%, Sp = 93%, PPV = 97%, NPV = 42%, Ac = 75%; Se = 63%, Sp = 96%, PPV = 99%, NPV = 38%, Ac = 70% respectively for nocturnal oximetry. The authors conclude that nocturnal oximetry is useful in subjects with suspected OSA who have normal spirometric values.

Fietze *et al.* [142] investigated the night-to-night variation in ODI for OSA subjects. 35 patients (32 m, 3 f; age = 58±11yrs; BMI = 26.0±3.0kg/m^2) were monitored at home for seven consecutive nights. The diagnostic accuracy of the system used was compared with the PSG outcome of 18 patients. For the home recordings the median ODI was 10.9 across all 35 patients. The reliability of the ODI was adequate, however there was a 14.4% probability of placing the patient in the wrong severity category (ODI≤15 or ODI>15) when a single recording was taken. The authors found that ODI variability was not significantly influenced by age, BMI, time spent in the supine position or mild alcohol consumption. The authors concluded that the ODI could be used as a screening tool based on a single night recording due to the small variability.

Vásquez *et al.* [143] automatically analysed oximetry to diagnose OSA. 245 subjects (192 m, 53 f; age = 45±11.3yrs; BMI = 30.8±5.9kg/m^2; neck = 40.3±3.7cm) with suspected OSA underwent PSG and the oximeter signal was analysed off-line. The PSG derived AHI and the oximeter derived RDI were highly correlated ($R = 0.97$). Using a cut-off of 15$events/h$ for

both AHI and RDI, the results achieved were Se = 98% and Sp = 88%. The authors concluded that off-line automated analysis of the oximetry provides a close estimate of AHI as well as good diagnostic sensitivity and specificity for OSA.

3.4.2 Other Methods

Al-Angari & Sahakian [144] analysed the PSG data from 50 OSA patients (AHI = 36.86±24.32 $events/h$) and 50 control subjects (AHI = 1.1±1.48$events/h$). The thoracic and abdominal respiratory effort signals, ECG and SpO_2 signals were extracted and divided into 1min segments, or analysed by subject. Respiratory features and SpO_2 features were analysed and for the minute classification, the respiratory features had the highest sensitivity while the SpO_2 features gave the highest specificity. When classifying on a subject basis, the SpO_2 features performed as well as the combination of features, Ac = 95%, Se = 100%, Sp = 90.2% and Ac = 95%, Se = 91.8%, Sp = 98% respectively.

Álvarez *et al.* [145] analysed the SpO_2 from nocturnal pulse oximetry to diagnose OSA. 148 subjects with suspected OSA (115 m, 33 f; age = 52.9±14.1yrs; BMI = 29.8±5.6kg/m^2) underwent full PSG and each 30s epoch was scored according to the R&K rules. 100 subjects were diagnosed with OSA (84 m, 16 f; age = 55.2±14.6yrs; BMI = 30.8±5.0kg/m^2; AHI = 40.9±27.6$events/h$), *i.e.* they had an AHI≥10, while 48 did not have OSA (32 m, 16 f; age = 48.3±11.8yrs; BMI = 27.3±6.3kg/m^2; AHI = 4.1±2.4$events/h$). 16 features from the time and frequency domain were computed including the first-fourth-order statistical moments in the time domain ($M1_t$:$M4_t$), first-fourth-order statistical moments in the frequency domain ($M1_f$:$M4_f$), median frequency, spectral entropy, total spectral power (P_T), local maximum of the apnoea frequency range (0.014-0.033Hz) (PA), relative power in the apnoea frequency band (P_R), sample entropy , central tendency measure (CTM) and Lempel-Ziv complexity (LZC). Feature selection was carried out using step-forward logistic regression with leave-one-out cross-validation. The optimal feature set ($M2_t$, $M4_t$, P_R and LZC) achieved Se = 92%, Sp = 85.4% and Ac = 89.7%.

Hornero *et al.* [146] applied approximate entropy ($\mathcal{H}_A$) to the SpO_2 signals from overnight pulse oximetry in order to determine whether it could be used to diagnose OSA. 187 subjects were analysed (111 OSA, 76 non-OSA). A training set of 44 OSA and 30 non-OSA sub-

jects (56 m, 18 f; age = 58.25±12.14yrs; BMI = 29.62±5.71kg/m^2) was used for algorithm development and optimal threshold selection. The data were sampled at 0.2Hz and divided into epochs of 200 samples (1000s). $\mathcal{H}_A$ was applied with a number of values for the parameters m and r. A validation set of 67 OSA and 46 non-OSA subjects (91 m, 22 f; age = 57.91±13.39yrs; BMI = 29.49±5.41kg/m^2) obtained Se = 82.09% and Sp = 86.96% using $m = 1, r = 0.25$.

Morillo *et al.* [147] analysed the SpO_2 using Poincaré plots to diagnose OSA. Poincaré plots are a geometrical representation of a time series into a Cartesian plane where the values of each pair of successive elements of the time series define a point in the plot. SpO_2 signals from 117 subjects (54 subjects with AHI≥ 15) were analysed; dividing into train data set of 70 subjects (54 m, 16 f; age = 59.1±9.8yrs; BMI = 30.5±6.4kg/m^2; 32 OSA, 38 non-OSA) and a validation data set of 47 subjects (35 m, 12 f; age = 58.1 ±12.5yrs; BMI = 32.2±7.4kg/m^2; 22 OSA, 25 non-OSA). The Poincaré plots were generated using the SpO_2 signals, which were recorded at 8Hz but filtered using a 30^{th} order FIR filter. The train data set was used to tune thresholds of the Poincaré descriptors which were the standard deviation around the Y-axis (σ_1) (indicating the level of short-term variability of the SpO_2 signal), the standard deviation around the X-axis (σ_2) (related to the long-term variability of the SpO_2 signal). σ_1 and σ_2 allow an ellipse to be fitted to the Poincaré cloud and thus the area to be calculated ($A = 2\pi \times \sigma_1 \times \sigma_2$). On the validation set, the results were Se = 90.0%, Sp = 84%; indicating that Poincaré analysis could be useful in screening for OSA.

Sepúlveda-Cano *et al.* [148] used PPG envelope-based dynamic features for detecting OSA in children. 21 PSG recordings of children (age = 4.5±2yrs) with suspected SDB were analysed. A time-evolving version of the standard liner multivariate decomposition was used to perform stochastic dimensionality reduction. Using a subset of cepstral-based dynamic features it was possible to classify patients with Ac = 83.3%.

Gil *et al.* [149] automatically detected sleep apnoea in children using decreases in PPG amplitude fluctuations. 25 children (16 m, 9 f; age = 4.56±1.79yrs) with suspected OSA underwent full PSG. A detector was developed which preprocessed the PPG using a moving average filter. An envelope detector was used (RMS detector and Hilbert transform detector) to get an adequate signal for comparison with a threshold. A decision was made using an

adaptive threshold, where an event was classed as apnoeic when the envelope is lower than the established threshold. Sensitivity and PPV were 76% and 73% respectively, while the PPG attenuation events per hour ratio E_h classified children as normal ($13.5 \pm 6.35 events/h$) or pathologic ($21.1 \pm 8.93 events/h$) with $p < 0.05$. In another work, Gil *et al.* [150] analysed HRV during DAP events for OSA. PSG recordings from 21 children (10 OSA, 11 normal; age = $4.47 \pm 2.04 yrs$) were analysed, although only 15 were used in the final analysis (8 OSA, 7 normal). DAP events were classified as apnoeic or non-apnoeic and the ratio of DAP events per hour to the ratio of apnoeic DAP events per hour was calculated. Using this ratio to classify subjects achieved Ac = 80%, Se = 87.5% and Sp = 71.4%. Using the same subject population, Gil *et al.* [151] then analysed the pulse transit time variability (PTTV) during DAP events. The authors found that PTTV reflects sympathetic changes more clearly than HRV. DAP events were classified as apnoeic or non-apnoeic from the PTTV indices. The ratio of DAP events per hour, the ratio after filtering based on HRV indices and the ratio after filtering based on PTTV indices were computed. The highest accuracy was obtained using DAP events filtered using PTTV indices (Ac = 75%, Se = 81.8%, Sp = 73.9%) when classifying $1h$ PSG excerpts as OSA or normal.

Monasterio *et al.* [152] used features from ECG, impedance pneumogram and PPG to reduce the false alarm rate for neonatal apnoeas. A total of 1616 desaturations from 27 neonates were annotated. Features related to SpO_2, HR, respiratory rate and signal quality were computed every $5s$ for the $300s$ interval before each desaturation. Feature selection occurred using maximum relevance minimum redundancy and the best classification performance on the test set was Se = 86%, Sp = 91% and Ac = 90%.

3.5 Summary

Table 3.3 summarises key previous studies which detected OSA using audio, actigraphy or PPG. It is clear that using audio to detect OSA can achieve 90% accuracy at most, with a wide range of sensitivities and specificities. Actigraphy can detect OSA with 93% sensitivity, however, in some instances there is a compromise between sensitivity and specificity, *i.e.* one is very high while the other is very low. This is common in the literature. PPG can detect OSA with up to 95% accuracy. Unlike actigraphy, there is no discrepancy between sensitivity and

specificity. It should be noted that none of the studies referenced in this chapter have more than 300 subjects in them. In many cases, there are less than 100 subjects.

[4]The accuracy of the *Hawke Index* can be calculated from the statistics given: $Ac = (Se)(prevalence) + (Sp)(1 - prevalence)$ where the prevalence is the number of people in the population with the disease divided by the total population, in this case $\frac{9}{9+18}$, so the accuracy is 89%.

Table 3.3: *Summary of previous studies on detecting OSA. Training = in-sample, Val (validation) = out-of-sample, N/A = not available*

Method	Paper	Subjects	Data Set	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)
Audio								
LPC	Ng *et al.* [83]	30 OSA, 10 snorers	Val	88	82	N/A	N/A	N/A
Hawke Index	McCombe *et al.* [86]	9 OSA, 18 snorers	Training	67	100	100	86	89[4]
Power spectrum	Ng *et al.* [94]	30 OSA, 10 snorers	Training	86	70	N/A	N/A	N/A
HMM	Duckitt *et al.* [96]	6 snorers	Val	82	N/A	N/A	N/A	N/A
Energy distribution	Cavusoglu *et al.* [98]	12 OSA	Val	86.9	94.8	93.8	88.7	86.8
		18 snorers	Val	90.3	98.8	98.7	90.9	90.2
Energy distribution	Jones *et al.* [99–101]	20 unspecified	Training	50	62.5	66.6	45.5	N/A
Pitch	Abeyratne *et al.* [104]	16 unspecified	Val	86	67	N/A	N/A	N/A
Active devices	Michaelson *et al.* [109]	N/A	Training	100	88.5	57	100	N/A
Actigraphy								
Duration of immobility	Middlekoop *et al.* [125]	20 OSA, 96 non-OSA	Val (AI < 1)	75	43	N/A	N/A	N/A
			Val (AI ≥ 5)	5	100	N/A	N/A	N/A
Activity counts within 1*min*	Elbaz *et al.* [126]	12 OSA, 8 non-OSA	Val	88	N/A	N/A	92.5	N/A
Frequency of respiratory events	Ayas *et al.* [127]	25 OSA, 5 non-OSA	Val (AHI = 10)	82.6	71.4	N/A	N/A	N/A
			Val (AHI = 15)	93.3	73.3	N/A	N/A	N/A
			Val (AHI = 20)	90.0	84.2	N/A	N/A	N/A
			Val (AHI = 30)	83.3	91.7	N/A	N/A	N/A
PPG								
ODI	Chiner *et al.* [141]	216 OSA, 59 non-OSA	Val (ODI ≤ 5)	80	89	97	48	81
			Val (ODI≤ 10)	71	93	97	42	75
			Val (ODI≤ 15)	63	96	99	38	70
ODI	Vázquez *et al.* [143]	142 OSA, 99 non-OSA	Val (ODI = 15)	98	88	N/A	N/A	N/A
SpO_2 features	Al-Angari & Sahakian [144]	50 OSA, 50 non-OSA	Val	80.0	90.2	N/A	N/A	95
Time & frequency features	Álvarez *et al.* [145]	100 OSA, 48 non-OSA	Val	92	85.4	N/A	N/A	89.7
$\mathcal{H}_A$	Hornero *et al.* [146]	111 OSA, 76 non-OSA	Val	82.09	86.96	N/A	N/A	N/A
Poincaré plots	Morillo *et al.* [147]	54 OSA, 63 non-OSA	Val	90	84	N/A	N/A	N/A
Envelope fluctuations	Gil *et al.* [150]	8 OSA, 7 normal	Val	87.5	71.4	N/A	N/A	80

Chapter 4

Classic audio analysis techniques

4.1 Introduction

As a baseline approach, existing algorithms reported in the literature were implemented and tested on the data collected for this thesis. Most methods are derived from the field of speech analysis, and it is possible to use them because of the similarities between speech and snoring. Speech is produced when the vocal tract acts as an acoustic filter on air forced from the lungs, with the vocal tract changing shape depending which sound is to be produced. Snoring is caused by changes in the configuration and properties of the UA; the UA essentially acts as an acoustic filter in the same way that the vocal tract does in speech production. Abeyratne *et al.* [104] explained snoring in terms of a source/vibration model whereas speech follows a source/vocal-tract model. Many studies have shown that it is possible to use speech analysis techniques to diagnose snoring in terms of OSA or non-OSA (see section 3.2). This section focuses on classifying sounds as the first breath after an apnoea or noise/snore. Classifying the sounds in this way corresponds to what has been done in the literature, where benign/simple snores and apnoeic snores are the two classes used. By including noise in the benign snores class, the classification problem becomes more realistic, while allowing for an event detector to be built where the two classes are parametrised by the features described in this chapter.

Figure 4.1: *Grey Flash device for home use. Adapted from* `www.stowood.co.uk/page24.html`.

4.2 Data

The data used in this study were provided by collaborators at the Respiratory Medicine Group at the Churchill Hospital (Oxford, UK)[1]. Each subject used a portable home sleep study device, Grey Flash (Stowood Scientific Instruments Ltd., Oxford, UK.), shown in Figure 4.1 ($115 \times 80 \times 25mm$, $220g$ including batteries), which recorded a finger photoplethysmogram from which oxygen saturation and pulse rate were derived, nasal airflow and nasal sound from a nasal cannula, body movement and body position from an accelerometer and audio from a microphone placed on the nasal cannula. It should be noted that the subject was in charge of connecting themselves to the device, and that each device was calibrated to have approximately constant gain for the audio signal.

The data were saved to a memory card at a variety of sampling frequencies (Table 4.1). The subject returned the device to the hospital where the data were downloaded by the nursing staff using the Visi-Download software (Stowood Scientific Instruments Ltd., Oxford, UK) and then interpreted by trained medical staff.

The data were collected retrospectively through the sleep clinic at the Churchill Hospital. Each subject experienced normal clinical practice, without any additions/changes for this study. In a three year period, 1354 subject records were collected through the sleep clinic. The

[1]This study was approved by the NHS HRC National Research Ethics Service (NRES) South West REC Centre, Bristol, UK (REC reference SW/12/0211)

Table 4.1: *Channels recorded by the Grey Flash device.*

Parameter	Description	Sampling Frequency
Saturation	Blood oxygen saturation	$4Hz$
Pulse	Heart/pulse rate	$4Hz$
Pleth	Volume of the lungs	$32Hz$
Airflow	Pressure	$32Hz$
Body Movement	Correlated to Body Position	$4Hz$
Body Position	Supine, prone, left, right	$4Hz$
CPAP	Current level of CPAP	$4Hz$
Sound	Audio	$16Hz$
Battery	Battery power	$4Hz$

clinic does not focus solely on OSA but diagnoses a wide range of sleep disorders. As such, only those who were diagnosed as normal, snorers or mild/moderate/severe OSA were used in the analysis, *i.e.* only those diagnoses relevant to this study. All other diagnoses were excluded, bringing the total population from 1354 to 1014. Diagnoses that were excluded were: study failed/too short and needs repeating ($n = 204$), hypoventilation (obesity, central, REM, nocturnal) ($n = 55$), CSA ($n = 21$), COPD ($n = 15$), CPAP check up ($n = 14$), ventilatory failure ($n = 6$), poorly controlled asthma ($n = 4$), lung disease ($n = 3$), lung cancer ($n = 2$), narcolepsy ($n = 2$), hypoxia ($n = 2$), PLMS ($n = 2$), heart failure ($n = 2$), UARS ($n = 2$), idiopathic essential hypertension ($n = 1$), TB ($n = 1$), myotonic dystrophy ($n = 1$), polyethaemia ($n = 1$), catathrenia ($n = 1$), tachypnoea ($n = 1$).

Subjects that had duplicate recordings only had one included. The duplicates occurred either because there was a problem on the first night and nothing/only a few minutes were recorded, or the subject had a second recording taken because they were diagnosed with OSA initially and they were having a check-up (to see how well CPAP was controlling the condition or to see how much worse the condition had gotten and whether it was time to trial CPAP). In the first instance, the second recording was used, while in the second instance the first recording was used, dropping the population down further to 920 subjects. At this point, those subjects who had short recordings (less than $4.5h$) and those who were missing relevant signals were excluded. This brought the population down to 858 subjects, which was the final subject set used in this work. There was no selection bias in the choice of these 858 subjects; all records that met the criteria were included in the analysis. The subject demographics for each of the groups can be found in Table 4.2. Height is the only demographic that is not significantly different between the five sub-groups ($p < 0.001$). See Appendix D for more

Table 4.2: *Subject demographics for each sub-group: normal, snorer, mild OSA, moderate OSA and severe OSA (mean±σ). neck = neck circumference, m = male, f = female.*

Group	Normal	Snorer	Mild	Moderate	Severe
Gender	80 m, 75 f	166 m, 91 f	79 m, 28 f	94 m, 30 f	167 m, 48 f
Age (yrs)	45.9 ± 17.1	46.5 ± 12.0	50.5 ± 11.4	53.1 ± 12.4	52.5 ± 12.6
Neck (cm)	39.4 ± 4.6	41.4 ± 4.3	41.9 ± 4.1	42.9 ± 3.8	45.0 ± 4.8
Height (cm)	171.2 ± 10.7	173.5 ± 10.4	174.2 ± 9.9	173.0 ± 9.7	175.0 ± 9.1
Weight (kg)	77.7 ± 23.0	96.0 ± 24.2	212.0 ± 48.8	221.2 ± 49.5	247.3 ± 74.4
AHI ($events/h$)	4.4 ± 7.5	6.4 ± 7.4	10.6 ± 9.0	21.5 ± 11.6	47.5 ± 24.5
ODI ($events/h$)	3.7 ± 3.5	6.0 ± 5.2	10.3 ± 7.0	22.0 ± 11.6	56.8 ± 32.4
BMI (kg/m^2)	29.6 ± 7.9	32.0 ± 8.4	31.9 ± 7.9	33.8 ± 8.5	36.9 ± 11.2
ESS	11.0 ± 5.6	12.0 ± 5.2	12.2 ± 4.7	12.7 ± 4.7	14.1 ± 5.3

details.

The AHI values in Table 4.2, and throughout the rest of the thesis, were generated automatically by the Visi-Download software. The exact method of calculation is proprietary but it is based on airflow from the nasal cannula and oxygen saturation from the pulse oximeter. Airflow is used to detect apnoeas and hypopnoeas, and cross-checked with SpO_2 dips to ensure that the events are real. Body position is then used to determine the percentage of time that apnoeas/hypopnoeas occur in different positions (prone, supine, right, left, sitting up) [153]. As it is not possible to record airflow on a phone overnight, the AHI is not used as a feature in the remainder of this thesis. The ODI value is also generated by the software, however it is re-derived in Chapter 7.

4.2.1 Annotation and Segmentation

Twenty-two subjects had specific events identified and labelled using the Visi-Download software; their demographics can be found in Table 4.3 and a detailed breakdown of each individual's demographics and number of events can be found in Table 4.4. These subjects were chosen because they were the first subjects that were collected that met the selection criteria, *i.e.* were diagnosed as normal, snorer, mild/moderate/severe OSA and were longer than 4.5 h with all signals present. The labelling of events followed a protocol which involved dragging an event marker across the relevant section of data. The marker appeared on all data channels (Figure 4.2). Each event was labelled as in Table 4.5 and although five categories of sound were annotated, only three were deemed relevant for this section of work: first breath after apnoea (or choke), snoring, and noise events. This process resulted in ASCII 'workpad' (.WP)

Table 4.3: *Demographics of annotated subjects (mean±σ), m=male, f=female.*

Parameter	**Subjects** (mean ± σ)
Gender	m: 17, f: 5
Age (yrs)	48.9 ± 15.3
Neck (cm)	45.7 ± 3.8
Height (cm)	177.3 ± 10.7
Weight (kg)	107.4 ± 24.4
BMI (kg/m^2)	34.3 ± 8.9
ESS	11.7 ± 5.3

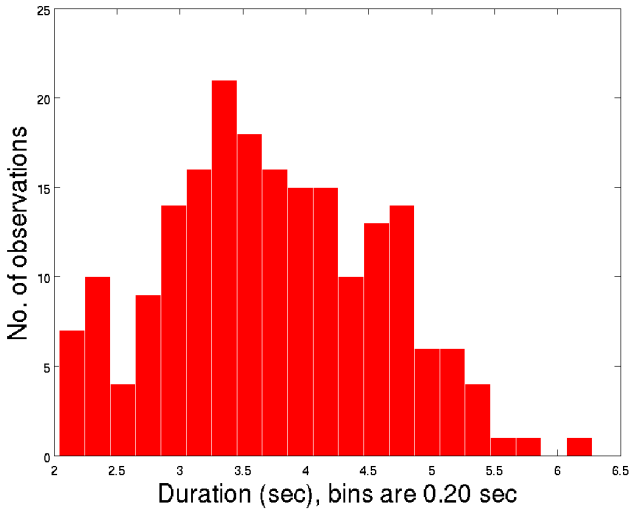
Table 4.4: *Detailed demographics of the annotated subjects, m=male, f=female, # n = number of noise events, # s = number of snoring events, # c = number of choke events.*

Subject	**Gender**	**Age** *yrs*	**Neck** *cm*	**Height** *cm*	**Weight** *kg*	**AHI**	**ODI**	**BMI** kg/m^2	**ESS**	**# n**	**# s**	**# c**
1	f	45.0	45.0	154.9	132.9	4.5	8.1	55.4	-	11	15	0
2	m	49.0	39.4	193.0	97.5	0.2	1.3	26.2	19.0	10	10	0
3	m	46.0	48.3	177.8	132.0	84.2	109.8	41.8	15.0	10	0	15
4	m	74.0	49.5	177.8	107.0	1.3	1.4	33.8	10.0	10	10	0
5	f	56.0	-	-	-	6.8	11.8	-	6.0	11	10	0
6	f	62.0	43.2	167.6	106.6	86.3	85.3	37.9	21.0	10	10	15
7	m	49.0	41.9	172.7	104.8	5.9	7.5	35.1	8.0	10	13	0
8	m	21.0	43.2	-	-	2.6	6.4	-	3.0	6	10	2
9	m	70.0	55.9	-	108.0	34.3	28.9	-	15.0	10	10	14
10	m	53.0	45.7	182.9	112.0	42.6	36.0	33.5	-	11	10	15
11	m	50.0	43.2	170.2	93.4	25.3	19.4	32.3	13.0	10	7	11
12	f	34.0	-	180.3	59.4	3.5	5.4	18.3	7.0	10	10	0
13	m	38.0	47.0	198.1	139.7	8.9	7.4	35.6	7.0	9	10	4
14	m	38.0	49.5	185.4	152.4	69.0	97.9	44.3	16.0	10	10	18
15	m	33.0	43.4	167.6	69.9	39.9	26.4	24.8	18.0	10	10	10
16	m	74.0	43.2	-	96.6	39.9	42.0	-	9.0	7	4	7
17	m	34.0	47.0	182.9	88.0	6.2	1.1	26.3	14.0	7	10	2
18	m	35.0	44.5	-	-	2.2	5.9	-	12.0	4	12	1
19	m	31.0	49.5	182.9	138.8	85.6	91.1	41.5	10.0	4	9	22
20	m	69.0	43.2	172.7	88.9	83.2	69.5	29.8	16.0	4	5	21
21	f	50.0	41.9	167.6	88.9	29.3	36.4	31.6	3.0	9	10	2
22	m	66.0	50.8	-	123.8	50.1	57.0	-	-	7	6	16

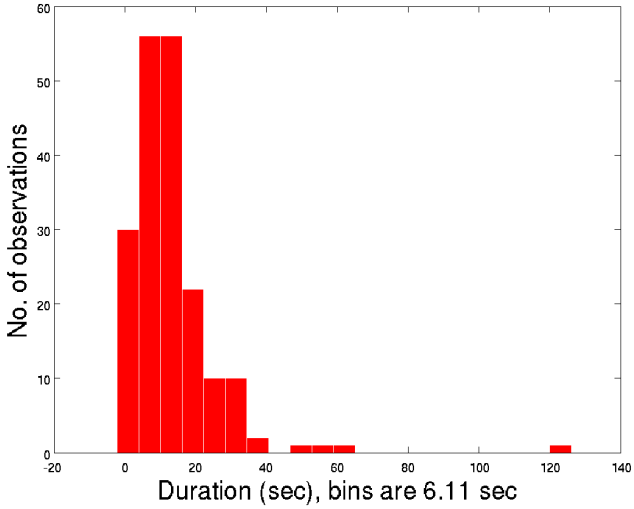
files being produced, where a start time, duration and label associated with each event were saved. The protocol used can be found in Appendix E. The annotations were carried out by a clinical research fellow, with two years of experience in sleep medicine. Only 22 records were annotated due to the time constraints, *i.e.* to annotate a single record took at least one hour. In addition, the annotator moved away and, although multiple attempts were made to find other annotators, none were found.

A total of 175 choke events, 190 noise events and 201 snoring events were annotated. A histogram of events' duration can be found in Figure 4.3.

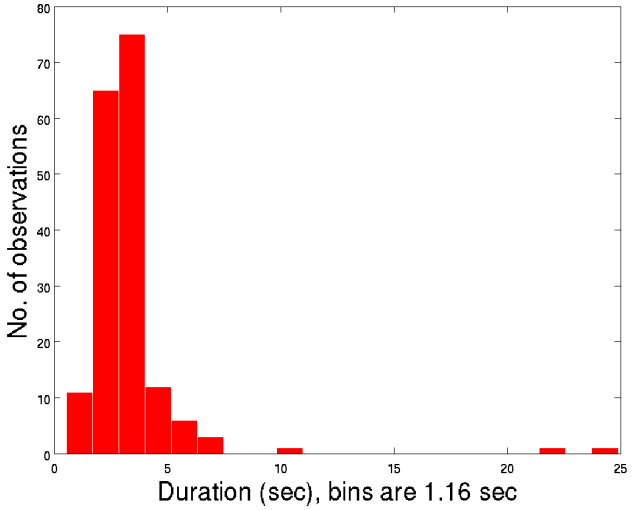
Figure 4.2: *Annotating the data using Visi-Download (Stowood Scientific Instruments Ltd., Oxford, UK) - a marker appears on all channels for the event in question. X-axis is time of day.*



(a) *Snore histogram*



(b) *Noise histogram*



(c) *Choke histogram*

Figure 4.3: *Histogram of events' duration.*

Table 4.5: *Labelling Protocol, A = apnoea, F = first breath following an apnoea, S = snoring, N = noise, B = breathing, U = uncertain, C = crescendo, S = simple, V = voice, T = t.v., R = radio, L = light, H = heavy.*

Event	**Label** (first letter)	Optional sub-label
Apnoea	A	U
First breath after apnoea	F	U
Snoring	S	U, C, S
Noise	N	U, V, T, R
Breathing	B	U, L, H

4.3 Methods

The following methods were used in the analysis of the events from the 22 subjects. Linear predictive coding coefficients and mel-frequency cepstral coefficients were used as features, while a support vector machine and linear discriminant analysis were used to classify the features as associated with OSA (choke) or not (noise or snore).

4.3.1 Linear Predictive Coding

Linear predictive coding (LPC) is a commonly used speech analysis technique as it provides an accurate representation of speech, particularly at low bit rates [154]. LPC has been used in speech analysis since the late 1960s. In 1966, Saito & Itakura [155] developed a method to discriminate between different phonemes (the elements of speech) which made use of maximum-likelihood. A proposal was made to encode speech in real-time by Atal & Hanauer [81].

Linear prediction is based on the source-filter model of speech. The sampled speech signal can be modelled as the output of a linear, slowly time-varying system excited by either quasi-periodic impulses (voiced speech), or random noise (unvoiced speech) [154]. The excitation of the vocal tract gives rise to the different sounds and so the shape of the vocal tract is key.

The simplest model of the vocal tract is to approximate it as a cylinder; a more complicated view sees the vocal tract as the concatenation of lossless tubes of uniform length. The shorter the length of the tubes, the closer the approximation is to the actual shape of the vocal tract. As the shape of the vocal tract changes throughout time so too will the diameters of the tubes.

At each of the interfaces to this model, the input signal is delayed and attenuated due to partial reflection of the sound wave. Therefore, the output may be approximated as the linear combination of the previous reflected outputs. Linear prediction can estimate the coefficients

of this sum, using filter theory. The bilateral z-transform $X(z)$ of a discrete-time signal $x(n)$ is given by:

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$$

where z is a complex variable and n is the sample number. Using this definition, the *transfer function* of a linear filter in the z domain is defined as:

$$H(z) = Y(z)/X(z)$$

where $Y(z)$ is the z-transform of the filter output signal $y(n)$, and $X(z)$ is the z-transform of the filter input signal $x(n)$.

For the vocal tract, the filter is time-varying, causing the transfer function to change at each discrete time-step. However, for a linear system the transfer function may be represented by its poles and zeros, because, for a linear filter the transfer function can be rewritten as $H(z) = a(z)/b(z)$, where a and b are both polynomials. For voiced sounds, the transfer function has no zeros; for unvoiced sounds, the zeros all lie inside the unit circle and so the zeros can be approximated by multiple poles. Therefore the vocal tract can be represented by an *all-pole filter* with a transfer function of the form:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{4.1}$$

The a_k's and G are constants, while p corresponds to the number of tubes used to approximate the vocal tract.

Linear prediction allows the values of some discrete-time signal to be written in terms of a linear combination of previous values of the signal [154]. Given a signal s with n^{th} element $s(n)$, a p^{th} order linear predictor is a system whose output is:

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k). \tag{4.2}$$

The coefficients a_k are chosen so that the prediction 4.2 is close to the actual value of the signal.

The vocal tract is being modelled as an all-pole linear filter with transfer function H given by 4.1. For p approximating tubes, the speech samples $s(n)$ are related to the normalised excitation $u(n)$ by the equation:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n)$$

where $u(n)$ is the n^{th} sample of the normalised excitation signal, and G is the *gain* of the signal. The prediction error, the amount by which $\tilde{s}(n)$ fails to exactly predict sample $s(n)$, is:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k). \tag{4.3}$$

From the definitions 4.2 and 4.3, the prediction error $e(n)$ is equal to the scaled normalised excitation $Gu(n)$. The coefficients of the linear predictor are chosen so that they minimise the mean-squared prediction error:

$$E_n = \sum_{m} e_n^2(m) \tag{4.4}$$

where $e_n(m)$ is shorthand for $e(n+m)$ and m is a neighbourhood around the analysis sample n for $\{-M_1 \leq m \leq M_2\}$. Using 4.3, rewrite 4.4 as:

$$E_n = \sum_{m=-M_1}^{M_2} \left(s_n(m) - \sum_{k=1}^{p} a_k s(m-k) \right)^2 .$$

Find the values of a_k that minimises E_n by setting each of the partial derivatives with respect to each a_k to zero. Using the chain rule:

$$\frac{\delta E_n}{\delta a_i} = -\sum_{m=-M_1}^{M_2} 2 \left(s_n(m) - \sum_{k=1}^{p} a_k s(m-k) \right) s_n(m-i), \quad \{1 \leq i \leq p\}. \tag{4.5}$$

Setting 4.5 equal to zero gives a formula for the optimal predictor coefficients $\hat{a}_k$:

$$\sum_{m=-M_1}^{M_2} s_n(m-i) s_n(m) = \sum_{k=1}^{p} \hat{a}_k \sum_{m=-M_1}^{M_2} s_n(m-i) s_n(m-k), \quad \{i = 1, \ldots, p\}. \tag{4.6}$$

This is a system of p equations, solvable for p unknowns. Defining the *short-term covariance*

of $s_n(m)$ as:

$$\varphi_n(i,k) = \sum_{m=-M_1}^{M_2} s_n(m-i)s_n(m-k),$$

then 4.6 can be rewritten as:

$$\varphi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \varphi_n(i,k). \tag{4.7}$$

Given a speech sample $s(n)$, calculate each of the quantities $\varphi_n(i,k)$ for $\{1 \leq i \leq p\}$ and $\{0 \leq k \leq p\}$. The method of choosing the range of m plays a key role in these calculations and in the solving of the equations; two possible methods are the covariance method and the autocorrelation method. For the work in this report, the latter method was used as it is less computationally intensive.

4.3.1.1 The Autocorrelation Method

The autocorrelation method enforces the condition that the segment $s_n(m)$ is equal to zero outside some interval $\{0 \leq m \leq N-1\}$. The condition can be enforced by multiplying the speech signal by some window function, $w(m)$, which is zero outside this interval:

$$s_n(m) = \begin{cases} s(m+n)w(m), & 0 \leq m \leq N-1 \\ 0, & \text{otherwise} \end{cases}$$

Given this weighted signal, the mean-squared error can be formally defined as: $E_n = \sum_{m=0}^{N-1+p} e_n^2(m)$. With the range of m thus defined, the short-term covariance can be rewritten as:

$$\varphi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k), \quad \{1 \leq i \leq p\}, \quad \{0 \leq k \leq p\}.$$

This is now a function of the single variable $i-k$, which is referred to as the short-term autocorrelation, $r_n(i-k)$. Given this new notation, the LPC equations 4.7 can be written in a form known as the *Yule-Walker equations* [154]:

$$\sum_{k=1}^{p} r_n(|i-k|)\hat{a}_k = r_n(i), \quad \{1 \leq i \leq p\}.$$

These equations can be written in matrix form as follows:

$$\begin{pmatrix} r_n(0) & r_n(1) & r_n(2) & \cdots & r_n(p-1) \\ r_n(1) & r_n(0) & r_n(1) & \cdots & r_n(p-2) \\ r_n(2) & r_n(1) & r_n(0) & \cdots & r_n(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(p-1) & r_n(p-2) & r_n(p-3) & \cdots & r_n(0) \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{pmatrix} = \begin{pmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(p) \end{pmatrix}$$

This is a Toeplitz matrix, which means that it is symmetric and all diagonal entries are equal. The column vector on the right-hand side contains entries which appear in the matrix. Therefore, the Durbin algorithm can be used to speed up the computational time for estimating the mean squared error as follows:

$$\begin{aligned} E^{(0)} &= r(0) \\ k_i &= \{r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|)\} / E^{(i-1)} \\ \alpha_i^{(i)} &= k_i \\ \alpha_j^{(i)} &= \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \\ E^{(i)} &= (1 - k_i^2) E^{(i-1)} \end{aligned}$$

This process is iterated for values of i in the range $\{1 \leq i \leq p\}$. The LPC coefficients are then given by $a_m = \alpha_m^{(p)}$.

4.3.2 Cepstral Analysis

In 1963 Bogert *et al.* [156] defined the *cepstrum* as the inverse Fourier transform of the log magnitude spectrum of a signal. Cepstral analysis makes use of the source-filter model of speech production. It is widely used in speech processing, particularly for pitch estimation [154].

For the source-filter model of speech production the output signal s is expressed in terms of a convolution of a rapidly-varying excitation e and a slowly-varying filter h:

$$s(n) = e(n) * h(n). \tag{4.8}$$

For the LPC analysis, the z-transform of this equation was computed and the transfer function of the filter was assumed to be of a certain type, corresponding to an all-pole model of the vocal tract. For the cepstral analysis the Fourier transform is computed instead. The discrete-time Fourier transform (DTFT) of a signal $x(n)$ of finite length L is:

$$X(e^{i\omega}) = \sum_{n=0}^{L-1} x(n)e^{-i\omega n}.$$

The discrete Fourier transform (DFT) provides a discrete sampling of this transform and is given by:

$$X(k) = \sum_{n=0}^{L-1} x(n)e^{-\frac{2\pi i}{L}kn}, \qquad \{k = 0, 1, \ldots, L-1\}.$$

As with the z-transform, the Fourier transform of a convolution of two signals is equal to the product of the Fourier transforms of the individual signals. Therefore, taking the DTFT of 4.8 gives:

$$S(e^{i\omega}) = E(e^{i\omega})H(e^{i\omega}) \tag{4.9}$$

in the frequency domain. The aim is to deconvolve the signal so as to obtain representations of the transfer function and excitation. One method is to take the absolute value of 4.9 and then take the logarithm, giving:

$$log|S(e^{i\omega})| = log|E(e^{i\omega})| + log|H(e^{i\omega})|.$$

The *complex cepstrum* $\hat{s}$ of a sequence $s(n)$ of finite length is defined as the Fourier transform of its log spectrum:

$$\hat{s}(n) = \int_{-\pi}^{\pi} log|S(e^{i\omega})|e^{-i\omega n}d\omega.$$

This is the definition given by Bogert *et al.* [156]. In other texts, the cepstrum is defined as the inverse Fourier transform of the log magnitude spectrum of a signal, $S(\omega)$.

Since the index n represents the time at which a sample is taken, it is clear that the unit of the cepstrum is time not frequency and is referred to as the *quefrency*. When the cepstrum of a signal is calculated, the Fourier analysis is carried out on the log spectrum itself. Therefore the cepstrum looks for periodicity in the spectrum, providing information regarding the rate of change in different frequency bands.

4.3.2.1 Mel-Frequency Cepstral Coefficients

Weighted cepstrum distance measures have a direct equivalent interpretation in terms of distance in the frequency domain. This is important in models for human perception of sound which are based on frequency analysis carried out in the inner ear [154]. Davis & Mermelstein [97] used this fact as the basis of the mel-frequency cepstrum coefficients (MFCCs). The idea behind MFCCs is to compute a frequency analysis based on a filter bank with approximately critical band spacing of the filters and bandwidths. The frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response better than the normal cepstrum. Generally, a short-time Fourier analysis is carried out first resulting in a DFT $X_{\hat{n}}[k]$ for time $\hat{n}$. These DFT values are then grouped together in critical bands and weighted by a triangular weighting function. The mel-frequency spectrum at time $\hat{n}$ for $r = 1, 2, \ldots, R$ is

$$MF_{\hat{n}}[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r[k] X_{\hat{n}}[k]|^2$$

where $V_r[k]$ is the triangular weighting function for the r^{th} filter ranging from DFT index L_r to U_r, where

$$A_r = \sum_{k=L_r}^{U_r} |V_r[k]|^2$$

is a normalising factor for the r^{th} mel-filter. This normalisation is built into the weighting functions so that a perfectly flat input Fourier spectrum will produce a flat mel-spectrum. For each frame, a discrete cosine transform of the log of the magnitude of the filter outputs is computed to form the function $mfcc_{\hat{n}}[m]$:

$$mfcc_{\hat{n}}[m] = \frac{1}{R} \sum_{r=1}^{R} log(MF_{\hat{N}}[r]) cos[\frac{2\pi}{R}(r + \frac{1}{2})m].$$

Usually, $mfcc_{\hat{n}}[m]$ is evaluated for a number of coefficients, N_{mfcc}, that is less than the number of mel-filters.

4.3.3 Classification

Two classifiers were used to classify the data: a simple linear classifier and a support vector machine (SVM). These methods were chosen as standard linear and non-linear benchmark

classifiers.

4.3.3.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a well-known scheme that draws a linear boundary between the values of the feature set, and has been used in a variety of applications such as image retrieval and face recognition. Classical LDA projects the data onto a lower-dimensional vector space such that the ratio of the between-class distance to the within-class distance is maximised, thus achieving maximum discrimination. The optimal projection can be computed by applying the eigen-decomposition on the covariance matrices [157].

If there are g groups, Bayes' Rule minimises the total error of classification by assigning the object to group i which has the highest condition probability where $P(i|x) > P(j|x), \forall j \neq i$. Since it is not possible to get $P(i|x)$ we obtain $P(x|i)$ and then use Bayes' Theorem:

$$P(i|x) = \frac{P(x|i).P(i)}{\sum_{\forall j} P(x|j).P(j)}.$$

Therefore, Bayes' Rule becomes: assign the object to group i if

$$\frac{P(x|i).P(i)}{\sum_{\forall k} P(x|k).P(k)} > \frac{P(x|j).P(j)}{\sum_{\forall k} P(x|k).P(k)}, \forall j \neq i.$$

The denominators are positive and the same, therefore they can be cancelled out to become:

$$P(x|i).P(i) > P(x|j).P(j), \forall j \neq i.$$

Assume the data comes from a multivariate normal distribution whose formula is:

$$P(x|i) = (\frac{1}{(2\pi)^{\frac{\pi}{2}}|C_i|^{\frac{1}{2}}})exp(-\frac{1}{2}(x-\mu_i)^T C_i^{-1}(x-\mu_i))$$

where μ_i is the vector mean and C_i is the covariance matrix of group i. Inputting the distribution formula into Bayes' Rule, cancelling factors of $(2\pi)^{\frac{\pi}{2}}$, taking the logarithm of both sides

and multiplying by -2 gives:

$$ln(|C_i|) - 2ln(P(i)) - (x-\mu_i)^T C_i^{-1}(x-\mu_i) < ln(|C_j|) - 2ln(P(j)) - (x-\mu_j)^T C_j^{-1}(x-\mu_j), \forall i \neq j.$$

Assuming all covariance matrices are equal, $C = C_i = C_j$, allows for further simplification. Writing $(x-\mu_i)^T C_i^{-1}(x-\mu_i)$ as $xC^{-1}x^T - 2\mu_i C^{-1}x^T + \mu_i C^{-1}\mu_i^T$ means that terms can be cancelled, and multiplying both sides of the inequality by $-\frac{1}{2}$ gives rise to:

$$f_i = \frac{1}{2}\mu_i C^{-1} x_k^T - \frac{1}{2}\mu_i C^{-1} x_i^T + ln(P(i)).$$

Therefore, assign object with measurement x to group i if:

$$f_i > f_j, \forall i \neq j.$$

4.3.3.2 Support Vector Machines

Support vector machines (SVMs) are a machine learning technique, originally introduced by Vapnik and co-workers in the 1990s [158]. When used for classification they separate a given set of binary labelled training data with a hyper-plane that is maximally distant from them. If no linear separation is possible, SVMs work in combination with the technique of *kernels* and automatically realise a non-linear mapping to a feature space. The hyper-plane found by the SVM in the feature space corresponds to a non-linear decision boundary in the input space [159].

The soft-margin SVM algorithm is one way to train a linear separation boundary. Ideally, if the data is linearly separable, the linear boundary which has maximal distance to the nearest training points is determined. Mathematically, calculate $\boldsymbol{w}$, a weighting vector of coefficients, and b, a bias term, such that the following two equations hold:

$$\boldsymbol{w} \cdot \boldsymbol{x_i} - b \geq 1\ , \quad y_i = 1$$

$$\boldsymbol{w} \cdot \boldsymbol{x_i} - b \leq -1\ , \quad y_i = -1$$

These two equations are combined to form a simplified representation of the SVM's ideal

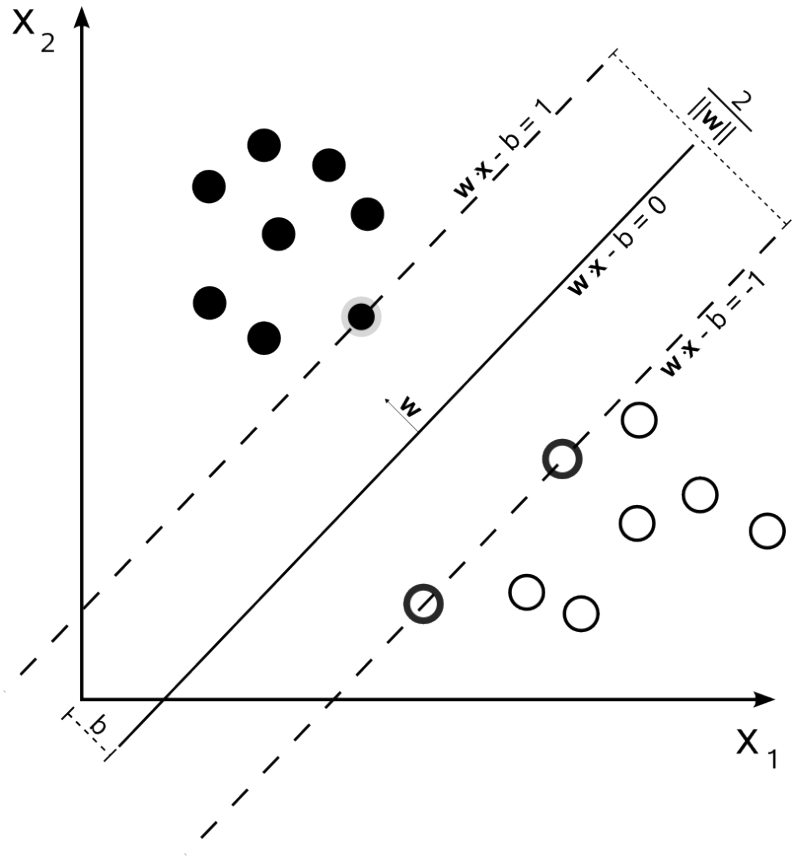


Figure 4.4: *Demonstration of a linear separation boundary and the maximum margin associated with it. Black circles are the positive class, $y = 1$, white circles are the negative class, $y = -1$. Circles on the dashed lines with an outline or thicker border are the support vectors. Taken from* `http://en.wikipedia.org/wiki/File:SVM_max_sep_hyperplane_with_margin.png`

output:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} - b) \geq 1 \tag{4.10}$$

where $\boldsymbol{w}$ is defined as:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

As stated earlier, it is desirable to maximise the separation boundary, or minimise $\|w\|$ subject to the specified constraint 4.10. This is shown graphically in Figure 4.4.

Since solving for the minimum of $\|w\|$ is computationally intensive, it is replaced by solving for the minimum of $\frac{1}{2}\|w\|^2$. This is mathematically equivalent, as both representations have the same minimum w. This is now a quadratic optimisation problem subject to the constraint shown in 4.10, which can be solved using the method of Lagrange multipliers. In the optimisation of SVM boundaries the only data dependency is on dot products between the support vectors, which is highlighted below. If $\boldsymbol{\alpha}$ represents a vector of coefficients, then the

Lagrange function can be defined as:

$$L(\boldsymbol{w}, b, \alpha_i) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_i \alpha_i \left[y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + b \right) - 1 \right] \tag{4.11}$$

where only $\alpha_i \geq 0$ represent the indetermined coefficients. If $\boldsymbol{w}$ and b take the optimal value, then the partial derivative of this equation can be set equal to 0 and solved. This results in:

$$\begin{aligned} \boldsymbol{w} &= \sum_i \alpha_i y_i \boldsymbol{x}_i \\ \sum_i \alpha_i y_i &= 0 \end{aligned} \tag{4.12}$$

Note that α_i is non-zero only for data points satisfying $y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} - b) - 1 = 1$, or more intuitively, only the data points on the maximum margin boundary are present in the solution $\boldsymbol{w}$. These are the support vectors referred to in Figure 4.4.

Substituting 4.12 into 4.11 and solving gives:

$$L(\boldsymbol{w}, b, \alpha_i) = -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_i \alpha_i. \tag{4.13}$$

In 4.13, it is desirable to maximise L, which is a quadratic programming problem. Formally, it involves maximising 4.14 subject to the constraint in 4.15:

$$\max(-\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_i \alpha_i) \tag{4.14}$$

$$\sum_i \alpha_i y_i = 0, \quad \alpha_i \geq 0 \tag{4.15}$$

In 4.14, it can be seen that the maximisation is only dependent on the inner product of the data vectors. This is what allows for the extension of SVMs to incorporate non-linear boundaries.

This is the general method used by SVMs to train a linear separation boundary if the data is completely linearly separable. All types of SVMs are solved through the use of Lagrange multipliers and quadratic programming solutions. In order to develop non-linear separation boundaries, the data are transformed into a higher dimensional space through a mapping function, $\phi(x)$. A linear separation boundary is then trained in this higher dimensional space using the methods presented earlier. The input is then mapped back to the original feature space,

resulting in a non-linear separation boundary.

Recall that in 4.14, the only input data dependency was on $\sum_i \sum_j x_i^T x_j$, or more compactly the inner product of the two vectors, $\boldsymbol{x_i} \cdot \boldsymbol{x_j}$. When transforming the data into a higher dimensional feature space, this inner product becomes $\boldsymbol{\phi(x_i)} \cdot \boldsymbol{\phi(x_i)}$. Though this is a computationally costly operation, it can be performed in the original feature space by use of a certain type of function. These functions, which represent a dot product of vectors in the high dimensional space using the vectors in the input space, are known as *kernels*. The *Kernel Trick* involves using kernels to transform computationally expensive high dimensional inner product calculations to easier operations.

A kernel is defined as a symmetric positive definite function K such that, for all $x, x' \in X$:

$$K(x, x') =< \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{x'}) > .$$

Similarly, the Kernel matrix (or Gram matrix) is defined as:

$$K = \begin{vmatrix} \phi(x_1)\phi(x_1') & \cdots & \phi(x_1)\phi(x_N') \\ \vdots & & \vdots \\ \phi(x_N)\phi(x_1') & \cdots & \phi(x_N)\phi(x_N') \end{vmatrix}$$

A wide variety of kernels that can be used by SVMs exist and the one that is used is dictated by the problem analysed. The most common kernels are:

$$K(x, x') = \boldsymbol{x} \cdot \boldsymbol{x'} \tag{4.16}$$

$$K(x, x') = (\gamma \boldsymbol{x} \cdot \boldsymbol{x'} + C_0)^d \tag{4.17}$$

$$K(x, x') = e^{-\gamma \|\boldsymbol{x} - \boldsymbol{x'}\|^2} \tag{4.18}$$

$$K(x, x') = tanh(\gamma \boldsymbol{x} \cdot \boldsymbol{x'} + C_0) \tag{4.19}$$

4.16 is a linear kernel, 4.17 is a polynomial kernel, 4.18 is the radial basis function kernel (or Gaussian kernel), and 4.19 is the sigmoid kernel.

Table 4.6: *The number of each event type at the four different window sizes used.*

Event	Window			
	$0.5s$	$1s$	$2s$	$3s$
F	175	175	155	82
S	201	201	201	159
N	190	189	185	167

4.4 Data Analysis Protocol

Based on the assumption that either an event detector would be used to find the sections of interest, or that the entire night would be analysed on an approximately second-to-second basis, only a specified amount of time for each event was analysed *i.e.* the first 0.5, 1, 2 or $3s$ of an event. If an event duration was less than the specified window size, it was not included in the analysis. This meant that as the window size increased, less data was analysed; Table 4.6 shows the number of each event type at the different window sizes.

4.4.1 Linear Predictive Coding

From speech analysis, a general rule of thumb is that for voiced sounds, two coefficients provide information about each formant frequency. It has been suggested that voiced sounds are identifiable from the first two or three formants [154]. Using a filter order of 12 ensures that the first three formants can be estimated, which is useful in identifying sections of speech in the audio signal. The work of Ng *et al.* [83] looked at the first three formants also and achieved promising results in distinguishing between apnoeic and non-apnoeic snoring. However, using a prediction order of 12 resulted in pole-splitting (which could be seen on the pole-zero plot). Instead a prediction order of 8 was used, giving four complex conjugate pairs, allowing the spectrum to be fully shaped. Figure 4.5 shows three examples of the pole-zero plots, using $p = 8$ resulting in all-pole plots (or three pairs of poles and one pair of zeros), and autoregressive (AR) spectra for the different events. The pole-zero plots show that none of the three events have poles in the same location, particularly the noise event whose poles have a much smaller radius. This results in different system responses for the three events. The AR spectra show how the power is distributed across the frequency range. Figures 4.5b, 4.5d and 4.5f clearly show that the choke event has a very different PSD to the other two events, with one very clear peak. As differences can be seen between the pole-zero plots and the AR spectra using this

prediction order, $p = 8$ was therefore used for the LPC analysis.

4.4.2 MFCC Analysis

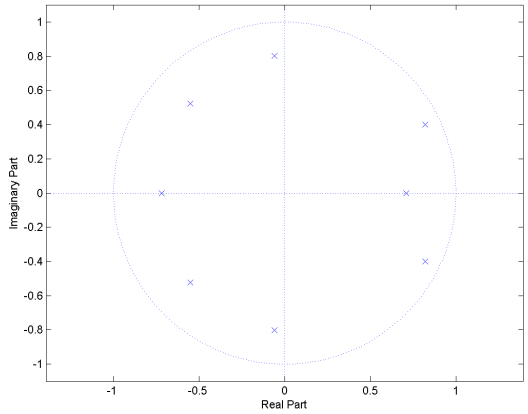
Each audio window of interest was detrended and then multiplied by a Hamming window of the same length. A filterbank with 24 filters was used and the entire length of each event was taken to be a single frame, resulting in 12 MFCCs per event.

4.4.3 Classification

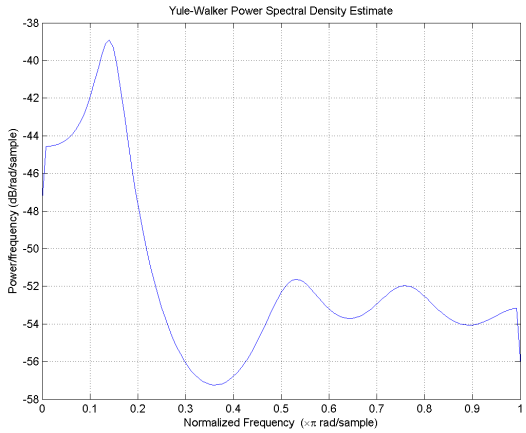
For both classifiers, five-fold cross-validation was carried out. Four folds were used as X_{design} and the remaining fold was used as X_{val}. Different combinations of features (LPC, MFCCs, and LPC+MFCCs) were used to find the optimal threshold for classification. For the SVM, the X_{design} data was used to find the best cost parameter for that data set. This was done by carrying out three-fold cross-validation on X_{design}, where one fold was used as X_{test} and the other two folds were used as X_{train}. The best cost parameter was found on X_{train} then used to to train the SVM on all of X_{design} before being validated on X_{val}. This process was carried out for every window size. Two approaches were used: the first involved events being divided into training and validation, the other divided subjects into training and validation. Dividing events into training and validation meant that events from a single subject appeared in both data sets which could lead to a bias in results. Dividing subjects into training and validation removed this bias, but led to uneven numbers of events in the two groups. For LDA, the discriminant function used fitted a multivariate normal density to each group, with pooled estimates of a diagonal covariance matrix (essentially a naive Bayes classifier). In the SVM analysis a linear kernel was used, as was a weighting function. Every time the data was divided into X_{design} and X_{val}, the ratio of chokes to snores/noise was calculated and the weighting function changed accordingly before being applied to the training data set.

4.5 Results

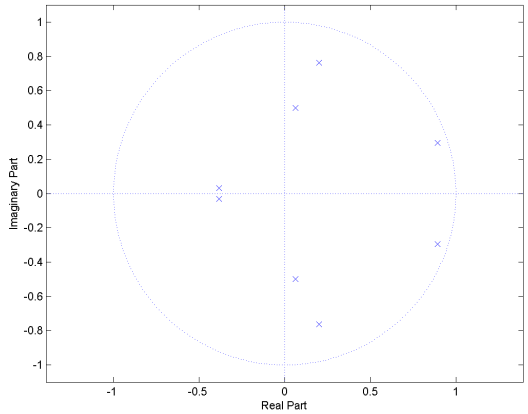
The results for LDA can be found in Tables 4.7 and 4.8 while the results for the SVM can be found in Tables 4.9 and 4.10. The best performance for each classification problem is high-



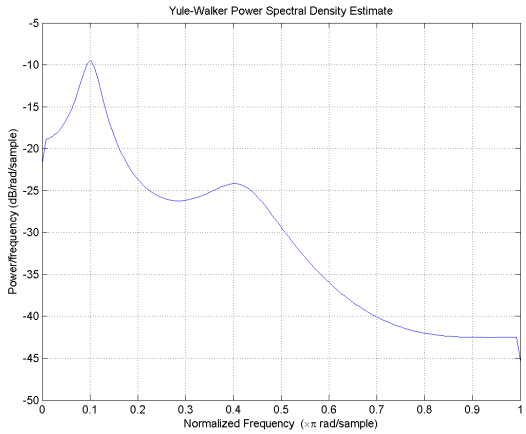
(a) *Pole-zero plot of a choke event*



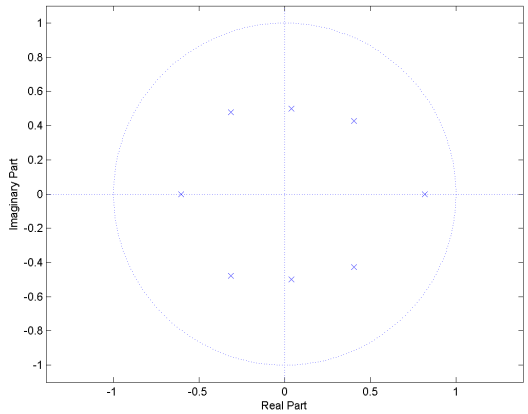
(b) *AR spectrum of a choke event*



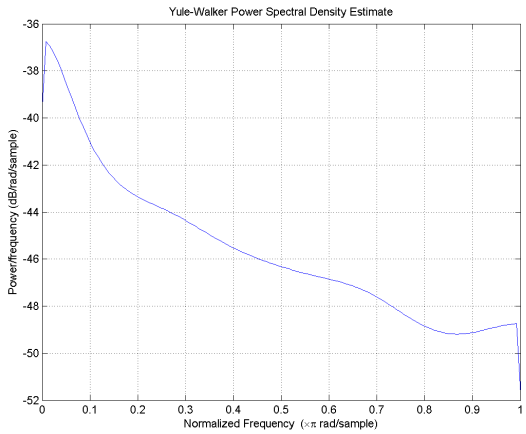
(c) *Pole-zero plot of a snoring event*



(d) *AR spectrum of a snoring event*



(e) *Pole-zero plot of a noise event*



(f) *AR spectrum of a noise event*

Figure 4.5: *Pole-zero plots and autoregressive spectra of a choke, snoring, and noise event. Differences can be seen clearly between the three event types.*

Table 4.7: *LDA statistics for LPC coefficients (LPC), MFCCs (CEP) and both combined (L&C) (mean±σ) for events divided into Training (Train) and Validation (Val).*

Inputs	τ (s)	Data	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
LPC	0.5	Train	70.0 ± 1.4	50.6 ± 2.3	38.8 ± 2.4	79.0 ± 0.7	56.6 ± 2.0	0.57 ± 0.02
		Val	68.0 ± 9.4	47.9 ± 8.6	37.0 ± 5.7	77.3 ± 4.4	54.4 ± 4.9	0.54 ± 0.06
CEP		Train	59.9 ± 1.8	59.8 ± 2.9	40.0 ± 2.1	76.9 ± 1.9	59.8 ± 2.3	0.56 ± 0.02
		Val	51.2 ± 11.8	58.8 ± 6.6	36.0 ± 10.5	73.1 ± 6.0	56.7 ± 6.3	0.50 ± 0.09
L&C		Train	70.0 ± 2.5	60.2 ± 2.0	44.0 ± 1.6	81.8 ± 1.2	63.2 ± 1.5	0.62 ± 0.02
		Val	65.2 ± 7.0	55.5 ± 4.6	39.6 ± 3.4	78.2 ± 3.2	58.5 ± 3.0	0.57 ± 0.04
LPC	1	Train	65.9 ± 3.3	48.5 ± 1.4	36.5 ± 2.3	76.1 ± 1.7	53.9 ± 1.8	0.53 ± 0.02
		Val	63.3 ± 11.3	48.0 ± 7.0	35.5 ± 7.7	74.9 ± 5.2	52.9 ± 6.2	0.52 ± 0.07
CEP		Train	54.7 ± 3.7	60.8 ± 2.2	38.5 ± 2.8	74.9 ± 1.6	58.9 ± 2.1	0.54 ± 0.03
		Val	51.9 ± 6.8	62.0 ± 5.1	38.0 ± 4.5	74.2 ± 4.4	58.9 ± 2.8	0.52 ± 0.06
L&C		Train	63.2 ± 7.3	59.8 ± 2.6	41.3 ± 1.9	78.5 ± 3.6	60.9 ± 2.2	0.58 ± 0.05
		Val	57.1 ± 9.4	54.9 ± 9.2	36.4 ± 5.7	73.8 ± 6.9	55.4 ± 6.2	0.52 ± 0.07
LPC	2	Train	58.8 ± 3.8	63.6 ± 3.9	39.4 ± 1.6	79.4 ± 1.0	62.2 ± 2.1	0.58 ± 0.01
		Val	59.2 ± 9.7	62.6 ± 9.3	39.4 ± 6.6	79.3 ± 4.1	61.4 ± 4.9	0.58 ± 0.04
CEP		Train	67.4 ± 1.7	67.5 ± 2.8	45.5 ± 2.3	83.8 ± 0.8	67.5 ± 1.8	0.65 ± 0.02
		Val	62.0 ± 10.4	68.1 ± 6.5	43.8 ± 7.6	81.7 ± 6.6	66.4 ± 4.7	0.63 ± 0.08
L&C		Train	63.9 ± 2.3	71.6 ± 1.7	47.5 ± 2.1	83.2 ± 1.2	69.4 ± 1.0	0.66 ± 0.01
		Val	60.3 ± 4.1	69.6 ± 7.4	44.6 ± 5.0	81.1 ± 5.2	67.3 ± 5.3	0.63 ± 0.04
LPC	3	Train	56.1 ± 5.5	72.5 ± 3.2	34.0 ± 2.9	86.8 ± 1.9	69.2 ± 2.8	0.66 ± 0.03
		Val	51.3 ± 17.9	71.4 ± 11.0	31.3 ± 8.7	85.5 ± 6.3	67.4 ± 7.5	0.63 ± 0.06
CEP		Train	66.7 ± 2.1	74.8 ± 2.3	40.1 ± 2.5	89.9 ± 1.1	73.2 ± 2.1	0.72 ± 0.02
		Val	66.0 ± 18.0	75.2 ± 4.6	39.2 ± 5.3	89.4 ± 7.0	72.8 ± 2.9	0.72 ± 0.06
L&C		Train	65.9 ± 1.7	81.0 ± 2.2	46.7 ± 2.1	90.4 ± 1.0	77.9 ± 2.0	0.75 ± 0.02
		Val	61.7 ± 12.1	81.3 ± 2.8	44.9 ± 5.0	89.3 ± 4.1	77.2 ± 2.6	0.73 ± 0.08

lighted in blue. In all instances (LDA vs. SVM and event vs. subject) the best performance is achieved using a 3s window and both LPC and MFCCs, indicating that LPC and MFCCs provide complementary information. There is also a large variability in the performance metrics, illustrated by the standard deviation in the tables (4.7, 4.8, 4.9 and 4.10), indicating that the results are dependent on the data used for training the model.

4.6 Discussion

For both classifiers the best results were consistently obtained using a combination of LPC and MFCCs and a window size of 3s regardless of whether the data were divided by subject or event. When divided by event, LDA achieved Ac = 77.9% in training and Ac = 77.2% during validation; when divided by subject, LDA achieved Ac = 77.4% in training, and Ac = 69.6% during validation. When divided by event, the SVM achieved Ac = 80.6% in training, and Ac = 80.7% during validation; when divided by subject, the SVM achieved Ac = 80.7% in training, and Ac = 80.9% during validation. It is clear that overtraining of the classifiers are

Table 4.8: *LDA statistics for LPC coefficients (LPC), MFCCs (CEP) and both combined (L&C) (mean±σ) for subjects divided into Training (Train) and Validation (Val).*

Inputs	τ (s)	Data	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
LPC	0.5	Train	69.5 ± 4.6	53.0 ± 2.3	39.7 ± 1.8	79.5 ± 3.5	58.1 ± 1.7	0.68 ± 0.04
		Val	66.0 ± 23.7	46.3 ± 12.1	33.8 ± 11.2	76.9 ± 14.6	50.4 ± 9.5	0.66 ± 0.18
CEP		Train	60.7 ± 3.5	59.1 ± 3.6	39.9 ± 2.3	77.0 ± 2.6	59.6 ± 1.7	0.66 ± 0.02
		Val	53.8 ± 14.7	53.1 ± 12.5	34.0 ± 7.7	71.0 ± 14.9	52.1 ± 6.1	0.63 ± 0.11
L&C		Train	69.1 ± 4.5	60.0 ± 1.7	43.4 ± 5.6	81.2 ± 4.4	62.7 ± 0.7	0.69 ± 0.02
		Val	63.4 ± 21.3	51.2 ± 12.5	34.6 ± 17.4	75.8 ± 22.4	52.8 ± 6.7	0.58 ± 0.11
LPC	1	Train	65.1 ± 3.5	50.7 ± 4.0	37.2 ± 1.5	76.4 ± 1.3	55.2 ± 2.1	0.65 ± 0.03
		Val	50.4 ± 20.8	49.8 ± 8.6	30.1 ± 9.7	70.1 ± 7.7	50.7 ± 3.6	0.58 ± 0.14
CEP		Train	55.2 ± 2.4	62.8 ± 2.3	39.9 ± 1.1	75.7 ± 2.5	60.5 ± 2.1	0.65 ± 0.03
		Val	48.0 ± 14.4	58.4 ± 13.2	34.0 ± 7.2	71.8 ± 7.1	54.3 ± 5.2	0.62 ± 0.13
L&C		Train	62.2 ± 8.6	61.3 ± 4.8	41.9 ± 4.2	78.4 ± 4.0	61.5 ± 3.6	0.67 ± 0.06
		Val	41.0 ± 17.6	57.8 ± 18.4	30.7 ± 9.1	68.9 ± 5.3	53.4 ± 7.8	0.59 ± 0.18
LPC	2	Train	60.3 ± 4.5	64.7 ± 5.2	40.5 ± 2.6	80.1 ± 4.5	63.4 ± 4.5	0.68 ± 0.04
		Val	50.0 ± 20.6	64.7 ± 4.1	33.9 ± 15.9	79.0 ± 11.5	61.1 ± 6.7	0.65 ± 0.15
CEP		Train	66.9 ± 4.4	68.0 ± 0.5	45.4 ± 5.1	83.6 ± 3.3	67.7 ± 1.4	0.72 ± 0.03
		Val	58.3 ± 10.6	65.8 ± 5.5	38.4 ± 21.7	80.2 ± 12.2	64.7 ± 5.7	0.65 ± 0.08
L&C		Train	64.5 ± 2.4	70.7 ± 1.1	46.9 ± 3.5	83.2 ± 1.0	68.9 ± 1.1	0.72 ± 0.02
		Val	52.1 ± 21.2	69.5 ± 9.2	40.1 ± 12.2	79.8 ± 8.0	65.3 ± 4.0	0.70 ± 0.13
LPC	3	Train	56.2 ± 12.9	72.1 ± 7.3	33.7 ± 7.9	87.2 ± 2.2	69.3 ± 4.5	0.66 ± 0.04
		Val	41.7 ± 36.3	72.3 ± 6.0	21.5 ± 20.1	83.7 ± 13.4	65.1 ± 5.1	0.62 ± 0.18
CEP		Train	68.8 ± 2.3	75.7 ± 2.8	41.7 ± 3.2	90.6 ± 0.8	74.3 ± 2.4	0.74 ± 0.03
		Val	60.2 ± 6.7	72.5 ± 13.2	36.8 ± 8.5	87.4 ± 5.1	70.2 ± 11.0	0.69 ± 0.16
L&C		Train	66.2 ± 4.8	80.2 ± 1.7	45.3 ± 5.4	90.4 ± 2.2	77.4 ± 1.9	0.75 ± 0.05
		Val	63.1 ± 28.0	74.7 ± 17.2	36.9 ± 23.0	88.2 ± 12.8	69.6 ± 13.5	0.75 ± 0.18

Table 4.9: *SVM statistics for LPC coefficients (LPC), MFCCs (CEP) and both combined (L&C) (mean±σ) for events divided into Training (Train) and Validation (Val).*

Inputs	τ (s)	Data	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
LPC	0.5	Train	0.0 ± 0.0	100.0 ± 0.0	NaN ± NaN	69.1 ± 1.3	69.1 ± 1.3	0.64 ± 0.01
		Val	0.0 ± 0.0	99.7 ± 0.6	0.0 ± 0.0	69.0 ± 5.2	68.9 ± 5.3	0.62 ± 0.05
CEP		Train	0.9 ± 0.8	99.2 ± 0.5	30.3 ± 10.0	69.1 ± 0.5	68.8 ± 0.6	0.66 ± 0.01
		Val	1.1 ± 1.5	98.7 ± 1.6	31.3 ± 47.3	69.0 ± 1.7	68.5 ± 1.9	0.63 ± 0.02
L&C		Train	2.4 ± 2.3	98.5 ± 1.4	40.8 ± 2.6	69.3 ± 0.9	68.8 ± 1.1	0.70 ± 0.01
		Val	3.5 ± 4.9	98.5 ± 1.7	41.7 ± 38.2	69.5 ± 4.4	69.1 ± 4.7	0.64 ± 0.06
LPC	1	Train	0.0 ± 0.0	100.0 ± 0.0	NaN ± NaN	69.0 ± 0.8	69.0 ± 0.8	0.62 ± 0.01
		Val	0.0 ± 0.0	100.0 ± 0.0	NaN ± NaN	69.0 ± 3.3	69.0 ± 3.3	0.56 ± 0.04
CEP		Train	0.6 ± 0.9	99.7 ± 0.5	50.0 ± 50.0	69.1 ± 1.2	69.0 ± 1.3	0.66 ± 0.02
		Val	1.3 ± 1.8	99.7 ± 0.6	75.0 ± 35.4	69.2 ± 5.3	69.2 ± 5.4	0.58 ± 0.05
L&C		Train	2.3 ± 2.5	99.2 ± 0.9	55.7 ± 5.2	69.3 ± 0.7	69.2 ± 0.6	0.69 ± 0.01
		Val	0.5 ± 1.2	98.7 ± 2.3	10.0 ± 14.1	68.9 ± 2.7	68.3 ± 3.0	0.59 ± 0.05
LPC	2	Train	5.3 ± 2.4	99.0 ± 0.7	71.3 ± 7.8	72.3 ± 0.8	72.2 ± 0.8	0.68 ± 0.01
		Val	5.2 ± 2.9	99.0 ± 1.6	78.3 ± 33.1	72.2 ± 3.6	72.1 ± 3.4	0.65 ± 0.03
CEP		Train	14.3 ± 7.3	96.6 ± 1.9	62.4 ± 7.5	73.8 ± 1.6	73.1 ± 1.9	0.73 ± 0.02
		Val	11.8 ± 4.2	94.9 ± 3.8	55.8 ± 29.3	72.8 ± 3.7	71.0 ± 4.0	0.69 ± 0.08
L&C		Train	15.3 ± 6.5	97.3 ± 1.1	69.1 ± 8.5	74.1 ± 1.3	73.8 ± 1.1	0.76 ± 0.01
		Val	17.2 ± 5.2	94.7 ± 5.2	61.0 ± 26.1	74.1 ± 2.5	72.5 ± 2.9	0.69 ± 0.05
LPC	3	Train	2.4 ± 1.8	99.4 ± 0.6	52.1 ± 17.2	80.2 ± 1.3	79.9 ± 1.3	0.72 ± 0.02
		Val	3.4 ± 4.8	99.0 ± 1.4	55.6 ± 50.9	80.3 ± 5.2	79.9 ± 5.6	0.67 ± 0.08
CEP		Train	18.0 ± 10.5	98.0 ± 1.2	70.1 ± 4.1	82.7 ± 0.4	82.0 ± 0.4	0.78 ± 0.03
		Val	14.8 ± 13.8	97.6 ± 2.2	54.6 ± 37.6	81.9 ± 7.3	80.6 ± 6.9	0.72 ± 0.13
L&C		Train	9.0 ± 5.1	98.7 ± 0.5	61.1 ± 7.9	81.2 ± 1.5	80.6 ± 1.3	0.81 ± 0.02
		Val	7.2 ± 9.1	99.1 ± 0.8	60.0 ± 17.3	81.0 ± 3.2	80.7 ± 3.1	0.72 ± 0.06

Table 4.10: *SVM statistics for LPC coefficients (LPC), MFCCs (CEP) and both combined (L&C) (mean±σ) for subjects divided into Training (Train) and Validation (Val).*

Inputs	τ (s)	**Data**	**Se** (%)	**Sp** (%)	**PPV** (%)	**NPV** (%)	**Ac** (%)	**AUC**
LPC	0.5	Train	0.5 ± 1.2	99.3 ± 0.6	11.1 ± 22.2	69.1 ± 2.3	68.8 ± 2.3	0.64 ± 0.03
		Val	0.8 ± 1.8	99.3 ± 1.5	25.0 ± 0.0	69.4 ± 9.0	69.0 ± 8.6	0.57 ± 0.08
CEP		Train	1.9 ± 3.5	98.6 ± 2.0	28.7 ± 34.0	69.3 ± 4.0	69.0 ± 4.3	0.67 ± 0.02
		Val	0.4 ± 1.0	98.1 ± 3.6	25.0 ± 35.4	71.1 ± 19.5	69.5 ± 16.9	0.57 ± 0.12
L&C		Train	2.7 ± 2.7	98.2 ± 1.6	30.7 ± 21.9	69.3 ± 2.1	68.7 ± 2.2	0.72 ± 0.02
		Val	1.4 ± 2.2	96.8 ± 3.6	17.5 ± 23.6	69.6 ± 9.9	67.7 ± 7.7	0.56 ± 0.07
LPC	1	Train	0.0 ± 0.0	100.0 ± 0.0	NaN ± NaN	69.1 ± 3.1	69.1 ± 3.1	0.62 ± 0.02
		Val	0.0 ± 0.0	100.0 ± 0.0	NaN ± NaN	71.1 ± 15.3	71.1 ± 15.3	0.53 ± 0.02
CEP		Train	0.5 ± 0.9	99.6 ± 0.4	25.0 ± 28.9	69.1 ± 2.9	69.0 ± 3.0	0.66 ± 0.02
		Val	0.8 ± 1.7	100.0 ± 0.0	100.0 ± 0.0	70.3 ± 11.6	70.4 ± 11.6	0.55 ± 0.10
L&C		Train	1.2 ± 1.2	99.2 ± 0.9	42.5 ± 6.6	69.1 ± 2.5	68.9 ± 2.7	0.70 ± 0.01
		Val	2.0 ± 3.5	98.0 ± 1.8	22.9 ± 31.5	69.6 ± 10.7	68.7 ± 10.2	0.56 ± 0.07
LPC	2	Train	7.9 ± 6.1	98.4 ± 2.2	72.6 ± 12.3	72.8 ± 2.5	72.7 ± 2.9	0.68 ± 0.02
		Val	2.4 ± 4.2	97.3 ± 4.2	43.8 ± 51.5	73.0 ± 15.3	70.9 ± 11.6	0.61 ± 0.05
CEP		Train	15.6 ± 7.7	96.3 ± 2.3	60.5 ± 16.3	74.0 ± 1.1	73.3 ± 1.7	0.73 ± 0.01
		Val	10.4 ± 7.7	94.8 ± 2.5	41.1 ± 33.2	72.4 ± 7.6	70.5 ± 6.0	0.63 ± 0.06
L&C		Train	16.6 ± 5.7	96.9 ± 2.0	71.8 ± 7.3	74.4 ± 2.8	74.2 ± 2.9	0.77 ± 0.01
		Val	17.0 ± 12.6	93.5 ± 7.0	53.6 ± 36.4	73.6 ± 15.6	71.0 ± 14.6	0.68 ± 0.12
LPC	3	Train	2.7 ± 2.2	99.6 ± 0.4	66.7 ± 23.6	80.3 ± 1.6	80.1 ± 1.7	0.73 ± 0.00
		Val	0.0 ± 0.0	98.3 ± 1.7	0.0 ± 0.0	79.4 ± 5.6	78.3 ± 5.3	0.57 ± 0.06
CEP		Train	19.3 ± 10.5	97.6 ± 1.5	67.7 ± 3.3	83.1 ± 3.0	82.2 ± 3.3	0.79 ± 0.02
		Val	17.1 ± 11.9	97.4 ± 4.2	58.3 ± 28.9	83.4 ± 14.0	81.9 ± 14.0	0.70 ± 0.13
L&C		Train	13.8 ± 9.5	97.5 ± 1.7	57.1 ± 8.7	81.9 ± 1.9	80.7 ± 1.6	0.81 ± 0.03
		Val	11.7 ± 14.9	98.3 ± 2.4	69.0 ± 27.0	81.8 ± 5.9	80.9 ± 5.2	0.70 ± 0.12

is taking place, indicated by the drop in performance from the training data to the validation data, as well as the large standard deviation values for some of the metrics. This is likely due to the lack of annotations *i.e.* there are insufficient data for this analysis or there are too many features. It should be noted that in X_{design} there are up to 24 features (when LPC+MFCCs are combined) but only 56 subjects. This causes the system to be under-specified, and leads to poor results on the validation data.

Based on the results in Tables 4.8 and 4.10, it is clear that it is not possible to do better than 75% on average when classifying by subject (69.6% for LDA, 80.9% for SVM). It is clear from Table 4.10 that the SVM is very specific (in the high 90s), but has very low sensitivity (less than 20%). In addition, the PPV could not be computed for some of the classifications using LPC indicating that in some instances the classifier never identifies a TP. It should be noted that, although accuracy is similar on the training data when dividing by subject and event, there is approximately a 4% difference when dividing by subject rather than event in the validation data, when using LDA. In addition, the performance when dividing by subject is more variable than when dividing by event (indicated by the consistently higher variance for all metrics) for both classifiers. Classifying by subject is more reasonable than classifying

by events as this ensures that there is no over-estimation of the results when there are events from a single subject in both the training and validation data sets.

The results were consistently better when the data were divided by event; there is less variability in the performance statistics than when dividing by subject. The fact that the combination of LPC and MFCCs provided the best accuracy in determining whether the event was either a choke or noise/snore indicates that the LPC coefficients and the MFCCs provide complementary information. A consideration for both classifiers is that less data was used at $3s$ than at $0.5s$. Table 4.6 shows that the number of events that are $3s$ or longer are actually very few (decreased from 391 noise/snore events, 175 chokes at $0.5s$ to 326 noise/snore, 82 chokes at $3s$). The ratio between the classes has completely changed at $3s$.

There are a number of limitations to this approach. Annotating the data is labour intensive and, ideally, there should be three annotators to ensure the quality of the annotation. In this work, one clinical research fellow with two years of training labelled the data which is not optimal.

It should be noted, that even with the higher accuracy obtained by dividing by event and assuming that the data were recorded in a low noise environment, it is insufficient for screening subjects. Even with the performance metrics as good as they are here, the accuracy needs to be much higher (at least 80%) for this approach to be clinically acceptable [55], although the ability of PMs to correctly diagnose subjects with OSA needs further validation (see section 2.5 for a full discussion).

The results presented here are lower than those reported in the literature (Table 3.3 shows that LPC [83] can achieve Se = 88% and Sp = 82%, while MFCCs [98] Se = 82%). However, many approaches consider in-sample classification which leads to a generous over-estimation of performance. In addition, the literature considers a different classification problem, *i.e.* thresholding on a given feature to differentiate apnoeic snores from benign snores, whereas, in the analysis above, the first breath after an apnoea has been differentiated from benign snores and noise, which will also influence performance. The poorer performance could be due to the lack of annotations, resulting in an under-specified system. In addition, none of the snoring events used in this analysis have been graded. The availability of a grading may have improved performance.

A two-step approach to classifying events could lead to an easier classification problem. The first step would involve removing artefacts, such as noise, followed by a two-class classification between snoring and choke events. This could lead to improved results, although the issue of sparse data remains. However, due to the lack of annotations/annotators this approach was not investigated.

An alternative approach was taken in the next chapter (Chapter 5) in an attempt to improve the classification accuracy sufficiently while minimising the pre-processing required.

Chapter 5

Processing audio data to classify OSA patients

5.1 Introduction

As demonstrated in Chapter 4, the classic speech analysis techniques only achieve Se = 66.2% and Ac = 69.6% during validation, when using LDA with LPC+MFCCs for a $3s$ window (Table 4.8). The large amount of pre-processing that was carried out, particularly in having to annotate individual events, is also a major drawback. In order to use these techniques, an event detector is required which would be another source of noise and errors. Even if a perfect detector could be built and there was no noise in the audio signal, the best possible classification accuracy would only be 80.7% for events (Table 4.9). In order to avoid, or at least limit this intensive process, and ideally improve classification performance, an alternative approach is needed; classifying by subject and not event. Classifying by subject does not require a detector and can use more contextual information regarding the changes in respiratory activity. In other words, using the information from multiple time scales which occurs in the audio signal [160] (as seen in Figure 5.3 there is clearly information on multiple time scales).

A multi-temporal method might provide a better screening metric to the classic techniques. Multiscale entropy (MSE) was originally applied to heart rate data, where visual inspection of the entropy values indicated that MSE coefficients could be used to distinguish between healthy and pathological groups. Heart rate data has similar issues to audio signals, such as non-stationarity. MSE measures the complexity of a signal, where more complex signals

Table 5.1: *Percentage of missing data for AHI, ODI and demographics.*

Feature	Missing data (%)
Gender	0.00
Age	0.00
Neck	8.39
Height	9.44
Weight	4.20
AHI	0.82
ODI	0.93
BMI	8.62
ESS	3.61

have higher MSE values. Complexity is associated with "meaningful structural richness" [161] incorporating correlations over multiple spatio-temporal scales. MSE also replicates the process that a clinician (Prof. John Stradling of the Respiratory Medicine Group, Churchill Hospital, Oxford, UK) uses to diagnose patients. Prof. Stradling [162] looks at how variable the signals are over the course of the night, with highly variable signals indicating that there is some form of sleep disturbance. However, MSE is an unintuitive way to look at multi-temporal information. A visual inspection of the audio signal in the time domain led to the idea that looking at the intervals between snores could provide a useful diagnostic tool (as suggested by Prof. Stradling [162]). There are very different patterns in the audio signal of apnoeics and snorers, for example. The intervals between snores could be a method of quantifying these differences.

5.2 Standard Metrics for OSA Diagnosis

As mentioned in Chapter 3, the AHI and ODI are two standard and commonly used metrics for identifying OSA severity, and are automatically calculated by the Visi-Download software. The patients in the Churchill Hospital are asked to fill out an ESS questionnaire as well as some basic information (age, gender, height, weight and neck size). Table 5.1 shows the amount of missing data for each of these features in the final population of 858 subjects; AHI was missing when actigraphy and body position were not recorded; ODI was missing when PPG was not recorded; if a subject did not answer all of the questions then those data were missing.

Using clinically accepted thresholds [163] for the AHI and ODI (both calculated by the software), and the STOP BANG questionnaire thresholds for the other features, the baseline

classification performance for these features can be found in Table 5.2. The STOP BANG uses one threshold for each of the features; if the feature is above the threshold the subject is said to have moderate or severe OSA. Both the AHI and ODI are associated with multiple thresholds used for classifying subjects into different categories. Subjects are said to be normal or a snorer if below a threshold of 5 and have mild, moderate or severe OSA otherwise; 10 is normal/snorer/mild vs. moderate/severe; 15 is also used for that division (normal/snorer/mild vs. moderate/severe); 20 is normal/snorer/mild/moderate vs. severe; 30 is used for that same division (normal/snorer/mild/moderate vs. severe). In this work, all thresholds have been tried to see how well they classify subjects. As can be seen in Table 5.2, the demographics are not very good at classifying subjects; they are either sensitive or specific, never both, and accuracy is only just better than random chance (ranging from 53% to 62%). This is not unexpected, particularly when subjects have been asked to note down these figures themselves rather than being measured and recorded by a healthcare professional; studies have shown that people are poor at self-reporting height and weight [164]. It would therefore be prudent to move away from the use of such information and perhaps rely on objective physiological signals only, if they provide a lower error rate.

Both the AHI and ODI are good classification features. Obviously, the higher threshold of 30 provides the best classification as this is the easiest classification problem of the commonly used thresholds, *i.e.* separating severe OSA from normal/snorer/mild OSA/moderate OSA. However, a threshold of 15, which classifies normal/snorer/mild vs. moderate/severe, is highlighted in blue in the table and is appropriate for this work, as this is the classification problem being addressed. Using this threshold gives performance statistics in the high-80 to low-90% range. The remainder of this work focuses on this type of classification which can be seen, in most countries, as non-treatment vs. treatment. These are the results to beat (Ac = 88.7%).

The results in Table 5.2 are based on clinical thresholds which have not been optimised for this data set. Table 5.3 shows the results across five folds, optimising the feature thresholds by maximising the accuracy. It is clear that optimising these thresholds improves classification accuracy in all cases. However, the increase is always small and the variation in the results means that these thresholds are specific to a particular data set. It is interesting to note that

Table 5.2: *Performance statistics when using clinical thresholds on the demographics, AHI and ODI where both AHI and ODI were automatically calculated by the software. The metrics in blue are the baseline to beat as this is the classification problem being addressed in this work (normal/snorer/mild OSA vs. moderate OSA/severe OSA).*

Feature	Threshold	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)
Gender	1.0	77.5	36.7	45.1	70.8	53.1
Age	50.0	61.7	59.0	50.2	69.6	60.0
Neck	40.0	84.9	40.4	51.7	78.1	59.5
BMI	35.0	45.0	73.8	53.4	66.8	62.3
ESS	15.0	46.4	66.4	48.5	64.5	58.3
AHI	5.0	97.4	55.3	59.2	96.9	72.1
AHI	10.0	92.9	80.1	75.6	94.4	85.2
AHI	15.0	83.5	87.6	81.7	88.8	86.0
AHI	20.0	71.4	94.4	89.5	83.2	85.2
AHI	30.0	53.0	97.9	94.4	75.8	80.0
ODI	5.0	97.6	54.2	58.7	97.2	71.6
ODI	10.0	94.0	81.1	76.8	95.3	86.3
ODI	15.0	85.3	90.9	86.2	90.3	88.7
ODI	20.0	74.3	96.0	92.5	84.9	87.3
ODI	30.0	56.2	98.6	96.4	77.2	81.6

some of the optimised thresholds are actually very close to those used clinically, specifically AHI (14.6) and ODI (16.9), both of which have a very tight range. The optimised thresholds differ from the clinical thresholds/values for the other features. Neck size, which also has a tight range, jumps from 40.0 to 43.4cm. Age, BMI and ESS all have larger ranges indicating that these mean threshold values are data set specific.

5.3 Data

The data described in section 4.2 were used in the audio analysis. See Table 4.2 for a breakdown of the demographics per subject diagnosis (normal, snorer, mild OSA, moderate OSA, severe OSA) and Appendix D for a breakdown of the statistical differences between the five groups. The demographics for each class (non-treatment vs. treatment) can be found in Table 5.4. The treatment group consists of moderate and severe OSA patients while the non-treatment group is made up of normal, snorer and mild OSA subjects. The distributions for each group can be found in Figure 5.1. Using the null hypothesis that the data in treatment and non-treatment groups are independent random samples from normal distributions with equal means and equal but unknown variances (a two-sample t-test), the hypothesis is rejected in

Table 5.3: *The performance statistics for each feature and the corresponding threshold per fold. Thold = threshold.*

Feature	Fold	Thold	Training					Validation				
			Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)
Age	1.0	68.5	12.5	93.7	56.7	62.1	61.7	14.7	92.3	55.6	62.3	61.6
	2.0	55.2	41.8	76.6	54.5	66.3	62.7	43.8	65.7	43.1	66.4	57.6
	3.0	65.3	17.6	92.0	57.5	64.4	63.6	20.8	84.0	51.6	56.4	55.6
	4.0	65.5	19.8	91.0	57.8	64.6	63.7	13.2	88.5	47.6	56.3	55.2
	5.0	55.0	42.1	71.9	51.5	63.7	59.5	42.6	82.9	53.5	75.8	70.2
		61.9 ± 6.3	26.8 ± 14.1	85.1 ± 10.0	55.6 ± 2.6	64.2 ± 1.5	62.2 ± 1.7	27.0 ± 15.0	82.7 ± 10.2	50.3 ± 5.0	63.4 ± 8.1	60.0 ± 6.2
Neck	1.0	42.9	65.4	66.8	58.7	72.8	66.2	66.2	63.7	57.7	71.6	64.8
	2.0	43.9	50.2	74.9	60.0	66.7	64.3	48.4	82.5	63.8	71.4	69.2
	3.0	43.2	67.2	67.0	58.1	75.0	67.1	60.0	62.0	60.0	62.0	61.0
	4.0	44.7	39.3	84.4	63.5	66.7	65.9	36.5	78.6	60.0	58.4	58.9
	5.0	42.9	66.7	65.2	60.4	71.1	65.9	61.5	66.3	47.8	77.5	64.7
		43.4 ± 0.8	57.7 ± 12.5	71.6 ± 8.0	60.1 ± 2.1	70.5 ± 3.7	65.9 ± 1.0	54.5 ± 12.0	70.6 ± 9.3	57.9 ± 6.1	68.2 ± 7.8	63.7 ± 4.0
Height	1.0	193.5	2.9	99.2	70.0	61.2	61.3	0.0	99.0	0.0	62.0	61.6
	2.0	193.0	2.1	98.9	55.6	61.1	61.1	3.3	100.0	100.0	62.2	62.7
	3.0	193.5	2.1	99.0	55.6	62.9	62.8	2.9	100.0	100.0	54.7	55.3
	4.0	193.0	2.5	99.2	66.7	62.1	62.2	1.6	98.9	50.0	58.3	58.2
	5.0	193.5	2.0	99.5	71.4	59.4	59.5	4.1	98.1	50.0	69.3	68.8
		193.3 ± 0.3	2.3 ± 0.4	99.2 ± 0.2	63.8 ± 7.8	61.4 ± 1.3	61.4 ± 1.3	2.4 ± 1.6	99.2 ± 0.8	60.0 ± 41.8	61.3 ± 5.4	61.3 ± 5.1
Weight	1.0	101.3	58.2	71.4	57.7	71.7	66.1	59.7	69.3	56.3	72.2	65.5
	2.0	101.4	57.8	71.5	58.1	71.3	66.0	61.3	68.7	55.1	73.9	65.8
	3.0	100.9	59.8	68.9	54.7	73.2	65.4	55.3	75.9	66.7	66.0	66.3
	4.0	113.7	34.9	82.5	56.2	66.3	63.9	33.3	80.2	57.1	60.3	59.5
	5.0	103.8	53.1	74.9	60.7	68.5	65.6	49.1	68.4	41.9	74.3	62.3
		104.2 ± 5.4	52.8 ± 10.3	73.8 ± 5.3	57.5 ± 2.3	70.2 ± 2.8	65.4 ± 0.9	51.7 ± 11.3	72.5 ± 5.3	55.4 ± 8.8	69.3 ± 6.0	63.9 ± 2.9
AHI	1.0	13.9	85.0	86.4	80.2	89.9	85.9	86.6	87.5	81.7	91.0	87.1
	2.0	15.5	81.2	89.5	83.7	87.8	86.2	84.1	91.6	85.5	90.7	88.8
	3.0	13.9	84.4	87.2	80.1	90.2	86.2	88.3	84.0	81.9	89.8	86.0
	4.0	13.9	86.2	86.7	80.0	91.0	86.5	82.4	86.3	82.4	86.3	84.6
	5.0	15.7	82.2	90.2	85.6	87.8	86.9	77.4	88.9	75.9	89.7	85.3
		14.6 ± 0.9	83.8 ± 2.1	88.0 ± 1.7	81.9 ± 2.6	89.4 ± 1.5	86.3 ± 0.4	83.8 ± 4.2	87.7 ± 2.8	81.5 ± 3.5	89.5 ± 1.9	86.4 ± 1.7
ODI	1.0	15.2	85.4	92.3	87.7	90.7	89.6	82.1	90.3	84.6	88.6	87.1
	2.0	17.5	79.0	94.6	90.7	87.2	88.4	82.5	95.3	91.2	90.3	90.6
	3.0	19.0	79.4	96.4	93.2	88.5	90.0	70.1	94.7	91.5	79.5	83.6
	4.0	16.1	82.7	92.6	87.4	89.7	88.8	83.8	93.7	91.2	88.1	89.3
	5.0	16.9	81.9	93.7	90.2	88.0	88.8	81.1	94.0	86.0	91.7	90.0
		16.9 ± 1.5	81.7 ± 2.6	93.9 ± 1.7	89.8 ± 2.4	88.8 ± 1.4	89.1 ± 0.6	79.9 ± 5.6	93.6 ± 2.0	88.9 ± 3.3	87.6 ± 4.8	88.1 ± 2.8
BMI	1.0	40.9	22.3	89.4	57.9	63.8	62.9	28.3	88.9	60.7	67.2	66.0
	2.0	41.0	21.2	89.2	55.9	63.7	62.6	29.0	90.7	66.7	66.7	66.7
	3.0	40.8	25.7	88.3	57.0	66.4	64.8	15.7	92.8	64.7	56.6	57.5
	4.0	31.3	65.7	62.8	52.5	74.5	63.9	66.2	62.2	55.8	71.8	63.9
	5.0	31.5	65.8	65.0	56.7	73.2	65.3	60.0	58.3	40.0	75.9	58.9
		37.1 ± 5.2	40.1 ± 23.4	79.0 ± 13.8	56.0 ± 2.1	68.3 ± 5.2	63.9 ± 1.2	39.8 ± 22.0	78.6 ± 16.8	57.6 ± 10.7	67.6 ± 7.2	62.6 ± 4.2
ESS	1.0	23.0	2.3	100.0	100.0	60.8	61.1	1.5	100.0	100.0	60.6	60.8
	2.0	17.0	24.7	84.8	52.4	62.5	60.6	19.4	88.3	50.0	64.5	62.4
	3.0	17.0	23.9	85.5	50.8	64.2	61.8	23.0	85.7	56.7	57.8	57.6
	4.0	18.0	17.7	90.6	54.2	63.8	62.6	16.0	87.0	50.0	55.9	55.1
	5.0	17.1	22.4	87.3	55.9	61.1	60.2	30.8	79.5	41.0	71.2	64.0
		18.4 ± 2.6	18.2 ± 9.3	89.7 ± 6.2	62.7 ± 21.0	62.5 ± 1.6	61.2 ± 1.0	18.1 ± 10.8	88.1 ± 7.5	59.5 ± 23.3	62.0 ± 6.1	60.0 ± 3.6

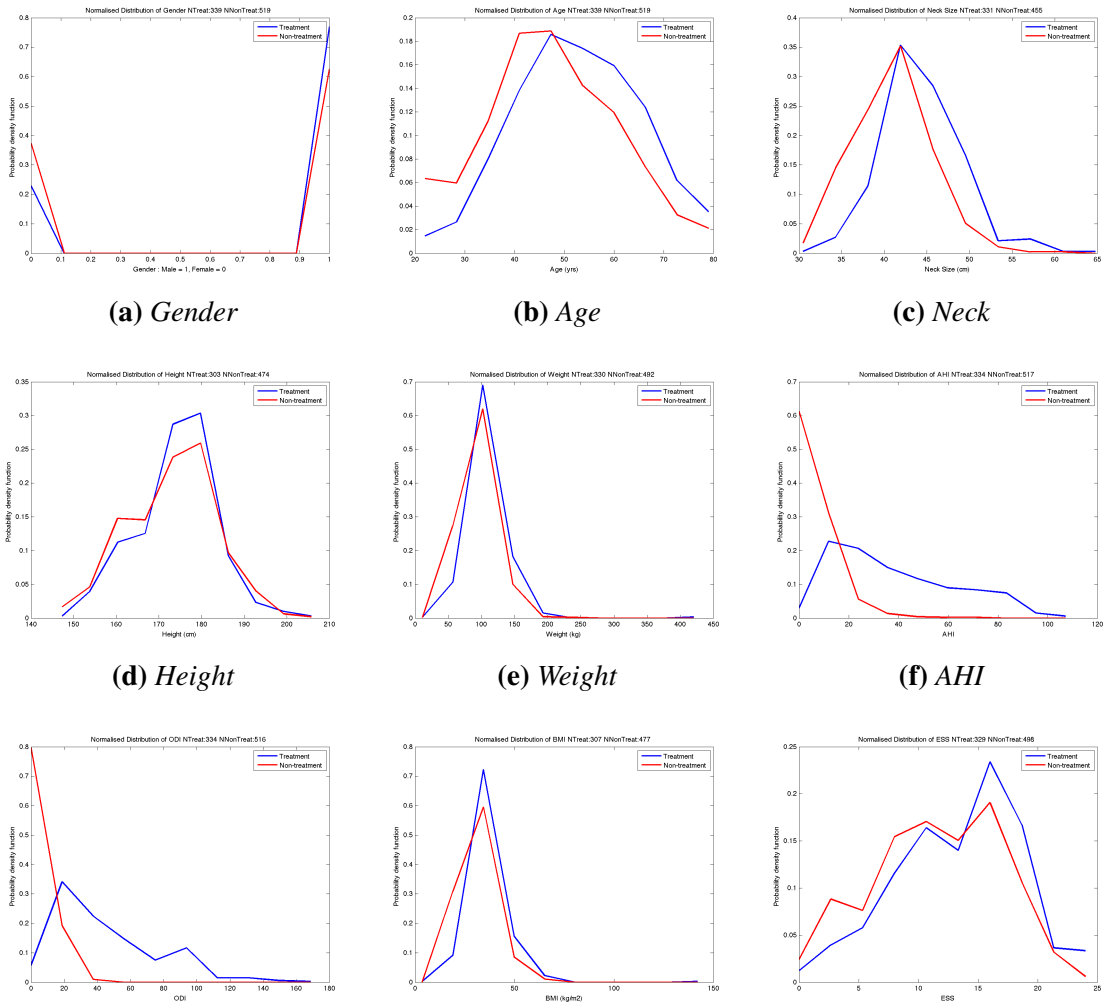
Table 5.4: *Patient demographics for the two classification groups: non-treatment (normal, snorer, mild OSA) and treatment (moderate OSA, severe OSA) (mean±σ) for the audio analysis. neck = neck circumference, m = male, f = female.*

	Non-treatment	**Treatment**
Gender	325 m, 194 f	261 m, 78 f
Age (yrs)	47.2 ± 13.7	52.7 ± 12.5
Neck (cm)	40.9 ± 4.6	44.2 ± 4.6
Height (cm)	173.0 ± 10.4	174.2 ± 9.4
Weight (kg)	93.5 ± 23.7	107.8 ± 30.6
AHI ($events/h$)	6.6 ± 8.0	38.1 ± 24.2
ODI ($events/h$)	6.2 ± 5.7	44.3 ± 31.6
BMI (kg/m^2)	31.3 ± 8.2	35.7 ± 10.4
ESS	11.8 ± 5.2	13.6 ± 5.1

Table 5.5: *Hypothesis (1: reject null hypothesis, 0: accept null hypothesis) and p value for the treatment vs. non-treatment demographics.*

Demographic	KS test			t-test	
	H	p	KS statistic	H	p
Gender	1	3.6×10^{-4}	0.14	1	8.9×10^{-6}
Age	1	7.3×10^{-7}	0.19	1	2.3×10^{-9}
Neck	1	1.0×10^{-17}	0.32	1	1.3×10^{-22}
Height	0	0.0579	0.09	0	0.0691
Weight	1	2.5×10^{-16}	0.30	1	1.2×10^{-13}
AHI	1	1.6×10^{-99}	0.74	1	2.3×10^{-118}
ODI	1	2.4×10^{-110}	0.78	1	5.9×10^{-115}
BMI	1	1.9×10^{-16}	0.31	1	5.1×10^{-11}
ESS	1	2.7×10^{-4}	0.15	1	1.7×10^{-6}

every case except for height at the 5% significance level with $p << 0.001$. In the case of height, the t-test found that the two groups come from a distribution with equal means and equal but unknown variances with $p = 0.0691$. This is not unexpected in that the typical OSA patient is older and heavier (and hence has a larger neck size, weight and BMI); as an OSA sufferer it would be expected that the AHI, ODI and ESS would be higher than for a subject without OSA. Full details for the paired t-test and the two-sample Kolmogorov-Smirnov test can be found in Table 5.5. Overall, there are 519 subjects in the non-treatment class (normal, snorer, mild OSA) (60.49%) and 339 subjects in the treatment class (moderate OSA, severe OSA) (39.51%).



(a) *Gender* **(b)** *Age* **(c)** *Neck*

(d) *Height* **(e)** *Weight* **(f)** *AHI*

(g) *ODI* **(h)** *BMI* **(i)** *ESS*

Figure 5.1: *Probability density functions for both classification groups for the normalised demographics. Treatment is shown in blue and non-treatment in red.*

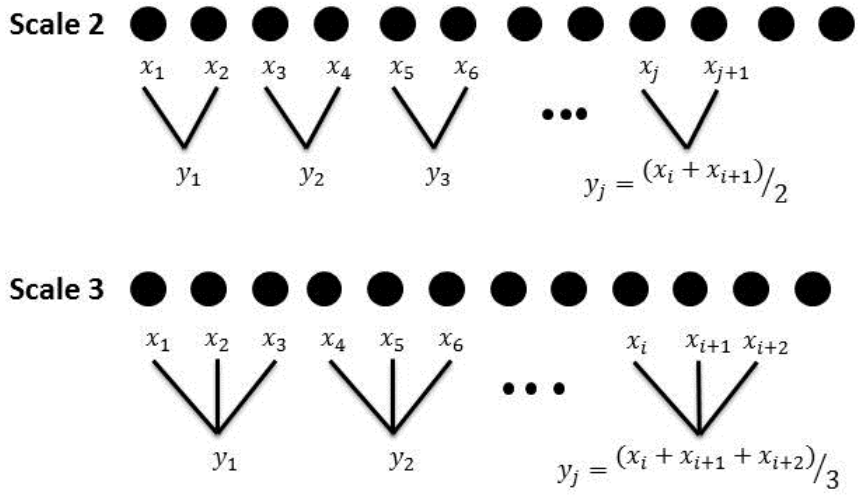


Figure 5.2: *Illustration of the coarse-graining process. Adapted from [169].*

5.4 Methods

As stated in section 5.1, multiscale entropy was chosen to replicate the variability of the signals, while the inter-snore interval was used to identify two patterns: snoring and apnoea.

5.4.1 Multiscale Entropy

Multiscale entropy (MSE) is a method of measuring the complexity of a finite length time series [165–167]. MSE has been applied to heart rate data, which has similar issues to the audio signal, such as non-stationarity. Costa *et al.* [168] noted that traditional algorithms indicated that certain pathological processes had a higher complexity than healthy dynamics with long-range correlations. The authors suggested that this paradox was due to the fact that conventional algorithms fail to account for the multiple time scales inherent in healthy physiological dynamics. Due to this hypothesis, MSE was developed and was found to robustly separate healthy and pathological groups. Given an N-point time series, $\{x_1, \ldots, x_i, \ldots, x_N\}$, a consecutive coarse-grained time series can be constructed by averaging a successively increasing number of data points in non-overlapping windows (illustrated in Figure 5.2).

Each element of the coarse-grained time series, $y_j^{(\tau)}$, is calculated according to the equation:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i \tag{5.1}$$

where τ represents the scale factor and $\{1 \leq j \leq N/\tau\}$. The length of each coarse-grained time series is N/τ. For scale $\tau = 1$, the coarse-grained time series is simply the original time series. The sample entropy ($\mathcal{H}_S$) is then calculated for each of the time series and can be plotted as a function of the scale factor. $\mathcal{H}_S$ quantifies the regularity of a time series and is the negative natural logarithm of the probability that two sequences similar for m points remain similar at the next point, where self-matches are not included. Given N data points from a time series $x(n) = x(1), x(2), \ldots, x(N)$, the algorithm forms $N - m + 1$ vectors $X(1), \ldots, X(N - m + 1)$ defined by $X(i) = [x(i), x(i + 1), \ldots, x(i + m - 1)]$, for $\{1 \leq i \leq N - m + 1\}$. The vectors, X, represent m consecutive values of the signal, commencing with the i^{th} point. The distance between $X(i)$ and $X(j)$, $d = [X(i), X(j)]$, is then calculated as the maximum absolute difference between their respective scalar components as follows:

$$d[X(i), X(j)] = max_{k=1,2,\ldots,m}(|x(i + k) - x(j + k)|) \tag{5.2}$$

For a given $X(i)$, the number of j's $\{1 \leq j \leq N - m, i \neq j\}$ are counted, such that the distance between $X(i)$ and $X(j)$ is less than or equal to r standard deviations and the following function is calculated:

$$B_r^m(i) = \frac{1}{N - m - 1} \sum_{j=1, j\neq i}^{N=m} \Theta(r.\sigma - d[X(i), X(j)]) \tag{5.3}$$

where Θ is the Heaviside function ($\Theta(z \geq 0) = 1$) and ($\Theta(z \leq 0) = 1$), σ is the standard deviation of the signal $x(n)$ and r is a tolerance window. B_r^m is calculated as follows:

$$B_r^m = \frac{1}{N - m} \sum_{i=1}^{N-m} B_r^m(i) \tag{5.4}$$

The dimension is then increased to $m + 1$ and $A_r^m(i)$ is calculated as:

$$A_r^m(i) = \frac{1}{N - m - 1} \sum_{j=1, j\neq i}^{N=m} \Theta(r.\sigma - d[X(i), X(j)]) \tag{5.5}$$

A_r^m is then given by:

$$A_r^m = \frac{1}{N - m} \sum_{i=1}^{N-m} A_r^m(i) \tag{5.6}$$

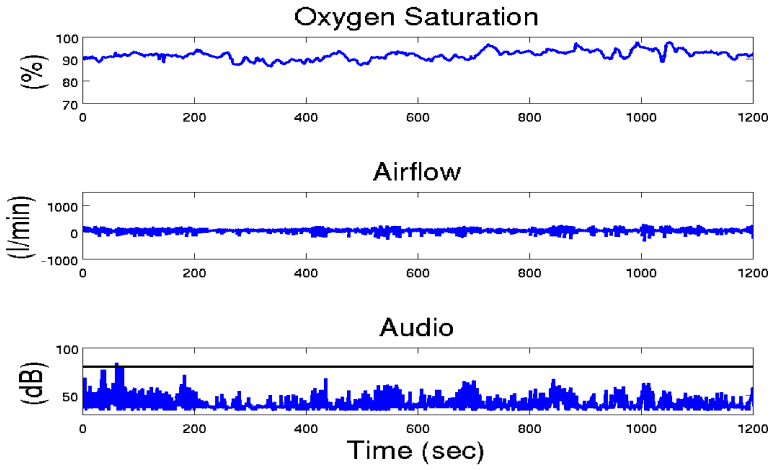
and the sample entropy is given by the negative logarithm of the ratio of A_r^m to B_r^m:

$$\mathcal{H}_S(m, r, N) = -\ln\left(\frac{A_r^m}{B_r^m}\right) \tag{5.7}$$

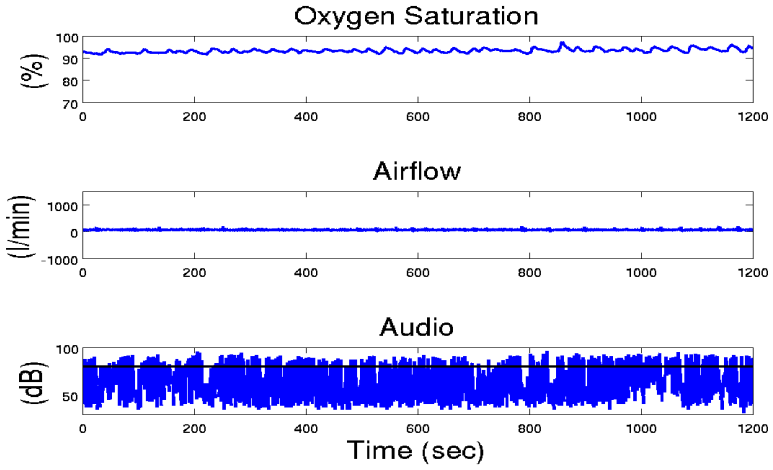
From these equations it is clear that $\mathcal{H}_S$, and hence, MSE is a function of three parameters: m, r and N.

5.4.2 Inter-snore Interval

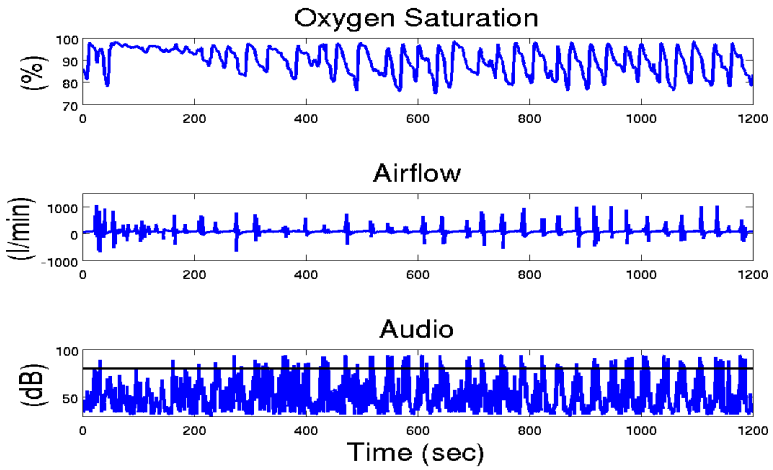
The inter-snore interval (ISI) can be viewed as an intuitive measure of sleep disturbance. This metric was proposed by our clinical collaborator (Prof. Stradling) through years of experience. He felt that it would be possible to distinguish between apnoeics and non-apnoeics by looking at the interval between snores (or at least loud noises assumed to be snoring). Figure 5.3 shows the SpO_2, airflow and audio signals for three subjects; a normal subject, a snorer and a subject with severe OSA. On the audio channel for each subject there is a black line: data above this line would be taken as snoring peaks. It is clear that there are three very different patterns: one for each of the subjects. Using this as the starting point, four metrics were derived that quantify the differences between the two classes. An apnoea event ends when the subject experiences a momentary arousal (not enough to wake them up fully) in order to resume breathing. This causes sleep fragmentation which translates to a characteristic pattern in the audio signal of loud snoring or brief gasps with intervals of silence [16, 160, 170]. These micro arousals can be seen as a series of peaks in the signal. Sleep fragmentation (D_f) can be measured by counting the number of peaks and dividing by the number of hours of sleep. The severity of OSA and the number of apnoeic events per hour of sleep are directly related, *i.e.* the more severe the condition, the more apnoea events per hour of sleep the subject experiences. Therefore, the intervals between the peaks may provide information regarding OSA severity. The more severe the apnoea, the more the interval between the apnoeic events decreases leading to less variability in the intervals between the peaks. The intervals between the peaks can be characterised by taking the standard deviation (D_s) and the mean absolute deviation (D_m) of the peak intervals. It is possible that the amplitude of the peaks represents the amount of activity per arousal and so the mean amplitude of the peaks (D_h) may also be useful. Two methods, described in section 5.5, were used to calculated these metrics.



(a) *Normal subject*



(b) *Snoring subject*



(c) *Subject with severe OSA*

Figure 5.3: *Three examples of Grey Flash data illustrating the ISI process; a normal subject, a snoring subject and a subject with severe OSA. The black line on each of the audio channels is used to represent the peaks that the ISI captures; data points above this line would be taken to be peaks and used to calculate the four metrics. The audio signal for the snorer and normal subject have similar patterns to each other, whereas the audio signal for the subject with severe OSA has a completely different pattern.*

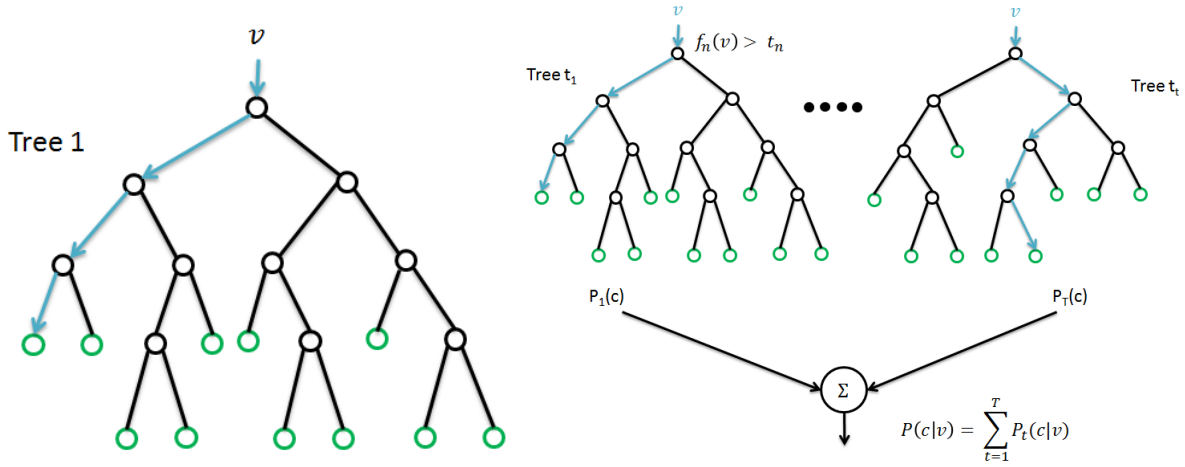
5.4.3 Classification

LDA, as described in section 4.3.3 in Chapter 4, was used in this section of the analysis. Another classifier, random forests (RFs), was also used and the details can be found below. RFs can handle larger numbers of features with small numbers of observations. This is unlike SVMs and so a RF was used instead of a SVM in the rest of the analysis.

5.4.3.1 Random Forest

Random forests (RFs) are a type of ensemble classifier based on decision trees [171]. Decision trees form a predictive model which uses a set of binary rules to calculate a target value. Training data is passed to the decision tree, which builds a model determining which variable to split on at a given node, what the value of the split is, whether to stop or to split again and when to assign a terminal node to a class. When a large number of trees have been generated, they vote for the most popular class. For the k^{th} tree, a random vector Θ_k is generated, independent of the past random vectors $\Theta_1, \ldots, \Theta_{k-1}$ but with the same distribution. A tree is grown using the training set and Θ_k, resulting in a classifier $h(\mathbf{x}, \Theta_k)$ where $\mathbf{x}$ is an input vector. A RF is a classifier consisting of a collection of tree-structured classifiers $h(\mathbf{x}, \Theta_k), k = 1, \ldots$ where the Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$. Figure 5.4a shows an individual tree while Figure 5.4b shows the forest that is made up of t trees.

A tree based classifier was developed using a Bayesian framework by Johnson *et al.* [172]. The algorithm has many advantages, including high overall performance and automatic handling of missing data, outliers and normalisation. Each tree selects a subset of observations via two regression splits. These observations are then given a contribution, equal to a random constant times the observation's value for a chosen feature plus a random intercept. Furthermore, the tree also assigns a contribution to missing values for this chosen feature based upon a scaled surrogate. The contributions across all trees are summed to provide the contribution for a single "forest", where a "forest" refers to a group of trees plus an intercept term. The predicted probability output by the forest is the inverse logit of the sum of each tree's contribution plus the intercept term. The intercept term is set to the logit of the mean observed outcome.



(a) *Decision tree* **(b)** *Random forest*

Figure 5.4: *An individual decision tree is shown in 5.4a while a forest comprised of a multitude of trees is shown in 5.4b.*

The core of the new model is the custom Markov chain Monte Carlo sampler which iteratively optimises the forest. This sampling process has a user defined number of iterations and a user defined number of resets (each reset involves reinitialising the forest and restarting the iterative process). After mapping the training data onto the quantiles of a normal distribution, the forest is initialised to a null model, with no contributions assigned for any observations.

At each iteration, the algorithm selects two trees in the forest and randomises their structure. That is, it randomly reselects the first two features which the tree uses for splitting, the value at which the tree splits those features, the third feature used for contribution calculation, and the multiplicative and additive constants applied to the third feature. The total forest contribution is then recalculated and a Metropolis-Hastings acceptance step is used to determine if the update is accepted. The Metropolis-Hastings algorithm is a Markov chain Monte Carlo (MCMC) method[1] for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult [173, 174]. If the update is accepted, the two trees are kept in the forest, otherwise they are discarded and the forest remains unchanged. After a set fraction of the total number of iterations to allow the forest to learn the target distribution (20%), known as the burn-in period, the algorithm begins storing forests at a fixed interval, *i.e.* once every set number of iterations. Once the number of user-defined iterations are reached,

[1]MCMC methods involve sampling from probability distributions by constructing a Markov chain (a memoryless mathematical system that undergoes transitions from one state to another, among a finite number of possible states) that has the desired distribution as its equilibrium distribution.

the forest is reinitialised as before and the iterative process restarts. Again after the set burn-in period, the forests begin to be saved at a fixed interval. The final result of this algorithm is a set of forests, each of which will contribute to the final model prediction.

5.5 Data Analysis Protocol

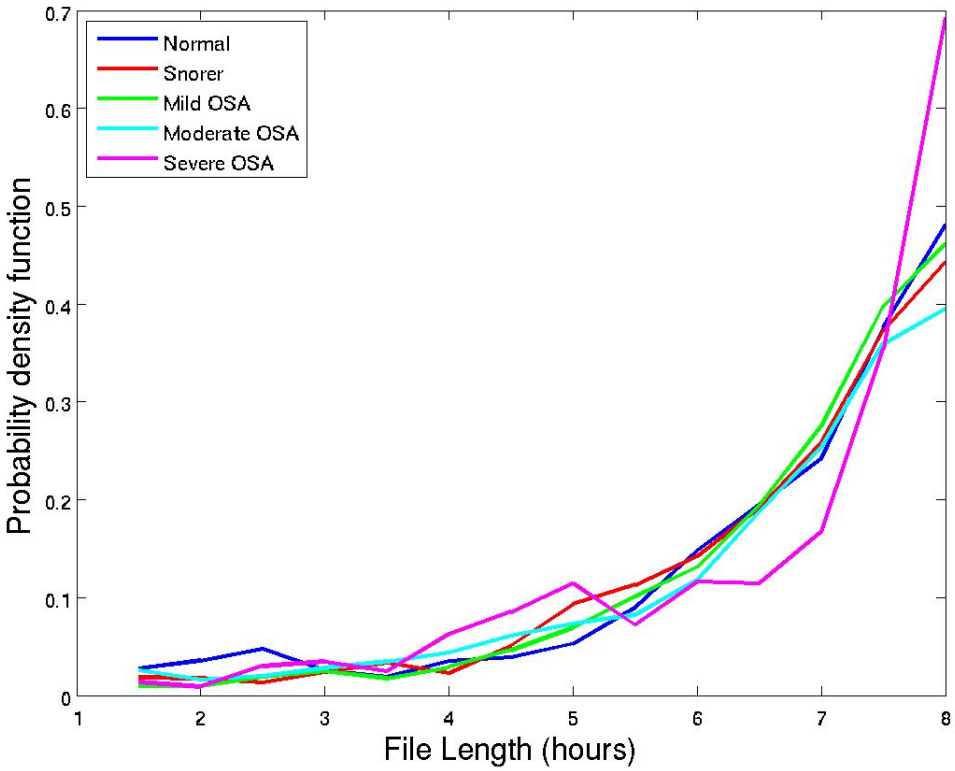
This section details the exact process that was followed and the values of parameters used in analysing the audio signals.

5.5.1 MSE

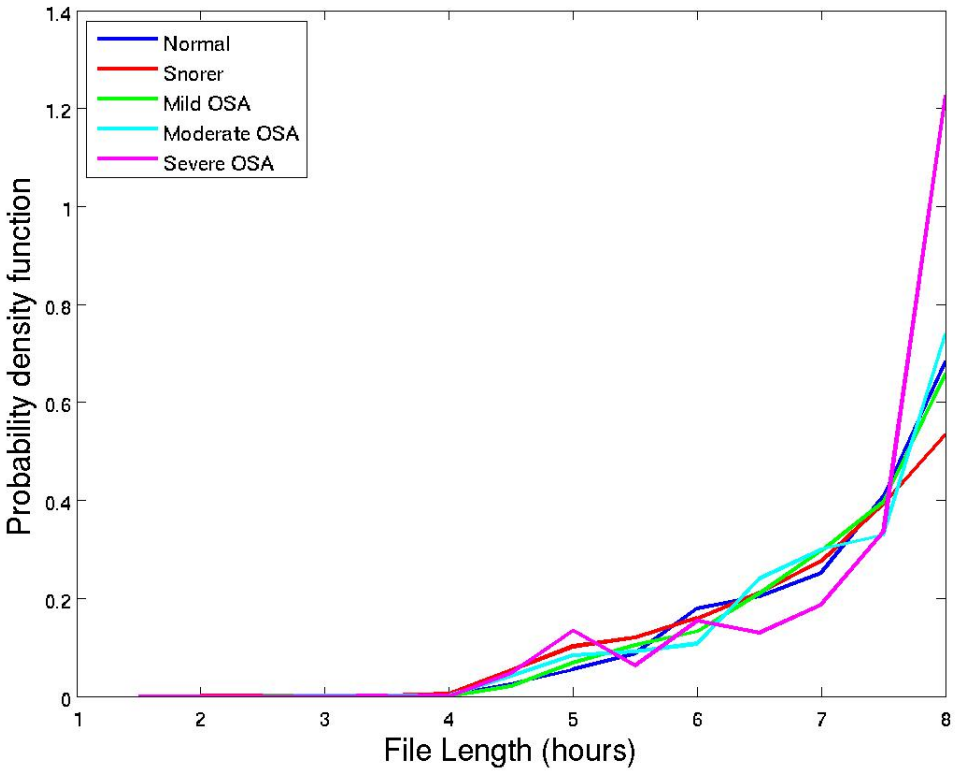
Only 240min of the audio data were analysed, beginning 30min into the recording and ending at 4.5h. This maximised the data, as well as only looking at 4h of data as suggested by a clinical expert [162]. Expert opinion indicates that if a subject has OSA, it will be apparent in 4h of sleep. This prevents the exclusion of short recordings, which are associated with OSA. Figure 5.5 shows the probability density function of the file lengths for the five different diagnoses (normal, snorer, mild OSA, moderate OSA and severe OSA). The start time was taken 30min into the recording to allow the user to fall asleep, although this assumes that all subjects are asleep within 30min. The data were preprocessed by taking the variance every 0.5, 1 or 2s and then the natural logarithm of that time series was taken. This process highlighted the peaks in the signal, which can be seen in Figure 5.6. Nine MSE coefficients were calculated per subject ($\tau = 1, 2, 4, 8, 16, 32, 65, 130, 180$) for $m = 1 : 1 : 8$ and $r = 0.1 : 0.05 : 0.25$. The scales chosen attempted to capture the time scales that occur during repeated apnoeas at both short and long time scales. The values used for m and r are based on reasonable ranges for physiological data taken from [166].

5.5.2 ISI Method 1

This method (ISI1) is a quick and easy way of finding the peaks in the audio signal. The data were preprocessed from 4kHz to 1Hz by taking the variance per s of data and then normalised over the whole night by subtracting the mean and dividing by the standard deviation. All local peaks were found that had a minimum distance of 10 samples between them; this equates

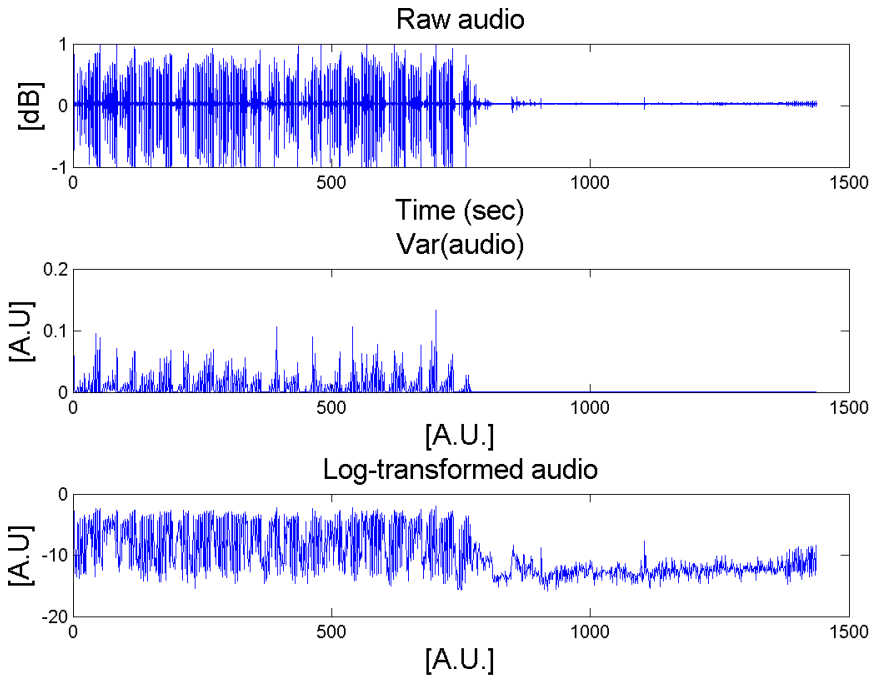


(a) *Full population*

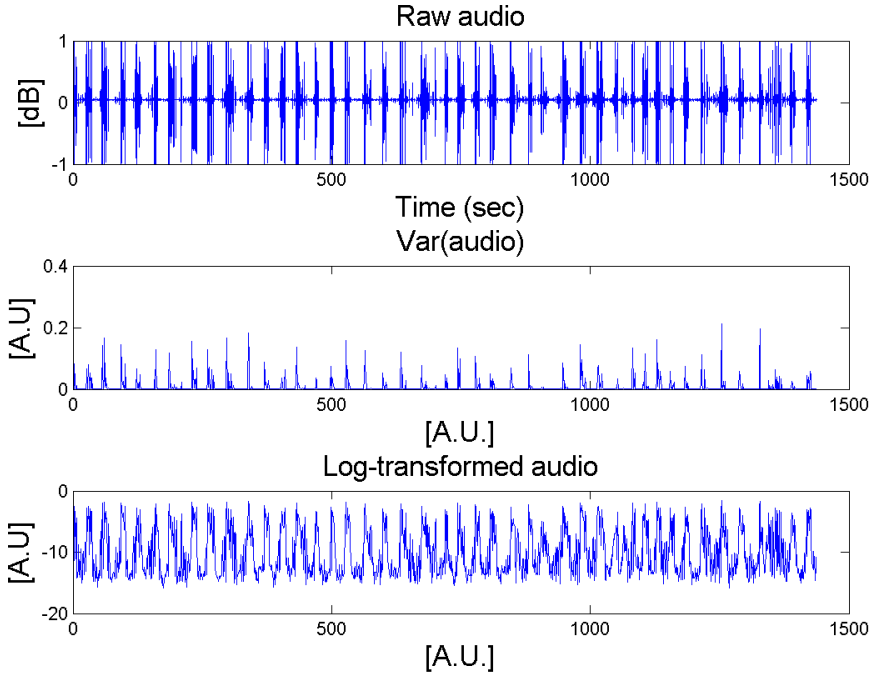


(b) *Final population*

Figure 5.5: *File durations for normal subjects (blue), snoring subjects (red), mild OSA subjects (green), moderate OSA subjects (cyan) and severe OSA subjects (magenta). 5.5a shows the distributions for the full population collected with appropriate diagnoses (1014 subjects) while 5.5b shows the distributions for the final population used in the analysis (858 subjects).*



(a) *Snorer*



(b) *Apnoeic*

Figure 5.6: *The raw audio signal, the downsampled audio signal and the downsampled and* ln*-transformed audio signal for a snorer and an apnoeic subject. A.U. = arbitrary units.*

to a minimum of $10s$ between peaks, which was chosen given the definition of an apnoea (the cessation of airflow at the nose and mouth lasting at least $10s$ [16]). Each peak was then analysed and if it was below a local threshold, it was removed. It was then possible to calculate the statistics described in section 5.4.2 *i.e.* D_f, D_s, D_m and D_h.

5.5.3 ISI Method 2

This second method (ISI2) is based on the Pan-Tompkins QRS detector [175]. This method has a longer computational time than ISI1 but involves no preprocessing of the data. It also has the advantage of finding more peaks; ISI1 finds local peaks but the actual highest peak may be discounted as there must be a minimum of 10 samples, which due to the preprocessing equates to $10s$, between peaks. ISI2 removes this uncertainty about whether the peak is the actual maximum or not.

The Pan-Tompkins algorithm is essentially an energy detector. For this application the algorithm is changed slightly. The data were read in; detrended (mean removed), squared and integrated over a $500ms$ window; then filtered and differentiated. Peaks that were caused by the data decreasing were removed and the delay caused by the filtering was removed by scanning back through the signal. A threshold was set (15% of the 99^{th} percentile) and an array of segments built. In each segment, the maximum was found and those peaks that were too close *i.e.* physiologically impossible, were removed. The minimum number of samples between detected peaks was $0.4 \times f_s$, where f_s is the sampling frequency, which equates to $0.4sec$ between detected peaks. The amplitude and location of each peak was noted and the statistics (D_f, D_s, D_m, D_h) described in section 5.4.2 were calculated.

5.5.4 Classification

Performance for individual demographics is presented in Table 5.2. For a full comparison, different combinations of ODI and demographics were investigated using LDA. Five-fold cross-validation was carried out on the data, where one fold was used as X_{val} while the remaining four folds were used as X_{design}.

The same process (five-fold cross-validation) was carried for the MSE analysis. In this case, one fold was held separately to be the validation set (X_{val}) while the other four folds

were used as the design data set. X_{design} was further divided into training and test data sets (70% and 30%) n times in order to find the best MSE downsampling rate (dsr), m value and r value. This was done by using the current classifier to classify every possible combination of dsr, m, r. For LDA this was carried out 100 times where the MSE coefficients were normalised (zero mean and unit variance based on the training data); for the RF, this process was carried out 5 times. The difference was due to the computational time. The highest accuracy was noted along with the dsr, m, r combination that it corresponded to. The best overall combination was taken to be the one that was chosen most often in the n iterations.

When the best dsr, m, r combination had been found using the training and testing data, the full four folds (*i.e.* all of X_{design}) were used to train the classifiers and the untouched fold was used for validation (X_{val}). Both classifiers were used to test all combinations of features, namely: MSE, $ISI1$, $ISI2$, $MSE + ISI1$, $MSE + ISI2$, $MSE + ISI1 + ISI2$, $MSE + ISI1 + ISI2 + demographics$, $MSE + ISI1 + ISI2 + ODI$, and $MSE + ISI1 + ISI2 + demographics + ODI$. In each case, the MSE coefficients that were used were those found in the first stage, *i.e.* the MSE coefficients that corresponded to the dsr, m, r combination that gave the highest accuracy for each classifier. The results were noted for each feature combination, classifier and fold.

For LDA, any missing data in X_{design} were mean imputed and then normalised (zero mean and unit variance based on X_{design}). A multivariate normal density was fitted to each group with a diagonal covariance matrix estimate (essentially a naive Bayes classifier). Naive Bayes classification assumes that all features are independent within each class. Although this assumption is not true in this case, these classifiers are known to work well even when the independence assumption is not valid, and so it is possible to use them to estimate classification accuracy. For the RF, 500 trees were used in the forest. Each tree split on three variables/features. The process was repeated twice with a new seed for 2×10^6 iterations.

5.6 Results

Table 5.6 shows the performance metrics for different combinations of ODI and demographics when using LDA. The best performance is highlighted in blue.

Figure 5.7 shows the MSE values at the different scale factors (τ) for all combinations of

Table 5.6: *Performance statistics when using LDA on combinations of demographics and ODI to classify treatment vs non-treatment, where ODI was automatically calculated by the software. The best performance is highlighted in blue.*

Features	Data Set	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
ODI+gender	Train	57.0 ± 0.7	98.6 ± 0.2	96.6 ± 0.4	77.5 ± 0.4	82.0 ± 0.3	0.95 ± 0.00
	Val	57.3 ± 4.8	98.6 ± 0.9	96.6 ± 1.4	77.7 ± 1.8	82.2 ± 1.6	0.95 ± 0.01
ODI+age	Train	59.0 ± 1.8	98.6 ± 0.2	96.5 ± 0.6	78.3 ± 0.7	82.8 ± 0.7	0.94 ± 0.00
	Val	58.4 ± 4.6	98.6 ± 0.8	96.4 ± 2.2	78.1 ± 1.9	82.6 ± 2.0	0.94 ± 0.02
ODI+neck	Train	56.6 ± 1.4	98.1 ± 0.2	95.7 ± 0.4	75.2 ± 0.6	80.4 ± 0.5	0.95 ± 0.00
	Val	55.6 ± 5.9	98.2 ± 0.9	95.9 ± 1.4	74.9 ± 1.9	80.1 ± 1.9	0.95 ± 0.01
ODI+height	Train	56.5 ± 0.7	98.5 ± 0.2	96.3 ± 0.4	77.4 ± 0.4	81.8 ± 0.2	0.95 ± 0.00
	Val	56.7 ± 4.0	98.3 ± 0.8	95.7 ± 1.6	77.5 ± 0.6	81.8 ± 0.7	0.94 ± 0.01
ODI+weight	Train	56.3 ± 1.5	98.3 ± 0.2	95.9 ± 0.3	76.6 ± 0.5	81.2 ± 0.4	0.94 ± 0.00
	Val	56.0 ± 5.8	98.3 ± 0.8	96.0 ± 1.4	76.6 ± 1.1	81.2 ± 1.3	0.94 ± 0.01
ODI+BMI	Train	56.8 ± 1.3	98.3 ± 0.2	95.7 ± 0.3	77.4 ± 0.4	81.7 ± 0.3	0.94 ± 0.00
	Val	55.9 ± 4.6	98.3 ± 0.8	95.7 ± 1.4	77.1 ± 0.4	81.4 ± 0.6	0.94 ± 0.02
ODI+ESS	Train	54.4 ± 1.6	98.7 ± 0.2	96.5 ± 0.4	76.2 ± 0.4	80.8 ± 0.4	0.95 ± 0.00
	Val	54.4 ± 6.0	98.7 ± 0.5	96.7 ± 1.0	76.3 ± 1.5	81.0 ± 1.8	0.95 ± 0.00
ODI+demos	Train	59.8 ± 1.1	97.8 ± 0.4	95.2 ± 0.7	77.4 ± 0.4	82.0 ± 0.4	0.93 ± 0.00
	Val	60.6 ± 2.9	97.6 ± 1.2	94.9 ± 2.3	77.7 ± 1.0	82.3 ± 0.9	0.93 ± 0.01

m, r, dsr for a snoring subject and a severe apnoeic. This figure clearly indicates that there are differences in the MSE value between the two subjects. Figure 5.8 shows the number of times each MSE parameter combination was chosen per fold for both classifiers. Figure 5.8a shows that when using LDA to choose the best parameter combination the results were: for the first fold, the best MSE parameter combination was $dsr = 0.5, m = 8, r = 0.1$ which was chosen 36 times. In the second fold, the best combination was $dsr = 1, m = 8, r = 0.1$ chosen 29 times, while in fold three the best combination was $dsr = 2, m = 7, r = 0.1$ chosen 30 times. For folds four and five the best combination was $dsr = 2, m = 6, r = 0.1$ and $dsr = 2, m = 8, r = 0.1$ chosen 43 and 25 times respectively. It is clear that the r value is of most importance, with all combinations using $r = 0.1$, while an m of 7 or 8 was chosen for four of the five folds. The dsr value is the parameter that changes most across the folds, although a dsr of 2 is chosen in three of the five folds. Overall, all of the best combinations were for $r = 0.1$, with $m = 7$ being chosen 40% of the time and $m = 8$ being chosen for another 40% of the time.

Figure 5.8b shows that when using a RF to choose the best parameter combination the results were: for the first fold, the best MSE parameter combination was $dsr = 0.5, m = 1, r = 0.15$ which was chosen twice. For the second fold, the best combination was $dsr = 0.5, m = 1, r = 0.25$ chosen twice while $dsr = 0.5, m = 1, r = 0.1$ was chosen only once in the third fold. For the fourth fold, $dsr = 1, m = 1, r = 0.1$ chosen twice and in the fifth

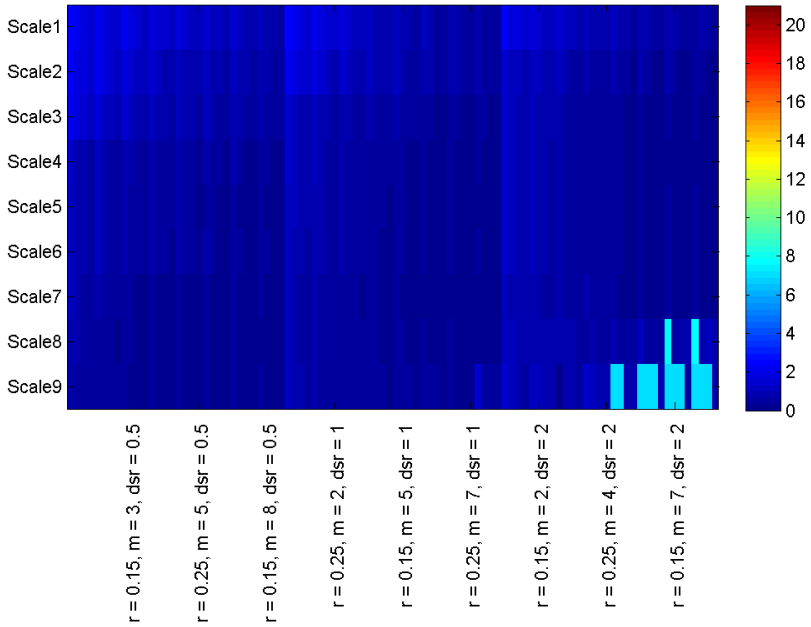
Table 5.7: *The average performance statistics over all five folds (mean±σ) for different audio feature combinations on both the training (Train) and validation (Val) data sets when using LDA to classify subjects.*

Features	Data Set	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{aud}$	Train	42.0 ± 14.0	81.8 ± 9.5	62.5 ± 7.2	68.7 ± 2.8	66.2 ± 1.1	0.66 ± 0.01
	Val	41.1 ± 14.3	78.5 ± 11.7	58.8 ± 15.6	67.1 ± 8.2	63.3 ± 5.2	0.64 ± 0.04
ISI1	Train	33.2 ± 3.9	85.2 ± 2.1	59.4 ± 1.0	66.1 ± 1.7	64.7 ± 1.5	0.65 ± 0.01
	Val	34.4 ± 6.4	85.5 ± 1.5	60.3 ± 4.9	66.5 ± 6.9	65.1 ± 5.5	0.65 ± 0.03
ISI2	Train	33.2 ± 3.9	85.2 ± 2.1	59.4 ± 1.0	66.1 ± 1.7	64.7 ± 1.5	0.65 ± 0.01
	Val	34.4 ± 6.4	85.5 ± 1.5	60.3 ± 4.9	66.5 ± 6.9	65.1 ± 5.5	0.65 ± 0.03
MSE$_{aud}$+ISI1	Train	48.5 ± 8.4	78.9 ± 5.8	60.5 ± 2.2	70.3 ± 2.4	67.0 ± 1.8	0.70 ± 0.01
	Val	48.1 ± 9.9	76.2 ± 8.3	57.6 ± 9.2	69.2 ± 7.0	64.8 ± 4.2	0.69 ± 0.05
MSE$_{aud}$+ISI2	Train	48.5 ± 8.4	78.9 ± 5.8	60.5 ± 2.2	70.3 ± 2.4	67.0 ± 1.8	0.70 ± 0.01
	Val	48.1 ± 9.9	76.2 ± 8.3	57.6 ± 9.2	69.2 ± 7.0	64.8 ± 4.2	0.69 ± 0.05
MSE$_{aud}$+ISI1+ISI2	Train	52.8 ± 4.9	76.4 ± 3.4	59.5 ± 2.1	71.3 ± 2.3	67.1 ± 2.0	0.70 ± 0.02
	Val	52.0 ± 8.2	75.2 ± 5.2	57.7 ± 7.5	70.4 ± 7.5	65.7 ± 4.9	0.70 ± 0.05
MSE$_{aud}$+ISI1+ISI2+demos	Train	61.9 ± 1.0	78.1 ± 2.3	64.9 ± 1.6	75.8 ± 1.1	71.7 ± 1.2	0.78 ± 0.01
	Val	60.3 ± 8.3	76.4 ± 3.1	63.4 ± 7.5	73.4 ± 7.1	69.5 ± 3.9	0.77 ± 0.04
MSE$_{aud}$+ISI1+ISI2+ODI	Train	63.6 ± 3.2	89.6 ± 3.9	80.4 ± 5.7	79.0 ± 1.5	79.3 ± 2.1	0.88 ± 0.01
	Val	65.3 ± 6.6	88.4 ± 4.8	78.6 ± 8.7	79.4 ± 6.4	79.1 ± 4.5	0.88 ± 0.04
MSE$_{aud}$+ISI1+ISI2+demos+ODI	Train	68.3 ± 1.8	89.6 ± 1.6	81.2 ± 2.0	81.2 ± 1.0	81.2 ± 0.9	0.88 ± 0.01
	Val	68.5 ± 4.6	89.3 ± 3.5	81.3 ± 7.6	80.1 ± 6.1	80.6 ± 3.6	0.88 ± 0.02

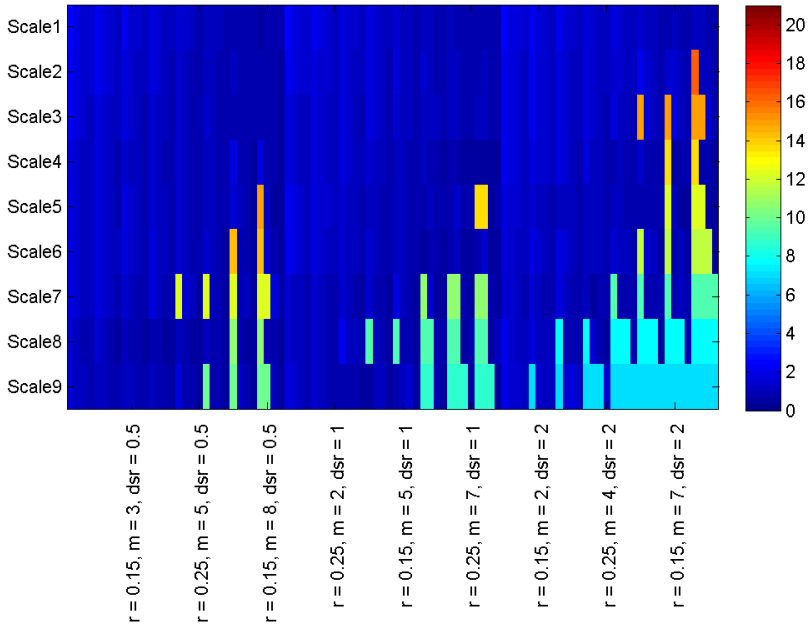
Table 5.8: *The average performance statistics over all five folds (mean±σ) for different audio feature combinations on the validation data set when using a RF to classify subjects.*

Features	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{aud}$	65.1 ± 7.1	89.0 ± 2.1	79.0 ± 6.0	79.8 ± 4.0	79.7 ± 3.5	0.86 ± 0.03
ISI1	47.7 ± 5.9	82.7 ± 6.2	64.7 ± 10.2	70.6 ± 5.7	68.5 ± 3.2	0.72 ± 0.03
ISI2	48.5 ± 4.1	82.3 ± 6.3	64.6 ± 10.4	70.8 ± 5.4	68.6 ± 3.1	0.72 ± 0.03
MSE$_{aud}$+ISI1	66.9 ± 5.1	87.6 ± 3.3	77.4 ± 8.5	80.2 ± 4.4	79.4 ± 3.6	0.87 ± 0.03
MSE$_{aud}$+ISI2	66.2 ± 5.1	87.3 ± 3.8	76.9 ± 8.7	79.7 ± 4.7	78.8 ± 3.0	0.87 ± 0.03
MSE$_{aud}$+ISI1+ISI2	65.4 ± 4.2	86.9 ± 3.3	76.0 ± 8.8	79.3 ± 3.9	78.3 ± 3.1	0.86 ± 0.03
MSE$_{aud}$+ISI1+ISI2+demos	71.8 ± 5.6	87.8 ± 3.2	79.1 ± 5.9	82.5 ± 5.4	81.5 ± 4.4	0.88 ± 0.04
MSE$_{aud}$+ISI1+ISI2+ODI	84.6 ± 3.0	91.2 ± 1.5	85.9 ± 4.1	89.8 ± 3.4	88.5 ± 1.0	0.96 ± 0.01
MSE$_{aud}$+ISI1+ISI2+demos+ODI	84.6 ± 4.1	91.3 ± 2.0	86.3 ± 3.1	89.9 ± 3.7	88.6 ± 1.8	0.96 ± 0.01

fold, $dsr = 0.5, m = 1, r = 0.15$ was chosen twice. Overall, approximately half of the best combinations were for $dsr = 0.5, m = 1$ with $r = 0.1 : 0.05 : 0.25$: 60% for the first fold, 100% for the second fold, 60% for the third fold, 20% for the fourth fold and 60% for the fifth fold. Although this is not consistent with the best combinations chosen by LDA, it is clear from Figure 5.7 that the two classifiers are picking up on differences between the MSE values, just at different parameter combinations. The average results over the five folds when using LDA can be found in Table 5.7; and the RF results can be found in Table 5.8. The best performance in both tables is highlighted in blue. Boxplots of the predictions for all combinations of parameters can be found in Figures 5.9 and 5.11, while the ROC curves for each combination of parameters can be found in Figures 5.10 and 5.12.

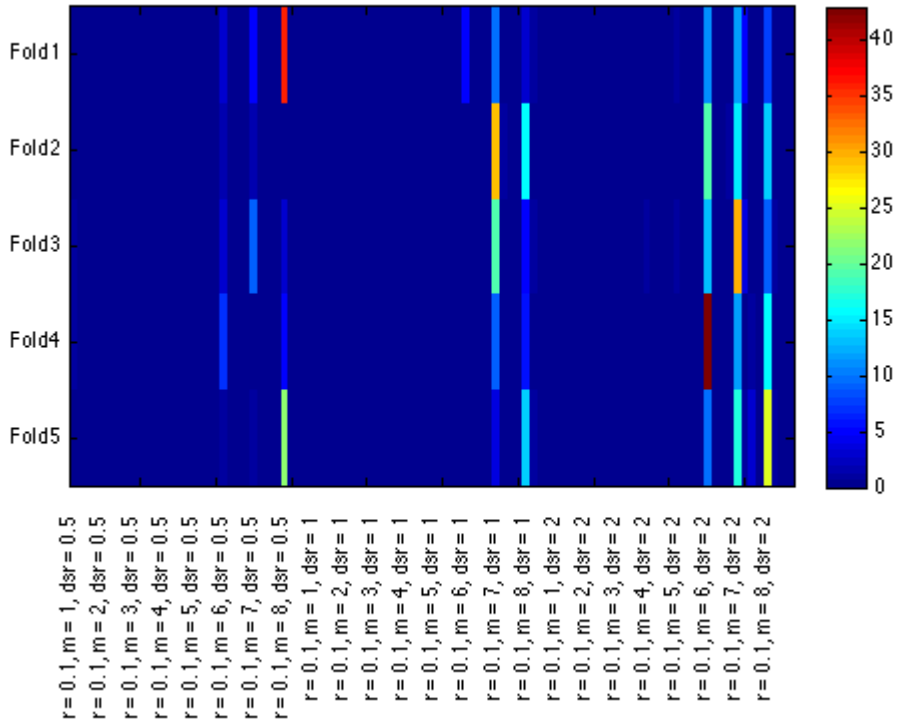


(a) *Snorer*

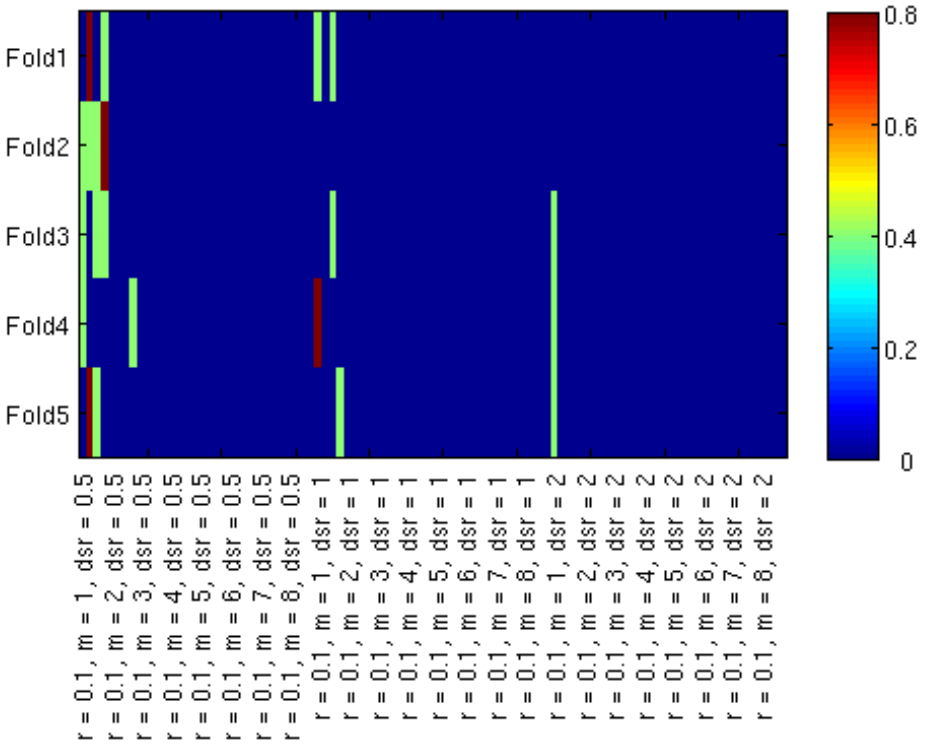


(b) *Severe apnoeic*

Figure 5.7: *Heatmaps illustrating the differences in MSE values at different scale factors for all combinations of m, r, dsr for a snoring subject and a severe apnoeic.*

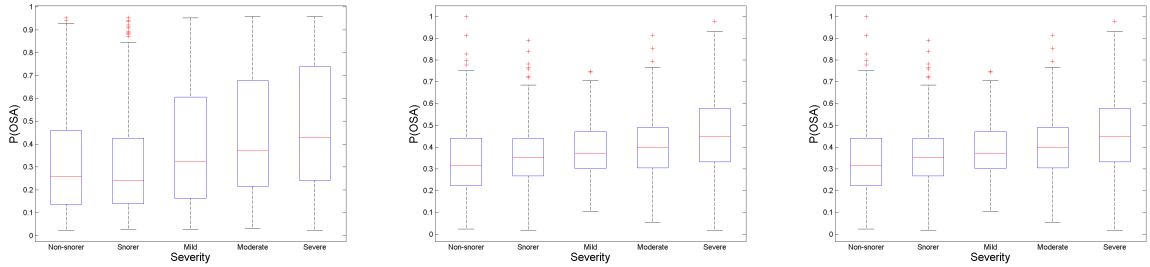


(a) *LDA*

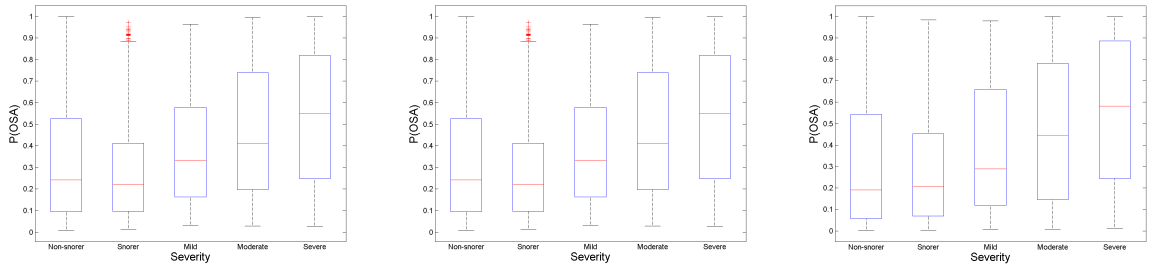


(b) *RF*

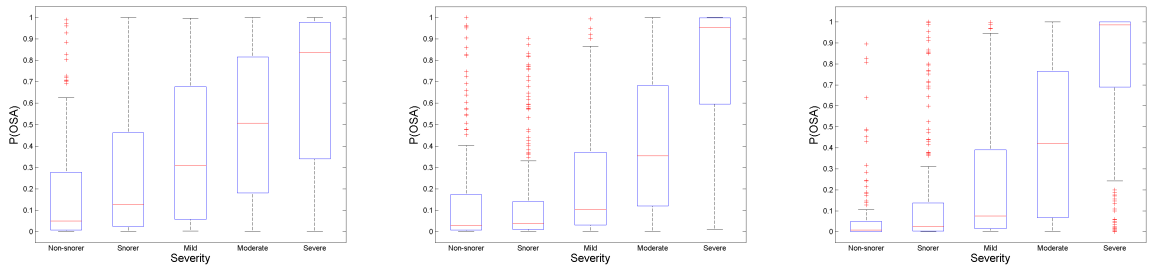
Figure 5.8: *Heatmaps illustrating how often MSE parameter combinations using audio data were chosen per fold for the different classifiers.*



(a) MSE_{aud} **(b)** *ISI1* **(c)** *ISI2*

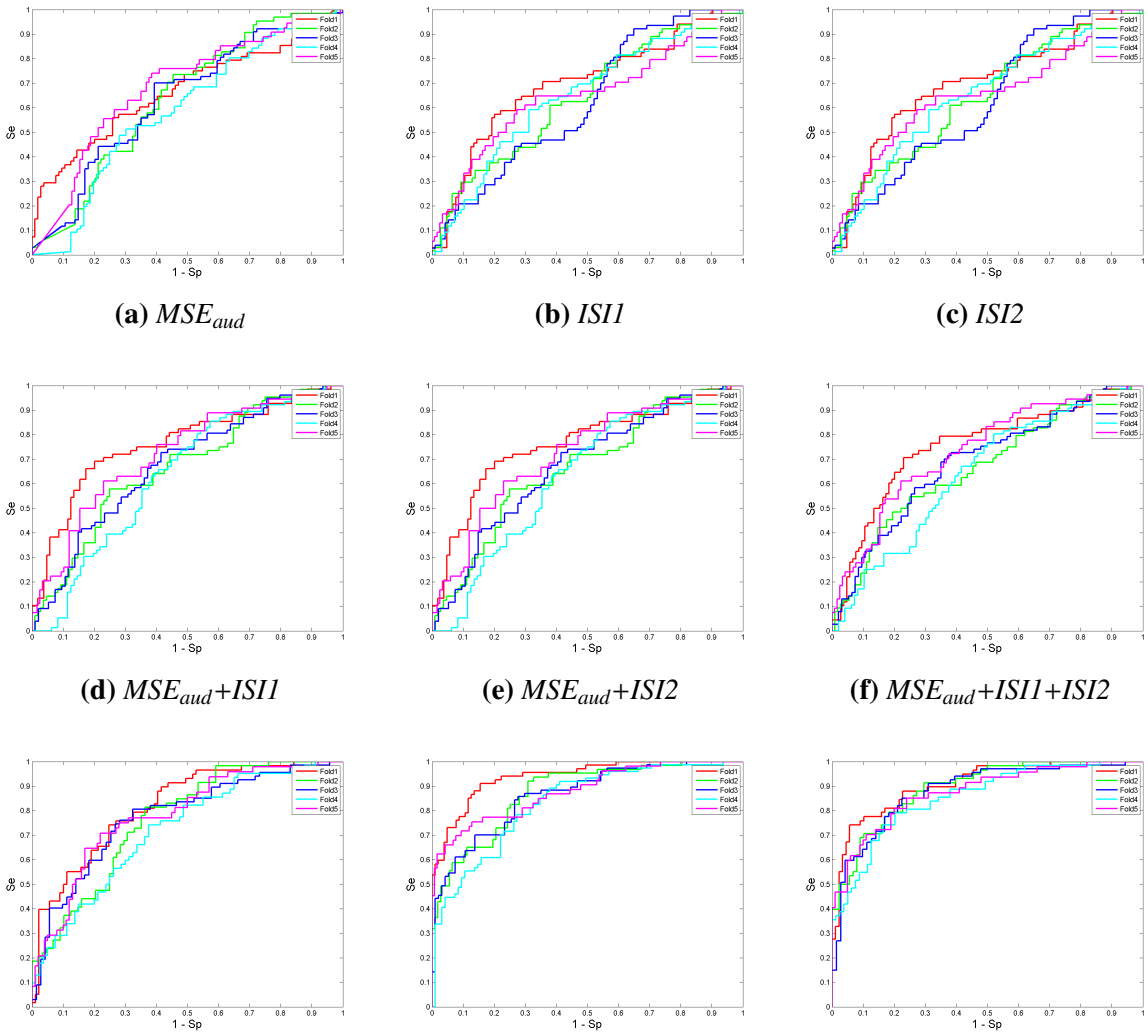


(d) MSE_{aud}+*ISI1* **(e)** MSE_{aud}+*ISI2* **(f)** MSE_{aud}+*ISI1*+*ISI2*



(g) MSE_{aud}+*ISI*+*demos* **(h)** MSE_{aud}+*ISI*+*ODI* **(i)** MSE_{aud}+*ISI*+*demos*+*ODI*

Figure 5.9: *Boxplots of the LDA predictions on the validation data over all five folds for the different audio feature combinations for the audio analysis.*

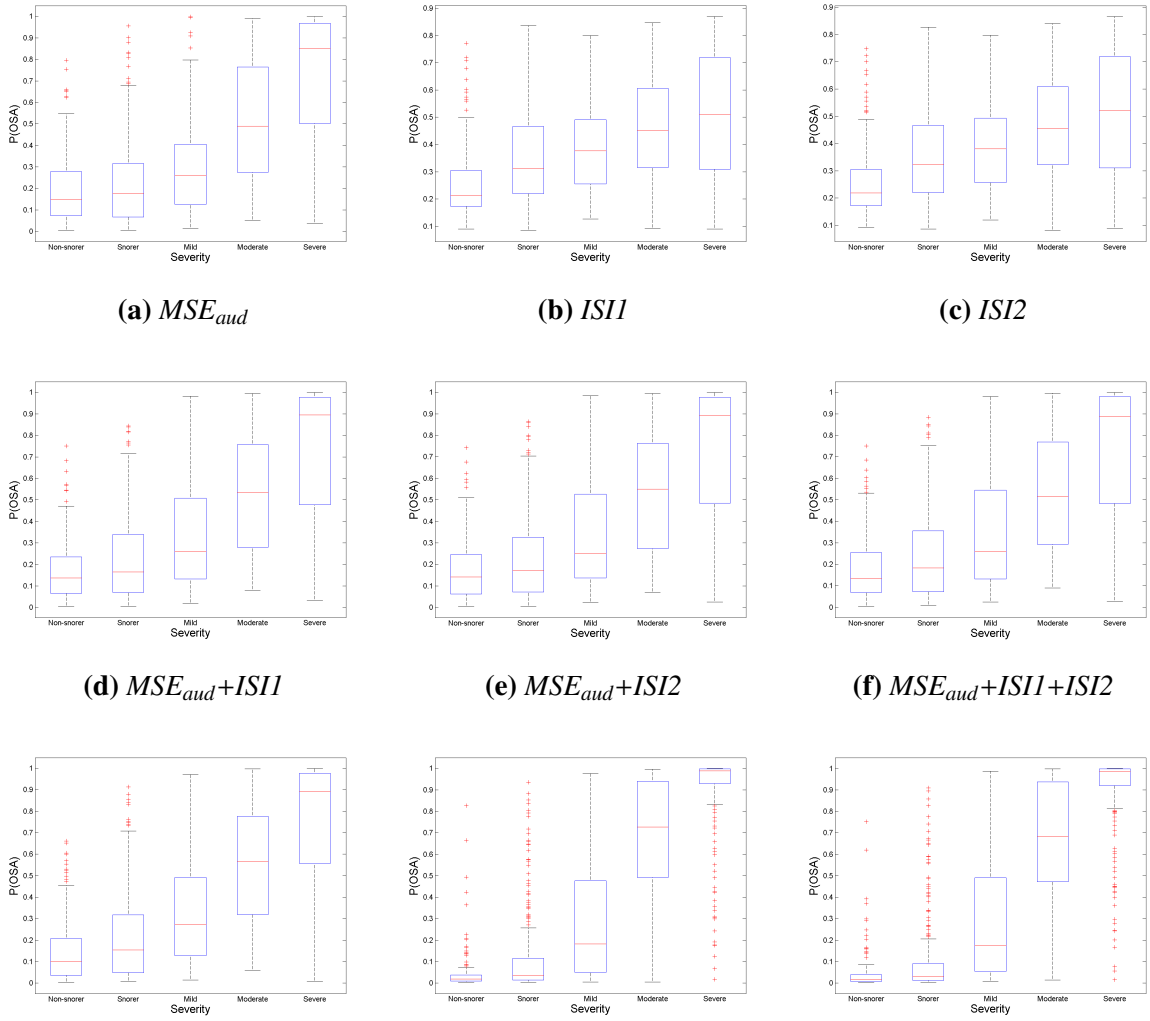


(a) MSE_{aud} **(b)** *ISI1* **(c)** *ISI2*

(d) MSE_{aud}+*ISI1* **(e)** MSE_{aud}+*ISI2* **(f)** MSE_{aud}+*ISI1*+*ISI2*

(g) MSE_{aud}+*ISI*+*demos* **(h)** MSE_{aud}+*ISI*+*ODI* **(i)** MSE_{aud}+*ISI*+*demos*+*ODI*

Figure 5.10: *ROC curves of the LDA predictions on the validation data over all five folds for the different audio feature combinations for the audio analysis. Red = fold 1, green = fold 2, blue = fold 3, cyan = fold 4, magenta = fold 5.*

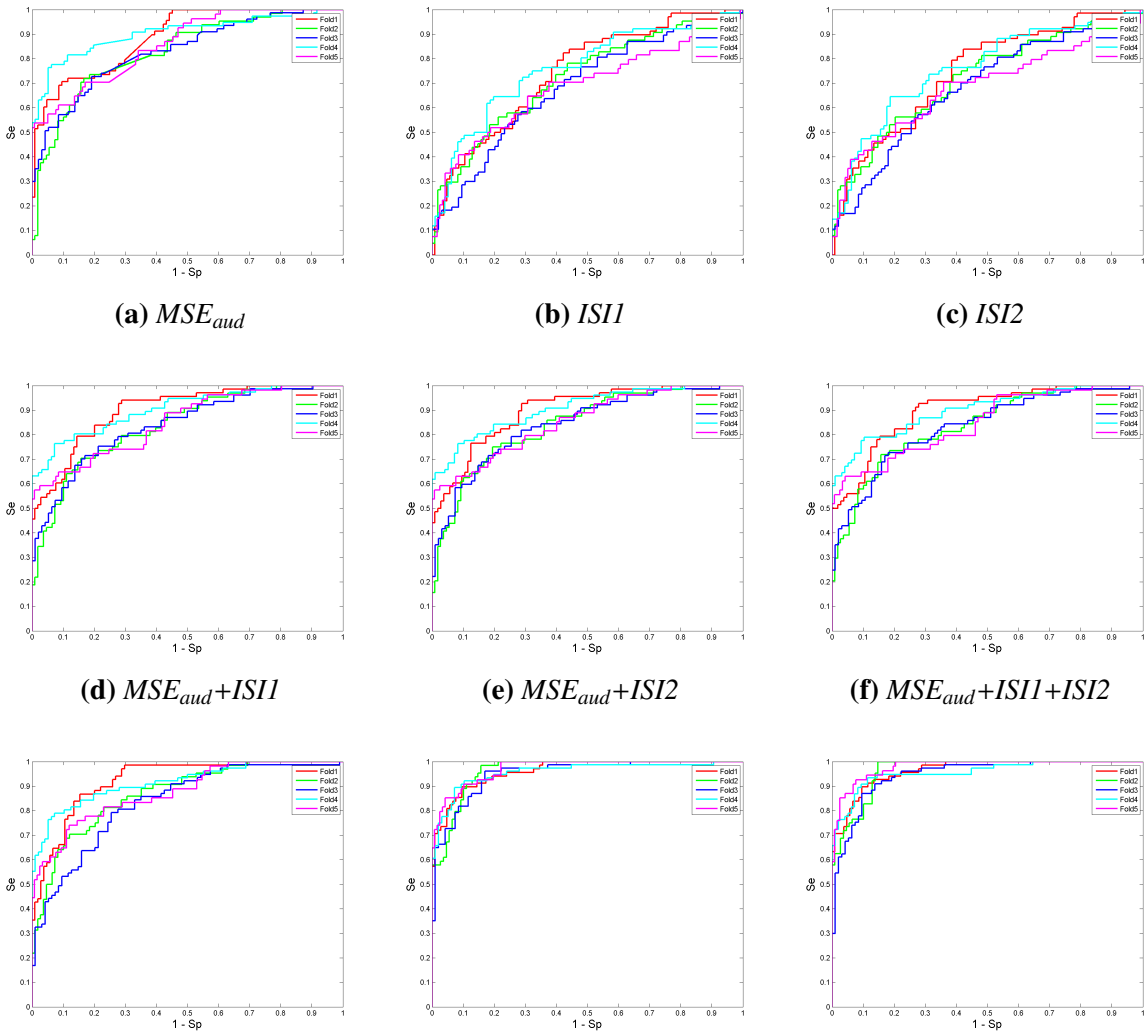


(a) MSE_{aud} **(b)** *ISI1* **(c)** *ISI2*

(d) MSE_{aud}*+ISI1* **(e)** MSE_{aud}*+ISI2* **(f)** MSE_{aud}*+ISI1+ISI2*

(g) MSE_{aud}*+ISI+demos* **(h)** MSE_{aud}*+ISI+ODI* **(i)** MSE_{aud}*+ISI+demos+ODI*

Figure 5.11: *Boxplots of the RF predictions on the validation data over all five folds for the different audio feature combinations for the audio analysis.*



(a) MSE_{aud} **(b)** *ISI1* **(c)** *ISI2*

(d) MSE_{aud}+*ISI1* **(e)** MSE_{aud}+*ISI2* **(f)** MSE_{aud}+*ISI1*+*ISI2*

(g) MSE_{aud}+*ISI*+*demos* **(h)** MSE_{aud}+*ISI*+*ODI* **(i)** MSE_{aud}+*ISI*+*demos*+*ODI*

Figure 5.12: *ROC curves of the RF predictions on the validation data over all five folds for the different audio feature combinations for the audio analysis. Red = fold 1, green = fold 2, blue = fold 3, cyan = fold 4, magenta = fold 5.*

5.7 Discussion

From Tables 5.2 and 5.6 it is clear that combining ODI with any other demographic actually worsens performance. In fact accuracy drops from 88.7% to 82.3%. This means that the baseline to beat is ODI on its own; or 88.7% accuracy.

From Figure 5.7 it is clear that MSE is capturing different information regarding the patterns in the data. Figure 5.7a shows that the MSE values for most of the scale factors and combinations of m, r, dsr are low, while Figure 5.7b shows that the MSE values are higher for many of the combinations. Visual examination of these plots indicate that MSE should be able to discriminate between the two classes at a number of the parameter combinations.

The search for the best dsr, m, r combination is consistent across all five folds for each classifier, which can be clearly seen in Figure 5.8. Based on the results, for LDA $r = 0.1$ and $m = 7$ or $m = 8$ is the optimal choice. For the RF, $dsr = 0.5, m = 1$ are the optimal choice, with all values of r suitable options. The m and r values are within the ranges specified by Costa *et al.* [166] for physiological data.

As can be seen from Tables 5.7 and 5.8, the best performance is obtained when $MSE_{aud} + ISI1 + ISI2 + ODI + demos$ are used in the analysis. LDA achieved Ac = 81.2% in training and Ac = 80.6% during validation. The RF achieved Ac = 88.6% during validation. Figures 5.9i and 5.11i both show that there is good separation between the two classes using this parameter combination.

For LDA, MSE is more specific (78.5%) than sensitive (41.1%). The results are better when using the RF, where both sensitivity and specificity improve (Se = 65.1% and Sp = 89.0%). The accuracy of MSE when using LDA is low (66.2%) however the RF performs better with Ac = 79.7%. It is clear that for audio MSE features, the RF is the better classifier. Figures 5.9a and 5.11a both show that, although there is separation between the two classes in terms of the median, the 75^{th} percentile of the mild OSA cases overlaps with the 25^{th} percentile of the moderate OSA cases. This is more pronounced in Figure 5.9a, and can also be seen in Figures 5.10a and 5.12a. The mild cases are the grey area where classification is most difficult. This is in keeping with diagnosis by clinicians. Mild cases, as standard, are not offered treatment, however when there is a high ESS for example, the subject is brought in to the hospital to talk with the clinician. A more detailed history is taken and depending on that

conversation, some subjects are offered treatment while others are not. It is heavily subject dependent.

There is no difference in the performance of ISI1 and ISI2 for LDA, and a slight difference for the RF. This indicates that the basic method (ISI1) is sufficient for calculating the ISI metrics (D_f, D_s, D_m, D_h). ISI2 takes much longer to compute and so ISI1 can be used to reduce processing time while achieving the same performance. For both classifiers ISI is more specific (85.2%, 82%) than it is sensitive (34.4%, 48%). Using LDA, the accuracy is better than for MSE (65.1%) but is worse (68.5%) when using a RF.

When combining features, including ODI consistently improves performance for both classifiers. Demographics also improve performance over that achieved by MSE and ISI. Both of these improvements can be seen in Figures 5.10, 5.10, 5.11 and 5.12. However, when all features are combined ($MSE + ISI + demos + ODI$) the best performance is achieved. For LDA this is 81.2% in training and 80.6% during validation, which has improved from 65.7% accuracy for $MSE + ISI$. The RF achieves 88.6% accuracy during validation, again improving from 78.3% accuracy for $MSE + ISI$. However, in situations when ODI is not available, demographics can be used, in combination with $MSE + ISI$ to give an accuracy of 81.5% when a RF is used.

These results are promising, and have improved greatly over the standard approaches to OSA diagnosis used in the previous chapter. The RF consistently achieves accuracy in the high-70 to high-80% range. The results have surpassed the performance of the AHI on its own (Ac = 86.4%), and are approximately the same as the performance of the ODI on its own (Ac = 88.1%) (see Table 5.2). This approach could potentially yield similar results when applied to other signals. The next step is to see how well a data combination approach works when MSE and ISI are applied to both audio and actigraphy data.

Chapter 6

Combining audio and actigraphy data

6.1 Introduction

It is possible that actigraphy could be used to improve performance over audio features alone. As can be seen in Chapter 5 audio features achieve reasonable results (Ac = 78.3%), and when combined with the ODI accuracy improves to 88.5%. However, it is important, for clinical practice, to achieve accuracy levels of 90% or higher [55] (see section 2.5). The accelerometer is another on-board smartphone sensor, from which actigraphy can be derived. Actigraphy is commonly used in sleep monitoring for sleep-wake identification [120]. It should be noted that actigraphy's use as a diagnostic tool for OSA has had variable success [127, 128, 132] (see section 3.3). Regardless, the Grey Flash device records actigraphy and so this can be used in the analysis. Ideally, the approach taken for audio would provide similar results when applied to actigraphy; and when the audio and actigraphy features are combined, performance could be improved.

This chapter investigates the application of MSE and ISI to actigraphy first and then combines audio, actigraphy, demographics, and ODI in order to discover the most predictive feature set for treatment vs. non-treatment.

6.2 Data

The data used in this chapter are the same as those in Chapter 5. See Table 5.4 for information on the class demographics, and Table 4.2 for the overall subject demographics. It is important

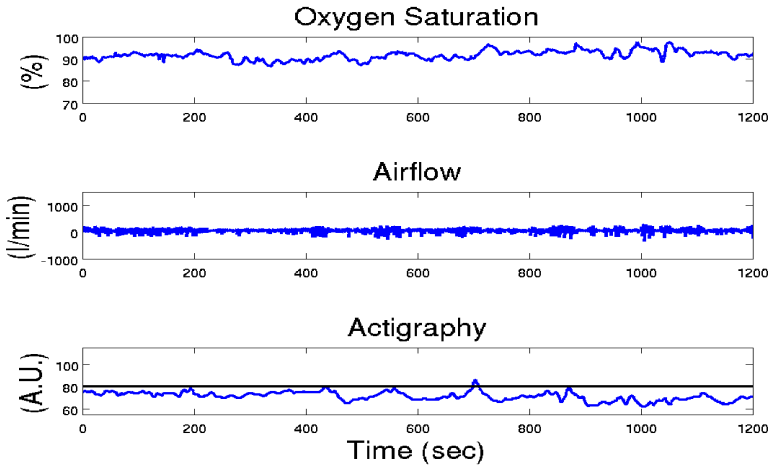
to note that the raw actigraphy signal is not available from the Grey Flash device, and that the signal that is available has undergone a lot of preprocessing. The exact process that is carried out is proprietary, but there is some form of averaging over the past x samples to compute the current sample. This is not ideal, but is the only movement signal available. In addition, the subjects are asked by the hospital to wear the box containing the accelerometer (see Figure 4.1) on their upper arm. The Grey Flash device has been designed to be worn on the chest not the upper arm, and so it is not as stable as if it was worn on the chest. By being less stable, the device moves more frequently, and movements not associated with OSA could be exaggerated. This could also be a factor in the quality and reliability of the signal.

6.3 Methods

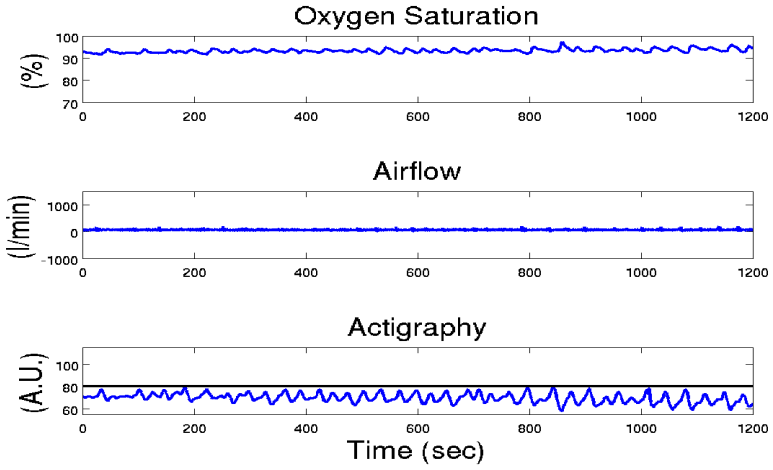
The methods used in this section of the analysis are actually the same as those used in Chapter 5. MSE is applied to the audio and actigraphy data; both ISI methods were also used on the audio and actigraphy data; the demographics did not undergo any preprocessing. All features from audio, actigraphy and demographics were classified using a RF, described in section 5.4. Figure 6.1 shows the SpO_2, airflow and actigraphy signals for three subjects; a normal subject, a snorer and a subject with severe OSA. On the actigraphy channel for each subject there is a black line: data above this line would be taken as the motion equivalent of snoring peaks. Instead of being peaks in the audio signal, there are peaks in the motion signal where a micro-arousal has occurred following an apnoea resulting in some physical activity [123]. It is clear that there are two very different patterns, one for the snorer and normal subjects and another for the apnoeic. However this is an unusual case; most of the severe apnoeic actigraphy data actually looks similar to the normal/snorer data, which can be seen in Figure 6.2. This figure clearly shows that for the apnoeic subject, the actigraphy does not reach the line used to identify peaks. This results in the four metrics for the apnoeic subject being similar to those of normal/snoring subjects.

6.4 Data Analysis Protocol

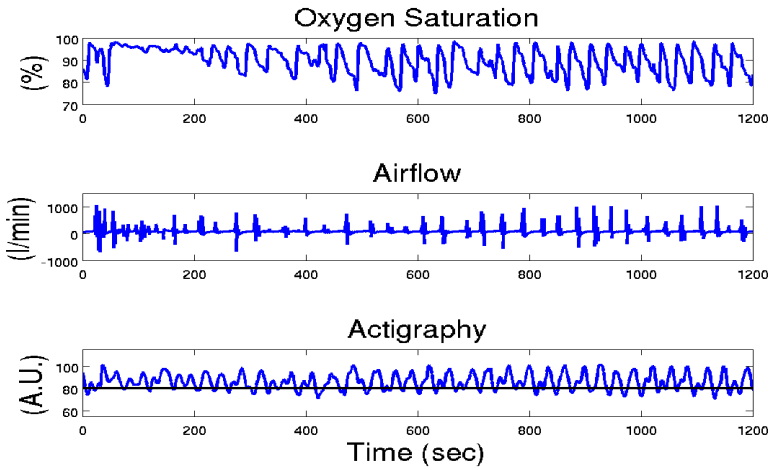
The procedure used to calculate the features has some overlap with section 5.5 in Chapter 5.



(a) *Normal subject*



(b) *Snoring subject*



(c) *Subject with severe OSA*

Figure 6.1: *Three examples of Grey Flash data; a normal subject, a snoring subject and a subject with severe OSA. The black line on each of the actigraphy channels is used to represent the peaks that the ISI captures; data points above this line would be taken to be peaks and used to calculate the four metrics. The actigraphy signal for the snorer and normal subject have similar patterns to each other, whereas the actigraphy signal for the subject with severe OSA has a completely different pattern. A.U. = arbitrary units.*

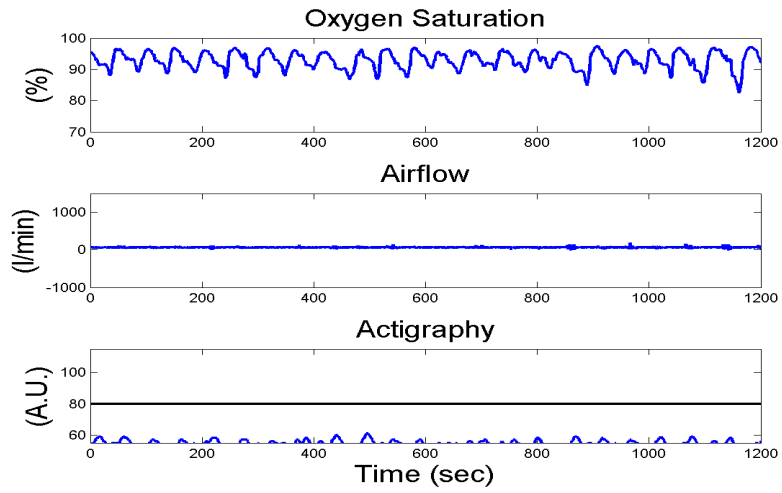


Figure 6.2: *Example of Grey Flash data for a subject with severe OSA. The black line on the actigraphy channel is used to represent the peaks that the ISI captures; data points above this line would be taken to be peaks and used to calculate the four metrics. The actigraphy signal for this subject has a similar pattern to that of a normal/snoring subject (which can be seen in Figure 6.1), which is the case for most of the subjects in the dataset. A.U. = arbitrary units.*

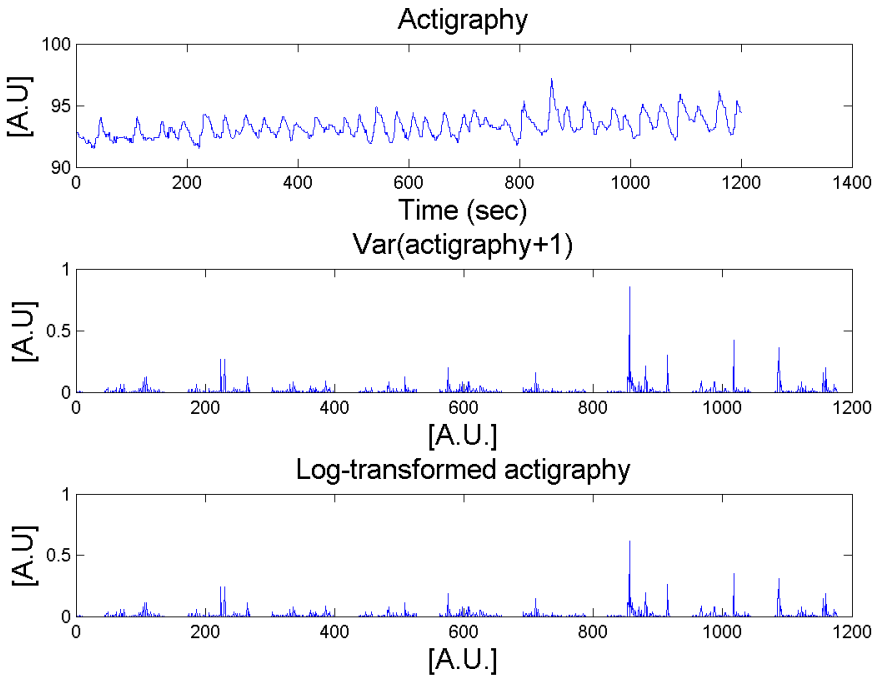
6.4.1 Audio

The same audio features that were calculated as per section 5.5 were used in this part of the analysis.

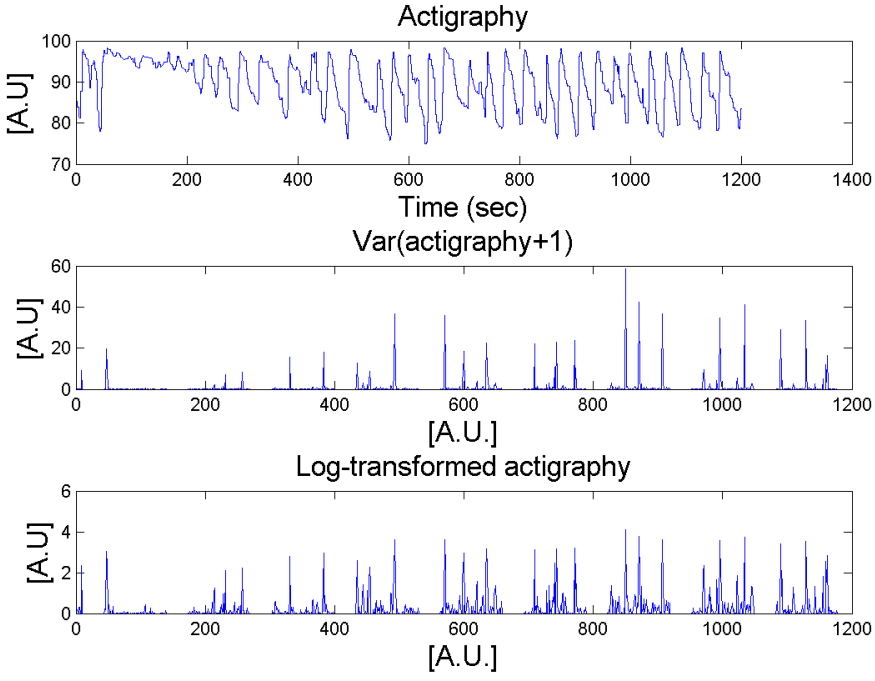
6.4.2 Actigraphy

The actigraphy data were analysed in a similar manner to the audio data. For MSE the same $240min$ that were used in the MSE calculation of the audio MSE coefficients were used, beginning $30min$ into the recording and ending at $4.5h$. The data were preprocessed by taking the variance every 0.5, 1 or $2s$ and then the natural logarithm of that time series plus one was taken. As can be seen in Figure 6.3, this process highlighted the peaks in the signal. Nine MSE coefficients were calculated per subjects ($\tau = 1, 2, 4, 8, 16, 32, 65, 130, 180$) for $m = 1 : 1 : 8$ and $r = 0.1 : 0.05 : 0.25$, which are the same parameter values as used in calculating MSE for audio data.

ISI1 was calculated in the same way as for the audio data: the actigraphy data were downsampled to $1Hz$, normalised over the entire night and the peaks were found. Similarly, ISI2 was calculated the same as for audio. In both cases, the same four metrics (D_f, D_s, D_m, D_h) were calculated.



(a) *Snorer*



(b) *Apnoeic*

Figure 6.3: *The actigraphy signal, the downsampled actigraphy signal, and the downsampled and* ln*-transformed actigraphy signal for a snorer and an apnoeic subject. A.U. = arbitrary units.*

6.4.3 Classification

The same classification procedure was used for actigraphy as for audio. Five-fold cross-validation was carried out on the data where the folds are the same as those used in the audio analysis. Each time, one fold was held separately to be the validation set (X_{val}) while the other four folds were used as the design data set. X_{design} was further divided into training and test data sets (70% and 30%) n times in order to find the best MSE downsampling rate (dsr), m value and r value. This was done by using the RF to classify every possible combination of dsr, m, r, carried out 5 times. The highest accuracy was noted along with the dsr, m, r combination that it corresponded to. The best overall combination was taken to be the one that was chosen most often in the n iterations.

When the best dsr, m, r combination had been found using X_{design}, the full four folds were used to train the classifiers and the untouched fold was used for validation. Both classifiers were used to test all combinations of features, namely: MSE, $ISI1$, $ISI2$, $MSE + ISI1$, $MSE + ISI2$, $MSE + ISI1 + ISI2$, $MSE + ISI1 + ISI2 + demographics$, $MSE + ISI1 + ISI2 + ODI$, and $MSE + ISI1 + ISI2 + demographics + ODI$. In each case, the MSE coefficients that were used were those found in the first stage, *i.e.* the MSE coefficients that corresponded to the dsr, m, r combination that gave the highest accuracy for both classifiers. The results were noted for each feature combination, classifier and fold. For the RF, 500 trees were used in the forest. Each tree split on three variables/features. The process was repeated twice with a new seed for 2×10^6 iterations.

The next classification stage involved combining the audio and actigraphy features. This was done using the best dsr, m, r combination per fold for both audio and actigraphy. These 18 features (9 audio MSE coefficients and 9 actigraphy MSE coefficients) were used as the combined MSE feature vector. Similarly, the ISI features for both audio and actigraphy were used as in the combined feature vector. It was then possible to test all combinations of combined features as described above. Again, the exact same folds that were used in the audio and actigraphy analysis were used for the combined features analysis.

Table 6.1: *The average performance statistics over all five folds (mean±σ) for different actigraphy feature combinations on the validation data set when using a RF to classify subjects.*

Features	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{act}$	6.9 ± 5.0	94.2 ± 4.5	43.6 ± 13.8	60.8 ± 5.7	59.4 ± 4.6	0.53 ± 0.03
ISI1	0.0 ± 0.0	99.4 ± 0.9	0.0 ± 0.0	60.3 ± 5.6	60.1 ± 5.8	0.49 ± 0.02
ISI2	0.6 ± 1.3	99.6 ± 0.9	50.0 ± 70.7	60.5 ± 5.6	60.5 ± 5.8	0.49 ± 0.03
MSE$_{act}$+ISI1	6.5 ± 2.9	94.6 ± 1.4	41.8 ± 14.3	60.8 ± 5.5	59.8 ± 4.9	0.53 ± 0.04
MSE$_{act}$+ISI2	5.5 ± 1.2	95.0 ± 1.2	42.5 ± 14.3	60.6 ± 5.1	59.7 ± 4.3	0.54 ± 0.04
MSE$_{act}$+ISI1+ISI2	4.1 ± 2.7	95.0 ± 2.2	32.1 ± 18.5	60.3 ± 5.3	59.3 ± 5.2	0.53 ± 0.04
MSE$_{act}$+ISI1+ISI2+demos	5.3 ± 1.7	95.4 ± 1.5	43.6 ± 14.2	60.7 ± 5.4	59.8 ± 4.8	0.51 ± 0.03
MSE$_{act}$+ISI1+ISI2+ODI	7.3 ± 3.2	95.0 ± 2.8	51.0 ± 21.2	61.1 ± 5.5	60.1 ± 4.5	0.54 ± 0.02
MSE$_{act}$+ISI1+ISI2+demos+ODI	6.7 ± 2.1	94.4 ± 4.4	49.0 ± 17.2	60.7 ± 5.4	59.6 ± 4.8	0.53 ± 0.02

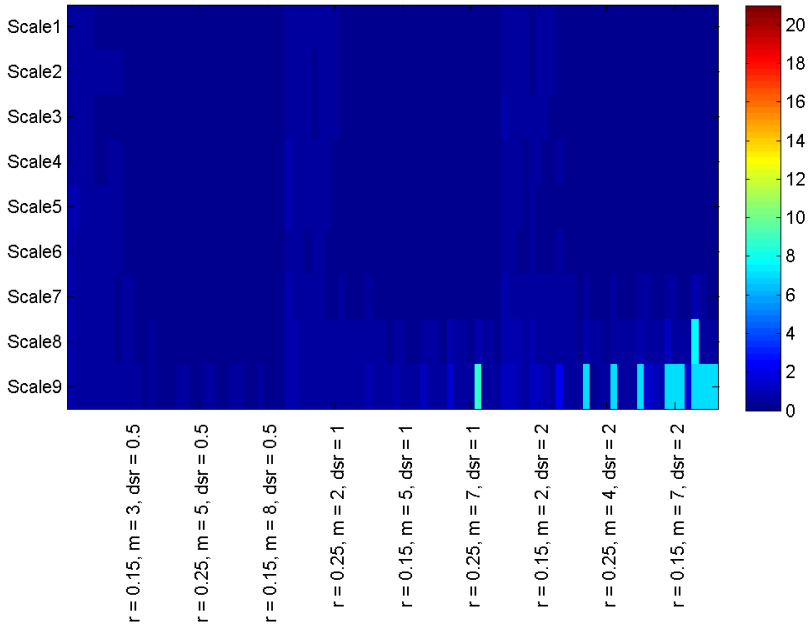
6.5 Results

The results are divided into two sections: the actigraphy analysis and the combined audio and actigraphy analysis.

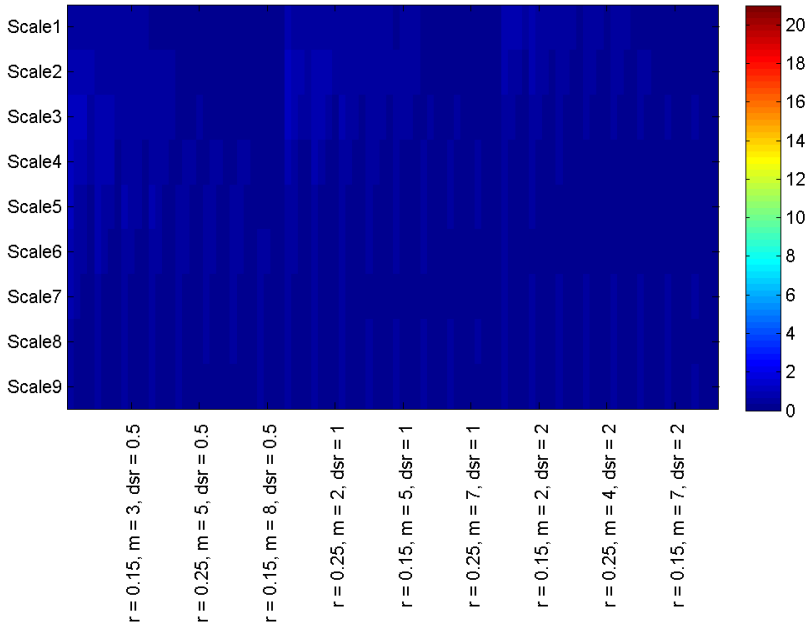
6.5.1 Actigraphy Results

Figure 6.4 shows the MSE values at the different scale factors (τ) for all combinations of m, r, dsr for a snoring subject and a severe apnoeic. This figure clearly shows that there are only minor differences in the MSE value between the two subjects. Figure 6.5 shows the number of times each MSE parameter combination was chosen per fold for both classifiers. It is clear that there is little consistency across the five folds, indicating that the best parameter combination for MSE of actigraphy is dependent on the data set, unlike MSE of audio. Figure 6.5 shows that when using RFs to choose the best parameter combination the results were: for the first fold $dsr = 0.5, m = 1, r = 0.15$ which was chosen once. For the second fold, the best parameter combination was $dsr = 0.5, m = 5, r = 0.15$, again chosen once. In the third fold, the best parameter combination, chosen once, was $dsr = 0.5, m = 2, r = 0.15$ while in the fourth fold, the best combination chosen once, was $dsr = 0.5, m = 3, r = 0.1$. In the fifth fold, the combination chosen once was $dsr = 0.5, m = 7, r = 0.1$.

The average results over the five folds can be found in Table 6.1. The best results have been highlighted in blue.

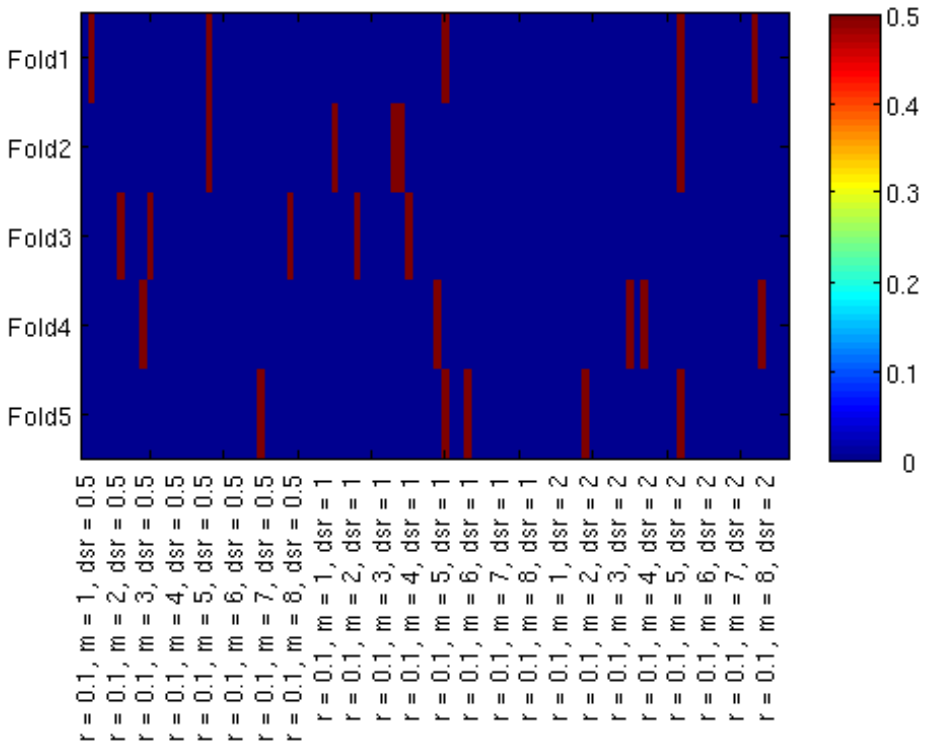


(a) *Snorer*



(b) *Severe apnoeic*

Figure 6.4: *Heatmaps illustrating the differences in MSE values at different scale factors for all combinations of m, r, dsr for a snoring subject and a severe apnoeic.*



(a) *RF*

Figure 6.5: *Heatmap illustrating how often MSE parameter combinations for actigraphy were chosen per fold for the RF.*

Table 6.2: *The average performance statistics over all five folds (mean±σ) for different audio and actigraphy feature combinations on the validation data set when using a RF to classify subjects.*

Features	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{aud+act}$	63.9 ± 7.1	89.4 ± 3.1	79.4 ± 6.6	79.3 ± 4.3	79.5 ± 4.0	0.86 ± 0.04
ISI1	47.7 ± 6.1	81.4 ± 4.5	62.7 ± 7.5	70.3 ± 6.4	67.8 ± 3.6	0.71 ± 0.03
ISI2	49.2 ± 6.4	81.4 ± 4.2	63.3 ± 7.6	70.9 ± 6.5	68.4 ± 3.7	0.71 ± 0.03
MSE$_{aud+act}$+ISI1	66.3 ± 5.5	86.7 ± 1.3	76.1 ± 5.4	79.7 ± 4.4	78.7 ± 2.7	0.87 ± 0.03
MSE$_{aud+act}$+ISI2	67.2 ± 5.6	87.0 ± 2.4	76.7 ± 6.6	80.2 ± 4.5	79.1 ± 2.9	0.87 ± 0.03
MSE$_{aud+act}$+ISI1+ISI2	66.0 ± 4.7	86.8 ± 1.4	76.0 ± 6.3	79.6 ± 3.9	78.6 ± 2.4	0.86 ± 0.03
MSE$_{aud+act}$+ISI1+ISI2+demos	65.7 ± 6.1	86.5 ± 1.7	75.7 ± 5.8	79.4 ± 4.4	78.3 ± 3.0	0.86 ± 0.02
MSE$_{aud+act}$+ISI1+ISI2+ODI	66.0 ± 4.9	86.1 ± 1.9	75.4 ± 5.1	79.5 ± 4.0	78.2 ± 1.5	0.86 ± 0.03
MSE$_{aud+act}$+ISI1+ISI2+demos+ODI	66.7 ± 3.9	87.3 ± 1.1	77.1 ± 5.1	80.0 ± 4.2	79.1 ± 2.3	0.86 ± 0.03

6.5.2 Combined Feature Results

Using the best MSE parameter combinations picked for audio and actigraphy per fold, the average results over the five folds can be found in Table 6.2, with the best results highlighted in blue.

6.6 Discussion

6.6.1 Discussion of Actigraphy Results

From Figure 6.4 it is clear that MSE is not capturing much information that is different between the two subject types. The MSE values for both the snorer (Figure 6.4a) and the apnoeic (Figure 6.4b) are low. The data used in these figures were taken from the same subject as in Figure 5.7. Heatmaps of the MSE values for the audio data showed clear differences between the two subjects, while there were only minor differences between the subjects for the actigraphy data.

The results of the search for the best dsr, m, r combination were very variable across the five folds which can be clearly seen in Figure 6.5; there was no consistency in any of the three parameters. With no consistency in the parameters chosen, it appears that MSE of actigraphy is heavily dependent on the data available for analysis, indicated by the range of combinations chosen per fold where different data is used in X_{design}. This is unlike MSE of audio using a RF, in which the value of dsr and m are consistent while the value of r is more variable, regardless of the data in X_{design}.

As can be seen in Table 6.1, actigraphy in not a good predictor for OSA. MSE is highly specific (94.2%) with very poor sensitivity (6.9%). For the RF, the best results obtained used

$MSE_{act} + ISI1 + ISI2 + ODI$, achieving Ac = 60.1% during validation.

There is no difference between the performance of ISI1 and ISI2. This is consistent with ISI for audio data, indicating that the basic method (ISI1) for calculating the features is sufficient. For the RF, performance is highly specific (99.5%) with very low sensitivity (0.3%) and an accuracy of 60.2% during validation. Including demographics in the analysis does not improve performance (Ac = 59.8%). In this case, including ODI in the analysis does not improve any performance metric.

These results are consistent with the literature and are definitely not good enough to use to classify subjects. It is possible that the preprocessing that is carried out internally by the Grey Flash device on the actigraphy data or the fact that the accelerometer is worn on the upper arm instead of the chest is causing the diminished performance. The preprocessing involves averaging over x samples to calculate the current sample. This could lead to small movements caused by micro-arousals being averaged out. Wearing the accelerometer on the arm instead of the chest could cause two different movement artefacts. In the first instance, by wearing the device on the upper arm, the accelerometer could be registering movements not related to arousals from apnoea but movements caused by the device slipping due to an ill-fitting connection. In the second case, wearing the accelerometer on the upper arm might lead to the device not registering movements that occur in the lower arm, legs or torso associated with arousal from apnoea. It is possible that the raw actigraphy signal could provide better classification accuracy, however, with that data unavailable it is not possible to confirm this hypothesis.

6.6.2 Discussion of Combined Feature Analysis

Table 6.2 shows that combining audio and actigraphy MSE features does not change the performance achieved when using audio MSE alone, achieving 79.5% accuracy during validation compared to 79.9%. The same can be said for combining audio and actigraphy ISI features; accuracy does not change (Ac = 68%). The best combined features were $MSE_{aud+act} + ISI + demos + ODI$, achieving an accuracy of 79.1%. However, this is worse than the accuracy achieved using the best features for the audio analysis ($MSE_{aud} + ISI + demos + ODI$) (88.6%) but better than that achieved using the best features for the actigraphy analysis ($MSE_{act} + ISI +$

ODI) (60.1%). It is clear that actigraphy on its own is not capable of reliably detecting OSA, and when combined with audio features, actually worsens the performance that was achieved when using audio features alone.

Given that the accelerometer data do not add value in identifying the subjects as requiring treatment or not, there is now a need to consider an external sensor. The obvious technology to use here is pulse oximetry, as it achieves good performance (see section 3.4) and can easily be recorded by a smartphone. The ODI used in the analyses thus far were computed by the Visi-Download software using proprietary algorithms, obviously based on SpO_2. It is possible that a new implementation of the ODI could improve performance by including a measure of signal quality to reject noisy segments of data. This technique has been investigated in the next chapter.

Chapter 7

PPG analysis

7.1 Introduction

PPG, from which SpO_2 is derived, is a very useful signal for diagnosing OSA, particularly for stratifying subjects' severity. As can be seen from Chapters 5 and 6, ODI on its own gives Ac = 88.7%; when combined with audio features, the accuracy remains at 88.6% with an AUC of 0.96. It is clear that using PPG, and ODI in particular, improves the classification accuracy to the high-80s. It is not clear whether signal quality is used in any way in the calculation of ODI as the Visi-Download software (from now on denoted ODI_{VISI}) is proprietary. This chapter looks at recalculating the ODI in two ways; with and without considering signal quality. Using signal quality of the PPG signal is not a new concept; Sukor *et al.* [176] used signal quality measures for pulse oximetry through waveform morphology analysis. The proposed algorithm for automatic rejection of artefact-contaminated pulse oximetry waveforms was compared with a manually annotated gold standard. Fingertip PPG signals were acquired from 13 healthy subjects (10 m, 3 f). Some unique waveform morphology features were extracted from the PPG signals; these were taken to be correlated with signal quality. A decision-tree classifier was used to arrive at a classification decision of whether to accept the pulse or not. The algorithm achieved a mean κ (Cohen's kappa coefficient) of 0.64±0.22 with Se = 89±10%, Sp = 77±19% and Ac = 83±11%. The authors concluded that it is possible to achieve automatic identification of signal artefact in the PPG signal through waveform morphology analysis.

The SpO_2 is not the only metric that can be derived from the raw PPG; it is also possible to

obtain pulse rate (PR). A considerable amount of work has been carried out applying MSE to HRV and there are distinct MSE curves for healthy, congestive heart failure and atrial fibrillation subjects, suggesting diagnostic use [168, 177]. OSA subjects have unusual HR dynamics over the course of the night, related to the sleep stages, so it is expected that HRV would be a useful feature. In particular, it is possible that MSE will reveal dynamics over multiple time scales as it picks up short term HRV changes as well as long term variations. In addition, MSE is robust to noise according to Costa *et al.* [177].

7.2 Data

The data used in this section of the analysis is the same as that used in the audio, actigraphy and combined analyses. See Table 5.4 for information on the class demographics, and Table 4.2 for the overall subject demographics.

7.3 Methods

7.3.1 Pulse Detection and Signal Quality

Pulse detection and signal quality were calculated using the method proposed by Li & Clifford [178]. Pulses are detected using wabp.c (an open source arterial blood pressure pulse detector [179][1]). Both time and amplitude thresholds are changed to fit PPG morphology (the slope width of the rising edge of the pulse is changed from 130 to 170ms; the eye-closing period[2] following each detected pulse is extended from 250 to 340ms in order to avoid the double detection of the possible secondary peak of a PPG pulse). The length of a PPG pulse is delimited by fiducial marks at the onset of the current pulse and the onset of the next pulse. If no pulse is found 3s after the onset of any given pulse, the end of the pulse window is truncated to 3s.

A PPG pulse template is generated by averaging every pulse in a window of 30s. PPG is assumed to be quasi-periodic, and so the autocorrelation of each 30s of data is taken and the length (L) between two main peaks of the autocorrelation sequence used to determine the

[1] From `www.physionet.org`

[2] A time period during which no pulse detection occurs

average period of PPG pulses. The length of the PPG template is then set to L. In order to derive the first template (T_1), all the pulses in the $30s$ window are averaged, with each pulse beginning at the fiducial mark (the onset of the pulse) and ending at the length of the template. The correlation coefficients (C) between T_1 and each pulse in the $30s$ window are calculated. Any pulse with $C < 0.8$ is removed from the template and the average pulse is recalculated from the remaining pulses to generate the second template (T_2). If more than half of the pulses are removed by this process T_2 is deemed untrustworthy, and the template from the previous window is used instead. If no previous window is available, the next $30s$ of data are used. Template updating can be performed on a pulse-by-pulse basis, but only after classification of a new incoming pulse is performed, which requires several other pulse analysis metrics first, as described below.

Three signal quality indices (SQIs) are defined as follows:

- Direct matching SQI: The sampling point series of each pulse within the $30s$ window is selected, beginning at the fiducial mark and ending at the length of the template (L). The correlation coefficient is calculated with the template giving the direct matching SQI (SQI_1).

- Linear resampling SQI: Each pulse between two fiducial marks is linearly stretched (if the length of the pulse is shorter than L) or compressed (if it is longer) to the length of the template. The correlation coefficient is calculated as the linear resampling SQI (SQI_2).

- Clipping detection SQI: Periods of saturation to a maximum or a minimum value are determined within each pulse. A hysteresis threshold (of 1 normalised unit) is defined to determine the smallest fluctuation that should be ignored. Such samples are defined as 'clipped'. The percentage of the pulse that is not clipped was defined to be the clipping detection SQI (SQI_3).

In SQI_1 and SQI_2, any negative values of the correlation coefficient (negative correlation) are set to zero, so the value of each SQI ranges between 0 and 1 inclusively.

7.4 Data Analysis Protocol

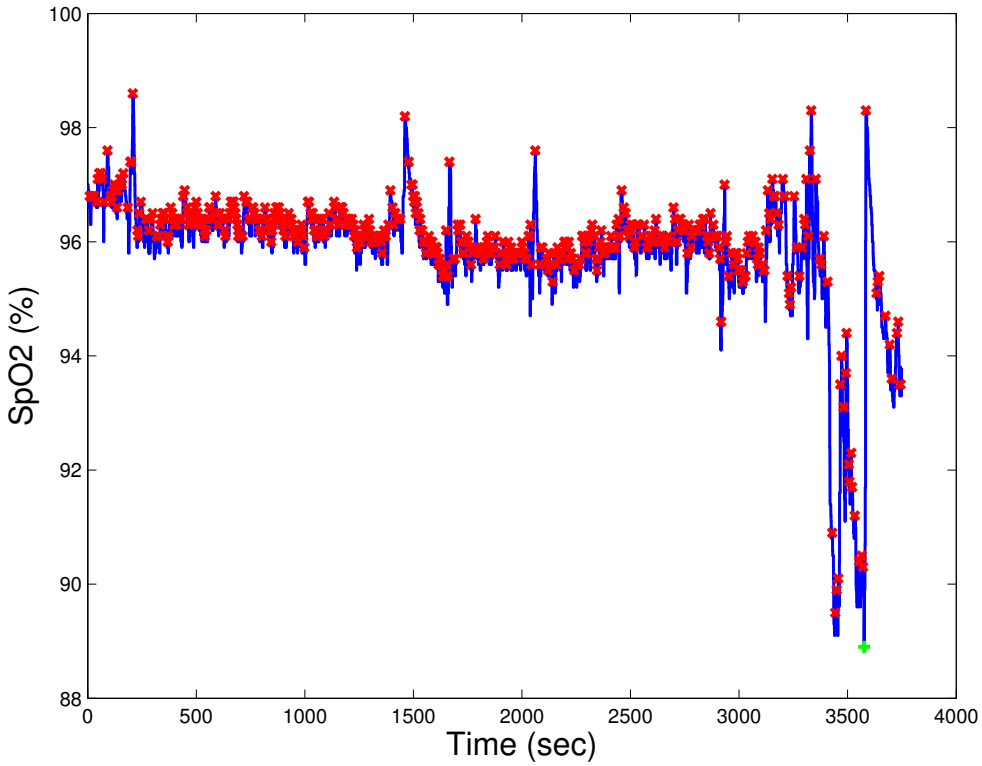
The first part of this work involved calculating the ODI and using SQIs to try to improve the classification accuracy. Next MSE was applied to an approximate of HRV (pulse rate variability (PRV)) data to distinguish between non-treatment and treatment subjects. Finally, a similar analysis to that in section 6.4 was carried out, combining audio and PPG features with the best ODI per fold replacing $\text{ODI}_{VIS\,I}$. From Chapter 6, it was clear that actigraphy features actually diminish classifier performance. Therefore, actigraphy features were not combined with audio and PPG features.

7.4.1 ODI

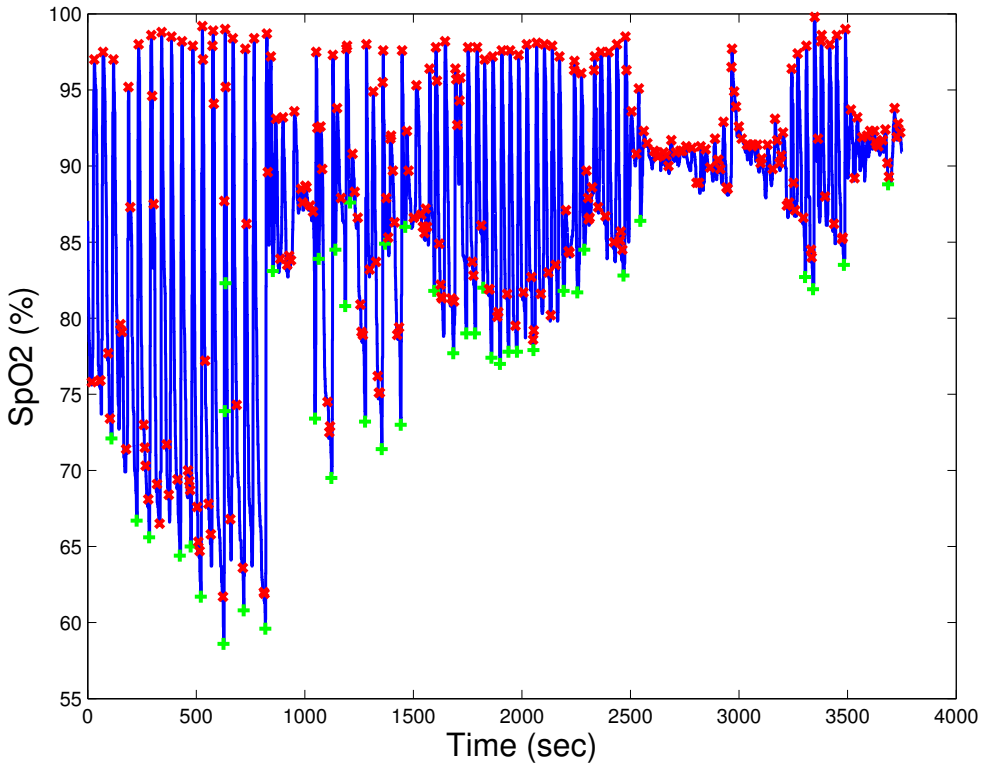
Four different ODIs were calculated: ODI3_b and ODI4_b are basic calculations of ODI without considering signal quality for dips of 3% or 4%; ODI3_{SQI} and ODI4_{SQI} use signal quality to count the number of 3% or 4% dips in saturation. Initially, the saturation signal was scanned through and the locations of all peaks were noted. The lowest point in the data between two consecutive peaks was noted. If that lowest point was more than either 3 or 4% from the first peak, with a following rise of $threshold-1\%$, it was taken to be an actual dip in saturation. The location of each of these dips was noted. For the basic ODI calculation (ODI3_b and ODI4_b), the number of dips was summed over the entire night and divided by the number of hours of sleep.

Figure 7.1 shows the dips that the dip detection algorithm identified for a normal subject and a severe apnoeic. Dips are noted by a green plus sign and peaks by a red cross. It is clear that the normal subject has one oxygen desaturation during this period and so only one dip is identified. The apnoeic subject experiences multiple apnoeas which the algorithm identifies.

The locations of the dips for the basic ODI calculation were the starting point for the SQI ODI calculation. For each dip, the average SQIs over the preceding $30s$ was noted. By varying the SQI levels, it was possible to exclude dips of low quality, *i.e.* those dips that were likely due to artefacts. Each SQI, as well as the average of all three SQIs, was varied between zero and 100 ($0:5:100$); if the average SQI for the preceding $30s$ to the current dip was less than the threshold, the dip was excluded from the analysis. This caused the ODI to decrease as the SQI threshold increased.



(a) *Normal subject*



(b) *Severe apnoeic*

Figure 7.1: *The dip detection algorithm identifying dips (green plus signs) and peaks (red crosses) for a normal subject and a severe apnoeic.*

7.4.2 Pulse Rate and MSE of Pulse Rate Variability

Costa *et al.* [177] applied MSE to the RR time series generated from ECG signals. In that instance, the R peaks of the ECG signal were detected and the RR time series was generated by taking the intervals between the R peaks. In this case, the locations (in samples) of each pulse were found using the PPG pulse detection algorithm described in section 7.3.1. Although not based on the ECG, an approximation of the RR interval time series was found by taking the time interval between successive pulses detected in the PPG. This resulted in an unevenly sampled time series. According to Costa *et al.* [177], outliers can be excluded one of two ways: by removing outliers prior to applying MSE or by using an r value that is calculated by excluding the outliers. The first method was chosen and any pulse-to-pulse (PP) intervals that had changed by more than 20% over the previous PP interval were removed. This method is based on the work of Clifford *et al.* [180] and removed any outliers, artefacts and ectopic pulses. This process, as well as the pulse detection algorithm, caused periods of missing data. In order to ensure that all subjects had four full hours of data for the MSE analysis, periods of missing data were identified by searching for PP intervals that were greater than two (which is physiologically impossible). The lengths of these missing data periods were replaced by multiple values of the mean PP interval, *i.e.* if the period was $10s$ long and the mean PP interval was $1s$ then ten PP interval values of $1s$ were entered into the time series in place of the missing data section. MSE was then applied to this time series in a similar manner to audio and actigraphy. The same $240min$ of data were analysed, beginning $30min$ into the recording and ending at $4.5h$. Nine MSE coefficients were calculated per subject ($\tau = 1, 2, 4, 8, 16, 32, 65, 130, 180$) for $m = 1 : 1 : 8$ and $r = 0.1 : 0.05 : 0.25$; all were used as inputs to the classifiers.

7.4.3 Classification

It should be noted that the exact same folds that were used in the audio analysis have been used in all parts of this analysis (ODI, PRV MSE and combination). The classifiers used the same parameter as in Chapters 5 and 6 for the MSE of PRV and data combination analyses.

7.4.3.1 ODI search

Five-fold cross-validation was carried out (the same five folds as in the other parts of the analysis). Each time, one fold was held separately to be the validation set (X_{val}). The other four folds (X_{design}) were used to find the best $ODI3_b$, $ODI4_b$, $ODI3_{SQI}$ and $ODI4_{SQI}$.

For the basic ODI, the training data was thresholded between the minimum and maximum values. The performance statistics were noted for each of the thresholds. The threshold that gave the highest Ac was taken to be the threshold for the validation data. The results for both training and validation data were noted for each fold. This was done for both $ODI3_b$ and $ODI4_b$.

The analysis for the SQI ODIs was more complex. For each of the four SQIs (three and the average of the three), the SQI threshold was varied, $0 : 5 : 100$, and those dips with a SQI below the SQI threshold were removed and the ODI recalculated. The data in X_{deisgn} for each of the 21 SQI thresholds were further divided into training (70%) and test (30%) 100 times. The training data was thresholded between the minimum and maximum values and the performance statistics noted. The threshold that gave the highest Ac was taken as the threshold for the test data and the performance statistics for both training and test was noted. The highest Ac on the training data for the SQI thresholds ($0 : 5 : 100$) was noted, as was the corresponding SQI threshold. When the 100 iterations were finished, the SQI threshold that occurred most often was taken to be the best SQI threshold for that ODI. Once that was found, the full training data, *i.e.* all of X_{design}, was used to find the threshold that gave the highest Ac. Once this had been done for each of the SQIs, the SQI with the highest Ac was taken to be the best ODI for that fold. This ODI was taken to be the ODI for the training data. Similar to the basic ODI analysis, the training data was thresholded between the minimum and maximum values. The performance statistics were noted for each of the thresholds. The threshold that gave the highest Ac was taken to be the threshold for the validation data. The results for both training and validation data were noted for each fold, along with the number of times each SQI was chosen and the corresponding SQI threshold that was used. This was done for both $ODI3_{SQI}$ and $ODI4_{SQI}$.

7.4.3.2 MSE of PRV

The same classification process was used for PRV as for audio and actigraphy, except there was no downsampling rate to find, just the best m, r combination. Five-fold cross-validation was carried out on the data. Each time, one fold was held separately to be the validation set (X_{val}). The other four folds (X_{design}) were further divided into training and test data sets (70% and 30%) n times in order to find the best MSE m value and r value. For LDA this was carried out 100 times where the MSE coefficients were normalised (zero mean and unit variance based on the training data); for the RF, this process was carried out 5 times. The difference was due to computation time. The highest accuracy was noted along with the m, r combination that it corresponded to. The best overall combination was taken to be the one that was chosen most often in the n iterations.

When the best m, r combination had been found using the training and test data, the full four folds (*i.e.* all of X_{design}) were used as training and the untouched fold was used for validation (X_{val}). Both classifiers were used to test all combinations of features, namely: MSE, $MSE + demographics$, $MSE + ODI$, and $MSE + demographics + ODI$. In each case, the MSE coefficients that were used were those found in the first stage, *i.e.* the MSE coefficients that corresponded to the m, r combination that gave the highest accuracy for each classifier. The results were noted for each feature combination, classifier and fold.

For LDA, any missing data in X_{design} were mean imputed and then normalised (zero mean and unit variance based on X_{design}). A multivariate normal density was fitted to each group with a diagonal covariance matrix estimate (essentially a naive Bayes classifier). Naive Bayes classification assumes that all features are independent within each class. Although this assumption is not true in this case, these classifiers are known to work well even when the independence assumption is not valid, and so it is possible to use them to estimate classification accuracy. For the RF, 500 trees were used in the forest. Each tree split on three variables/features. The process was repeated twice with a new seed for 2×10^6 iterations.

7.4.3.3 Feature combination

As seen in Chapter 6, combining audio and actigraphy features did not result in improved performance over that achieved using audio features alone. As such, only audio and PRV

Table 7.1: *The average performance statistics for training (Train) and validation (Val) over all five folds (mean±σ) for the basic ODI calculation.*

Features	Data Set	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
ODI3$_b$	Train	81.8 ± 4.4	89.2 ± 3.0	83.4 ± 3.6	88.3 ± 2.2	86.3 ± 0.6	0.93 ± 0.00
	Val	82.9 ± 5.3	89.9 ± 3.6	84.5 ± 4.3	88.7 ± 4.5	87.1 ± 2.3	0.93 ± 0.02
ODI4$_b$	Train	70.1 ± 3.1	94.7 ± 1.4	89.7 ± 2.0	82.9 ± 1.7	85.0 ± 0.6	0.91 ± 0.01
	Val	70.1 ± 3.5	94.5 ± 1.4	89.0 ± 3.5	82.7 ± 4.3	84.7 ± 1.9	0.91 ± 0.02

Table 7.2: *The average performance statistics from training (Train) and validation (Val) over all five folds (mean±σ) for the SQI ODI calculation.*

Features	Data Set	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
ODI3$_{SQI}$	Train	80.5 ± 0.9	92.4 ± 0.8	87.4 ± 1.1	87.9 ± 0.8	87.7 ± 0.6	0.93 ± 0.00
	Val	79.7 ± 3.6	92.0 ± 2.3	86.7 ± 3.6	87.2 ± 3.7	87.1 ± 2.7	0.94 ± 0.02
ODI4$_{SQI}$	Train	77.1 ± 0.8	92.5 ± 0.6	87.1 ± 1.0	86.1 ± 0.9	86.5 ± 0.4	0.92 ± 0.00
	Val	76.6 ± 1.3	92.2 ± 2.2	86.6 ± 2.9	85.6 ± 3.1	86.0 ± 1.9	0.92 ± 0.02

features were combined. The best dsr, m, r combination for audio per fold was combined with the best m, r combination for PRV into a MSE feature vector, *i.e.* the best nine audio MSE coefficients per subject were concatenated with the best nine PRV MSE coefficients. The ISI audio features stood alone, as the PRV analysis did not include them. Once this was done, it was possible to test all combinations of combined features[3]. However, the ODI used changed, depending on which approach gave the best classification accuracy for that fold.

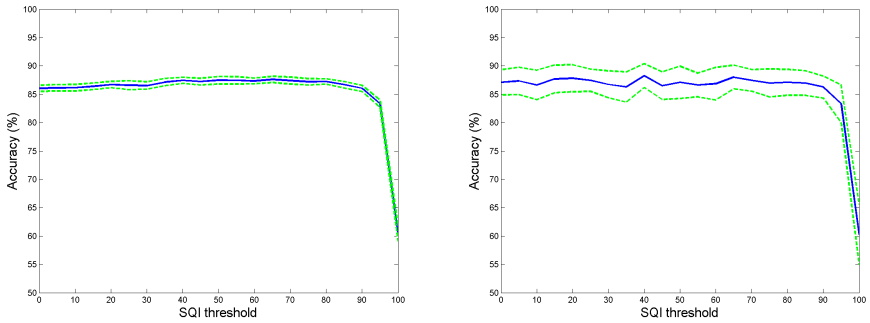
7.5 Results

Three different analyses have been carried out: using the different ODI calculations to classify subjects; applying MSE to the PP intervals obtained from pulses detected in the PPG signal; combining audio, actigraphy, PP features with the different ODI calculations.

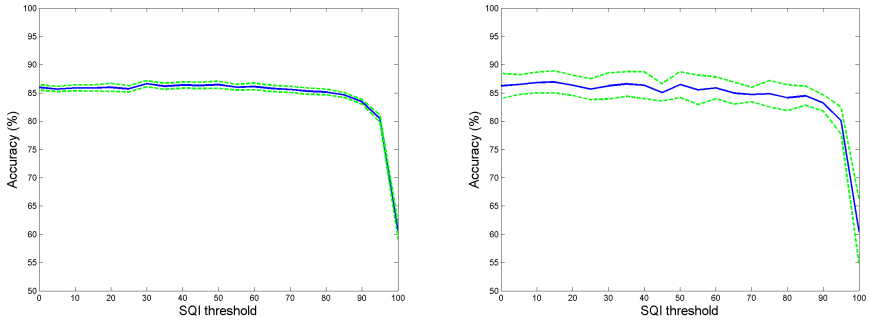
7.5.1 ODI Results

Figures 7.2 and 7.3 show how the accuracy of the ODI estimate changes depending on which SQI threshold is used, averaged across all five folds. The mean is the blue line, while the mean±σ are the dashed green lines. The results for the basic ODI calculation can be found in Table 7.1, while the results for the best SQI ODI can be found in Table 7.2.

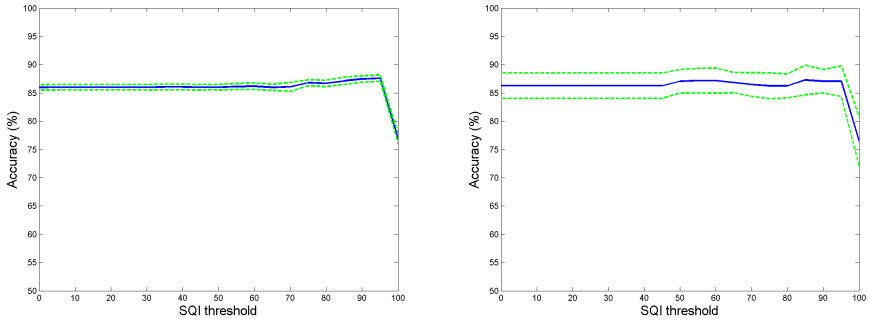
[3] MSE, $ISI1$, $ISI2$, $MSE + ISI1$, $MSE + ISI2$, $MSE + ISI1 + ISI2$, $MSE + ISI1 + ISI2 +$ *demographics*, $MSE + ISI1 + ISI2 + ODI$, and $MSE + ISI1 + ISI2 +$ *demographics* $+ ODI$



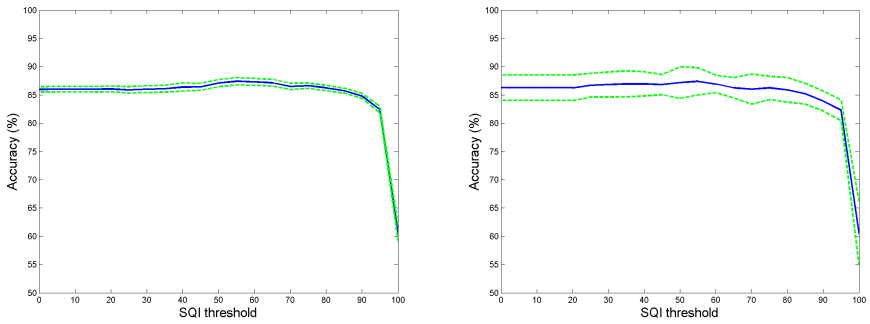
(a) SQI_1, *Train* **(b)** SQI_1, *Val*



(c) SQI_2, *Train* **(d)** SQI_2, *Val*

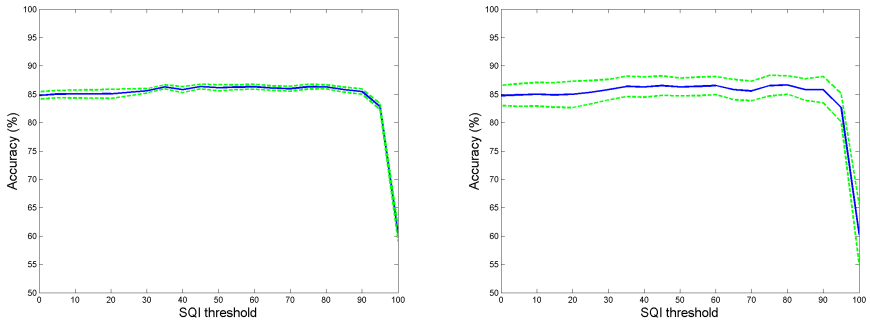


(e) SQI_3, *Train* **(f)** SQI_3, *Val*

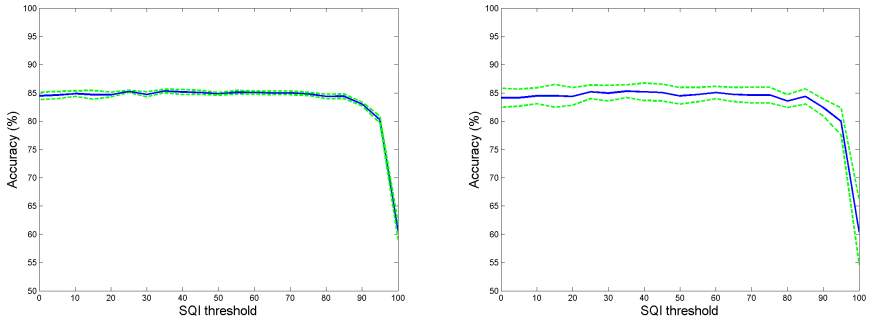


(g) SQI_4, *Train* **(h)** SQI_4, *Val*

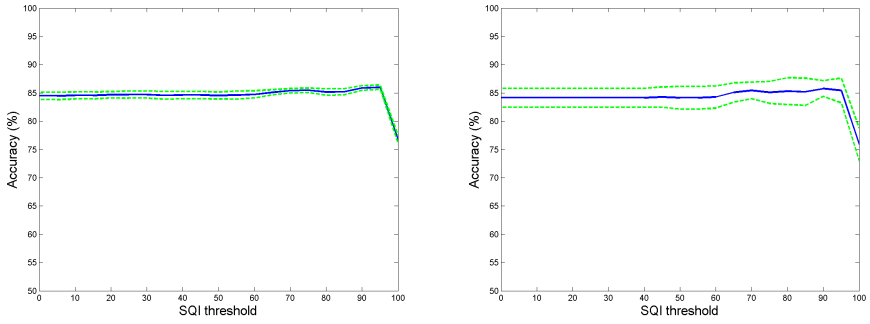
Figure 7.2: *The accuracy of the ODI at varying levels of the SQI threshold across the five folds. The results (mean $\pm\sigma$) are shown for training (Train) and validation (Val) using 3%, or more, dips.*



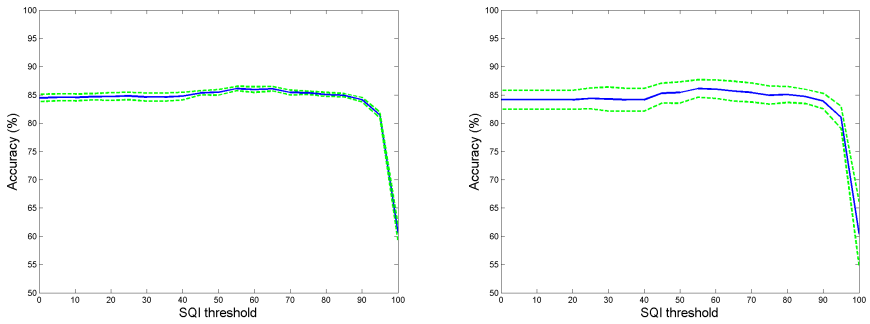
(a) SQI_1, *Train* **(b)** SQI_1, *Val*



(c) SQI_2, *Train* **(d)** SQI_2, *Val*



(e) SQI_3, *Train* **(f)** SQI_3, *Val*



(g) SQI_4, *Train* **(h)** SQI_4, *Val*

Figure 7.3: *The accuracy of the ODI at varying levels of the SQI threshold across the five folds. The results (mean $\pm\sigma$) are shown for training (Train) and validation (Val) using 4%, or more, dips.*

7.5.2 MSE of PRV Results

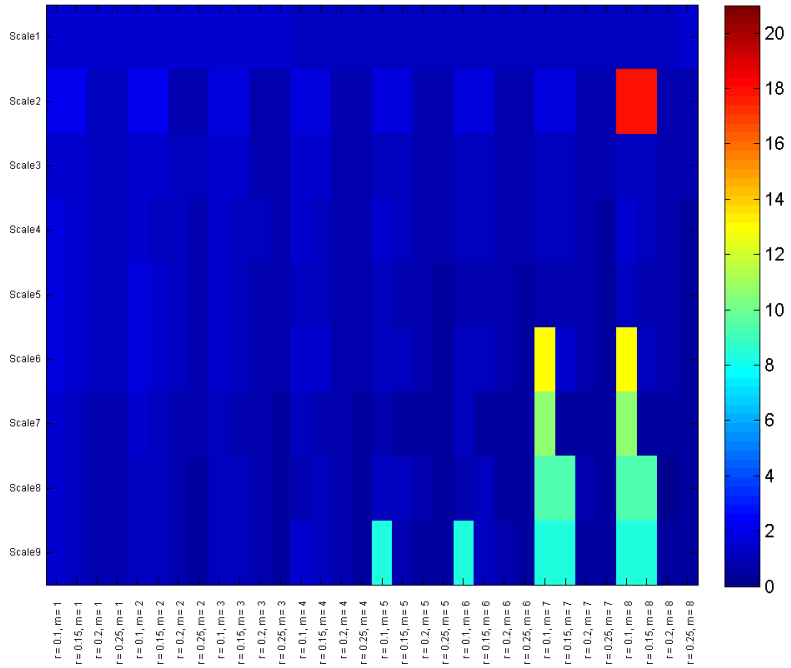
Figure 7.4 shows the MSE values at the different scale factors (τ) for all combinations of m, r for a snoring subject and a severe apnoeic. This figure indicates that there are differences in the MSE value between the two subjects.

Figure 7.5 shows the number of times each MSE parameter combination was chosen per fold for both classifiers. Figure 7.5a shows that when using LDA to choose the best parameter combination, the results were: for the first, second and fifth folds, the best MSE parameter combination was $m = 1, r = 0.2$ chosen 46, 38 and 56 times respectively. For the third and fourth folds the best combination chosen was $m = 1, r = 0.25$ chosen 42 and 34 times respectively. In fact, $m = 1$ was chosen more than 90% of the times across the folds (95, 91, 93, 92, and 93% respectively). Figure 7.5b shows that when using a RF to choose the best parameter combination the results were: for the first fold, the best MSE parameter combination was $m = 1, r = 0.2$ which was chosen 3 times. This combination was chosen again in the third (two times) and fourth (three times) folds. In the second fold, the best combination was $m = 1, r = 0.25$ chosen two times. For the fifth fold, the best combination chosen was $m = 1, r = 0.1$ chosen twice. As for LDA, the value of m appears to be more important than the tolerance (r). In each of the folds, $m = 1$ was chosen more than 80% of the time (100, 100, 80, 100 and 100% respectively).

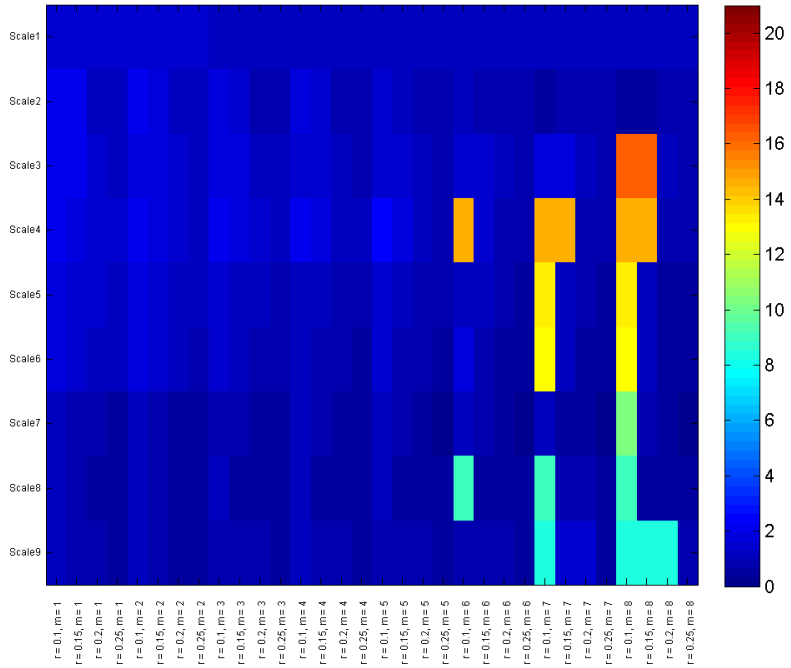
The average results over the five folds when using LDA can be found in Table 7.3 while the RF results can be found in Table 7.4. The best performance for both classifiers is highlighted in blue. Boxplots of the predictions can be found in Figures 7.6 and 7.8, while ROC curves can be found in Figures 7.7 and 7.9.

Table 7.3: *The average performance statistics over all five folds (mean±σ) for different PPG feature combinations on both the training (Train) and validation (Val) data sets when using LDA to classify subjects.*

Features	Data Set	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{prv}$	Train	50.4 ± 5.0	85.3 ± 2.7	69.3 ± 2.1	72.5 ± 0.8	71.6 ± 1.0	0.73 ± 0.01
	Val	50.5 ± 5.4	85.0 ± 7.6	69.3 ± 13.7	72.4 ± 4.6	71.0 ± 3.9	0.72 ± 0.04
MSE$_{prv}$+demos	Train	64.5 ± 2.1	81.1 ± 1.9	69.1 ± 1.4	77.8 ± 1.5	74.6 ± 1.4	0.79 ± 0.01
	Val	63.0 ± 7.6	77.3 ± 4.6	65.6 ± 8.0	75.0 ± 7.4	71.0 ± 3.8	0.78 ± 0.04
MSE$_{prv}$+ODI	Train	60.4 ± 0.7	99.2 ± 0.4	98.0 ± 0.8	79.3 ± 0.8	83.9 ± 0.6	0.93 ± 0.01
	Val	60.6 ± 3.5	99.2 ± 0.9	98.0 ± 2.0	79.4 ± 4.9	84.0 ± 3.5	0.93 ± 0.02
MSE$_{prv}$+demos+ODI	Train	65.0 ± 1.5	94.2 ± 0.7	88.0 ± 1.0	80.5 ± 1.0	82.7 ± 0.9	0.91 ± 0.01
	Val	63.3 ± 7.0	93.2 ± 3.8	86.2 ± 8.6	78.4 ± 7.0	80.8 ± 5.5	0.90 ± 0.04

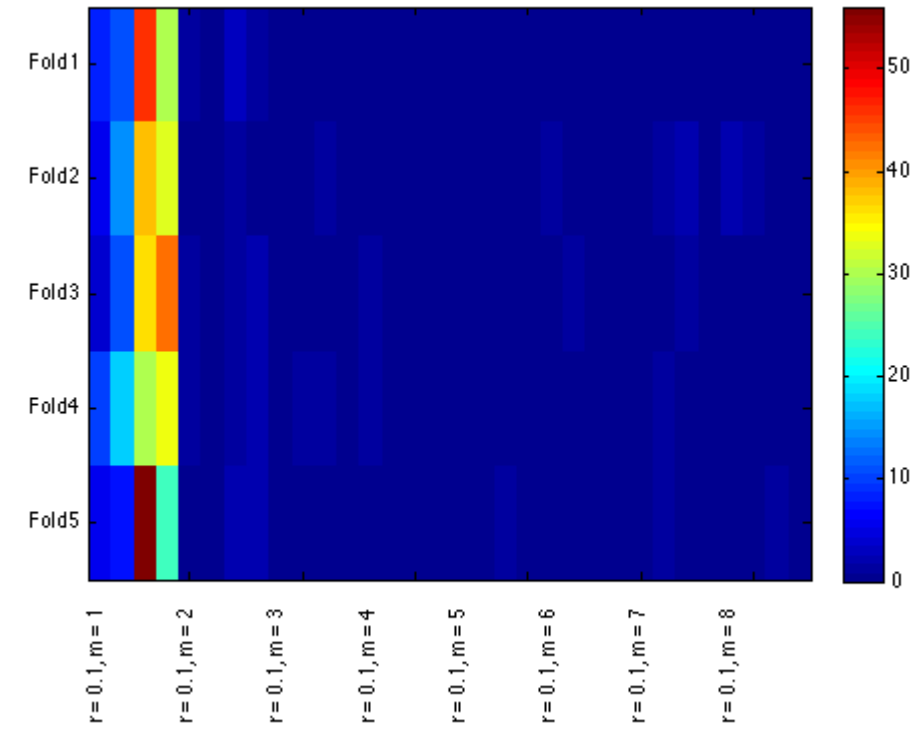


(a) *Snorer*

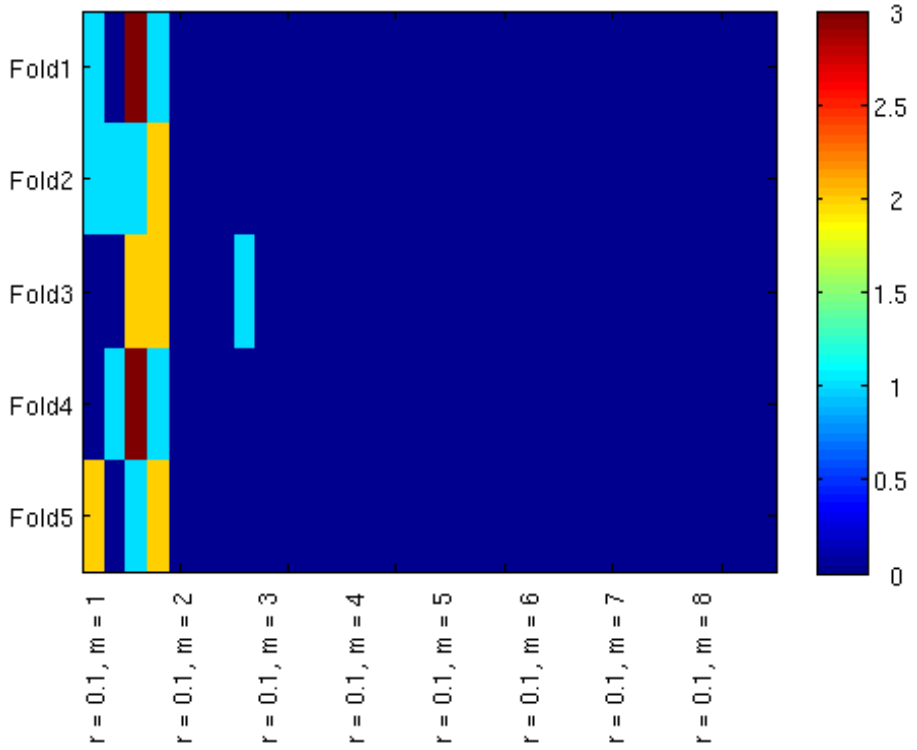


(b) *Severe apnoeic*

Figure 7.4: *Heatmaps illustrating the differences in PRV MSE values at different scale factors for all combinations of m, r for a snoring subject and a severe apnoeic.*

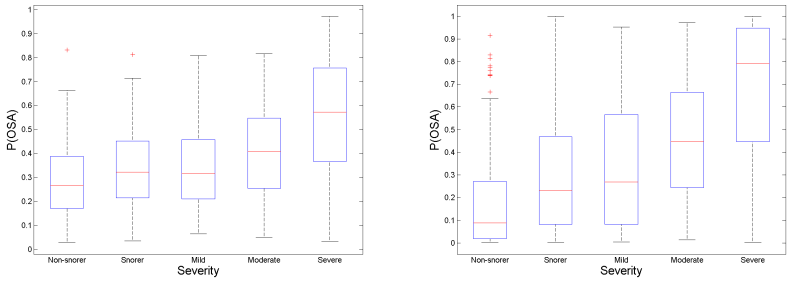


(a) *LDA*

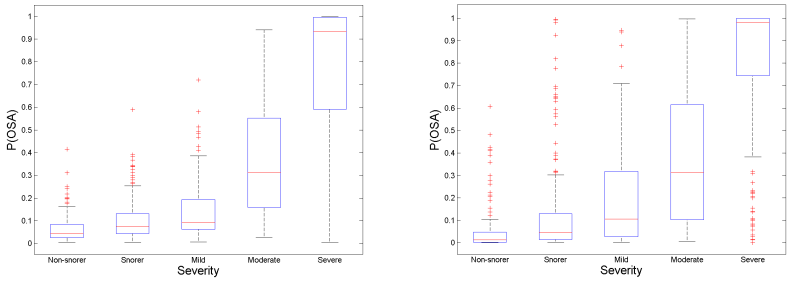


(b) *RF*

Figure 7.5: *Heatmaps showing how often MSE parameter combinations using PRV were chosen per fold for the different classifiers.*

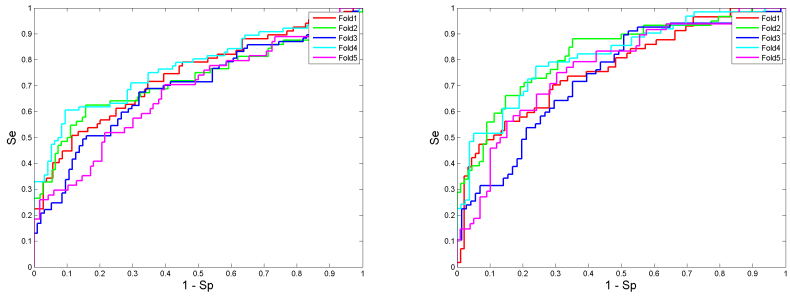


(a) MSE_{prv} **(b)** MSE_{prv}+*demos*

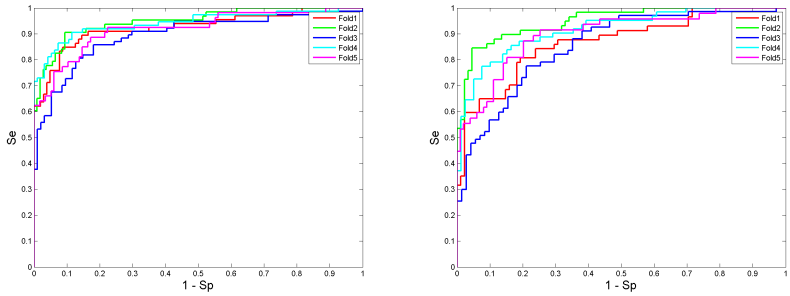


(c) MSE_{prv}+*ODI* **(d)** MSE_{prv}+*demos*+*ODI*

Figure 7.6: *Boxplots of the LDA predictions on the validation data over all five folds for the different PRV feature combinations for the PRV analysis.*



(a) MSE_{prv} **(b)** MSE_{prv}+*demos*

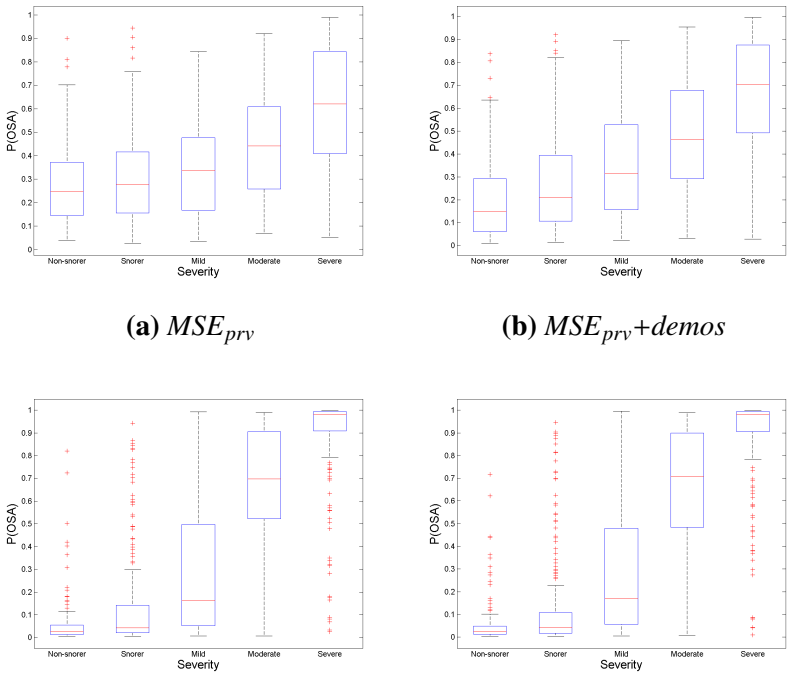


(c) MSE_{prv}+*ODI* **(d)** MSE_{prv}+*demos*+*ODI*

Figure 7.7: *ROC curves of the LDA predictions on the validation data over all five folds for the different PRV feature combinations for the PRV analysis. Red = fold 1, green = fold 2, blue = fold 3, cyan = fold 4, magenta = fold 5.*

Table 7.4: *The average performance statistics over all five folds (mean±σ) for different PPG feature combinations on the validation data set when using a RF to classify subjects.*

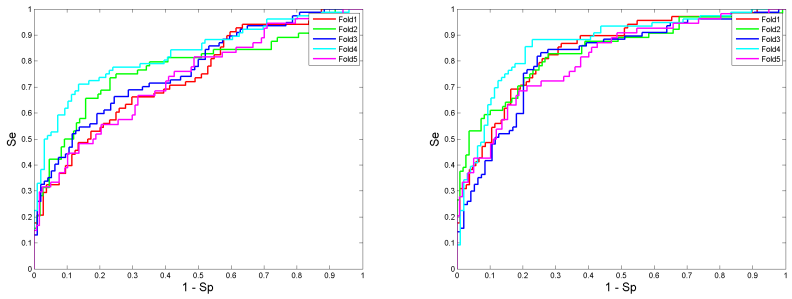
Features	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{prv}$	54.9 ± 4.2	85.3 ± 5.4	70.7 ± 12.0	74.3 ± 3.8	73.1 ± 2.8	0.77 ± 0.04
MSE$_{prv}$+demos	63.5 ± 5.6	84.1 ± 3.2	72.0 ± 7.3	77.8 ± 4.9	75.9 ± 3.1	0.83 ± 0.02
MSE$_{prv}$+ODI	87.4 ± 4.6	90.5 ± 2.6	85.7 ± 4.0	91.3 ± 4.2	89.2 ± 3.0	0.95 ± 0.01
MSE$_{prv}$+demos+ODI	86.1 ± 5.7	90.9 ± 1.9	86.0 ± 2.4	90.5 ± 5.0	88.8 ± 3.0	0.96 ± 0.01



(a) *MSE$_{prv}$* **(b)** *MSE$_{prv}$+demos*

(c) *MSE$_{prv}$+ODI* **(d)** *MSE$_{prv}$+demos+ODI*

Figure 7.8: *Boxplots of the RF predictions on the validation data over all five folds for the different PPG feature combinations for the PRV analysis.*



(a) MSE_{prv} **(b)** MSE_{prv}+*demos*

(c) MSE_{prv}+*ODI* **(d)** MSE_{prv}+*demos*+*ODI*

Figure 7.9: *ROC curves of the RF predictions on the validation data over all five folds for the different PRV feature combinations for the PRV analysis. Red = fold 1, green = fold 2, blue = fold 3, cyan = fold 4, magenta = fold 5.*

7.5.3 Combined Feature Results

When excluding actigraphy, *i.e.* using audio, PPG and demographics, the average results over the five folds when using LDA can be found in Table 7.5 while the RF results can be found in Table 7.7. The best MSE parameter combinations picked for audio and PRV per fold have been used, along with the best ODI calculation for that fold. The best results are highlighted in blue. Tables 7.6 and 7.8 show the NRI or IDI and corresponding p value obtained when the different ODI calculations are compared. In all cases the features used are $MSE + ISI + ODI$, where the only difference between the models is the ODI value.

Table 7.9 shows the best performance on the validation data set for each of the feature sets along with the performance obtained using ODI_{VISI} alone for comparison.

Figure 7.10 shows scatter plots of the prediction by MSE_{aud} vs the ODI3_b value. This figure illustrates whether MSE_{aud} adds any value to ODI3_b. Figures 7.10a and 7.10b show the TP, TN,FP, FN for MSE_{aud} and ODI3_b respectively, while Figures 7.10c and 7.10d highlight the cases where MSE_{aud} predicts incorrectly but ODI3_b predicts the outcome correctly and ODI3_b predicts incorrectly but MSE_{aud} predicts correctly, respectively.

7.6 Discussion

7.6.1 Discussion of ODI Results

Figures 7.2 and 7.3 demonstrate that, regardless of the SQI threshold used, the results remain similar up to a point. That is, for 3% dips, this SQI threshold is 90, 85, 80 and 85 while for 4% dips, this SQI threshold is 90, 85, 85 and 85 for SQI_1, SQI_2, SQI_3 and SQI_4 respectively. This is confirmed in Tables 7.1 and 7.2, where for both the basic ODI calculation and the SQI ODI, using 3% dips out-performs 4% dips. Table 7.1 shows that ODI3_b achieved Ac = 86.3% in training and 87.1% during validation, while Table 7.2 shows that ODI3_{SQI} achieves Ac = 87.7% in training and 87.1% during validation. Comparing the two tables shows that the use of SQIs only improves performance marginally, however performance is not as good as that achieved for ODI_{VISI} in table 5.2 (AC = 87.1% compared to 88.7%).

It is interesting to note that ODI_{SQI} consistently achieves better performance when combined with other features (audio-derived, actigraphy-derived, PPG-derived or demographics),

Table 7.5: *The average performance statistics over all five folds (mean±σ) for different audio and PPG feature combinations on both the training (Train) and validation (Val) data sets when using LDA to classify subjects.*

Features	Data Set	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{aud+prv}$	Train	52.9 ± 3.8	81.5 ± 5.9	65.8 ± 5.1	72.6 ± 1.5	70.3 ± 2.5	0.74 ± 0.02
	Val	54.4 ± 8.8	79.7 ± 5.4	63.8 ± 6.6	72.7 ± 6.8	69.3 ± 2.5	0.73 ± 0.03
ISI1	Train	33.2 ± 3.9	85.2 ± 2.1	59.4 ± 1.0	66.1 ± 1.7	64.7 ± 1.5	0.65 ± 0.01
	Val	34.4 ± 6.4	85.5 ± 1.5	60.3 ± 4.9	66.5 ± 6.9	65.1 ± 5.5	0.65 ± 0.03
ISI2	Train	33.2 ± 3.9	85.2 ± 2.1	59.4 ± 1.0	66.1 ± 1.7	64.7 ± 1.5	0.65 ± 0.01
	Val	34.4 ± 6.4	85.5 ± 1.5	60.3 ± 4.9	66.5 ± 6.9	65.1 ± 5.5	0.65 ± 0.03
MSE$_{aud+prv}$+ISI1	Train	57.4 ± 2.6	81.3 ± 3.7	67.0 ± 2.7	74.5 ± 1.2	71.9 ± 1.5	0.76 ± 0.01
	Val	57.4 ± 9.3	79.1 ± 6.8	64.5 ± 9.7	73.9 ± 6.7	70.3 ± 4.4	0.76 ± 0.04
MSE$_{aud+prv}$+ISI2	Train	57.4 ± 2.6	81.3 ± 3.7	67.0 ± 2.7	74.5 ± 1.2	71.9 ± 1.5	0.76 ± 0.01
	Val	57.4 ± 9.3	79.1 ± 6.8	64.5 ± 9.7	73.9 ± 6.7	70.3 ± 4.4	0.76 ± 0.04
MSE$_{aud+prv}$+ISI1+ISI2	Train	59.5 ± 1.9	79.1 ± 2.3	65.1 ± 1.4	74.9 ± 1.3	71.4 ± 1.2	0.76 ± 0.01
	Val	58.3 ± 11.2	79.5 ± 6.8	65.0 ± 11.1	74.4 ± 7.7	70.9 ± 6.5	0.76 ± 0.04
MSE$_{aud+prv}$+ISI1+ISI2+demos	Train	65.6 ± 1.0	81.3 ± 2.2	69.7 ± 1.3	78.3 ± 1.4	75.1 ± 1.5	0.81 ± 0.01
	Val	63.7 ± 9.1	80.4 ± 3.7	68.7 ± 8.5	76.1 ± 7.2	73.3 ± 4.5	0.81 ± 0.03
MSE$_{aud+prv}$+ISI1+ISI2+ODI$_{VISI}$	Train	65.2 ± 3.5	91.4 ± 2.7	83.4 ± 4.0	80.1 ± 1.4	81.1 ± 1.5	0.89 ± 0.01
	Val	64.7 ± 6.0	90.8 ± 3.3	82.0 ± 6.2	79.6 ± 5.8	80.5 ± 4.5	0.89 ± 0.03
MSE$_{aud+prv}$+ISI1+ISI2+demos+ODI$_{VISI}$	Train	68.5 ± 2.3	91.0 ± 1.2	83.3 ± 1.9	81.6 ± 1.5	82.1 ± 1.3	0.89 ± 0.01
	Val	67.1 ± 3.4	90.8 ± 2.1	83.3 ± 4.3	79.7 ± 5.3	81.0 ± 2.9	0.89 ± 0.02
MSE$_{aud+prv}$+ISI1+ISI2+ODI3$_b$	Train	61.5 ± 3.3	82.9 ± 3.3	70.4 ± 3.4	76.7 ± 1.6	74.5 ± 1.7	0.80 ± 0.01
	Val	59.2 ± 8.9	82.8 ± 6.2	69.4 ± 9.6	75.5 ± 6.6	73.2 ± 4.8	0.79 ± 0.05
MSE$_{aud+prv}$+ISI1+ISI2+demos+ODI3$_b$	Train	67.0 ± 1.0	84.1 ± 1.9	73.5 ± 1.2	79.6 ± 1.1	77.4 ± 1.1	0.84 ± 0.01
	Val	65.4 ± 7.6	83.8 ± 4.1	73.3 ± 8.1	77.5 ± 7.1	76.0 ± 4.7	0.83 ± 0.03
MSE$_{aud+prv}$+ISI1+ISI2+ODI4$_b$	Train	60.6 ± 2.9	81.6 ± 3.1	68.4 ± 2.7	76.0 ± 1.7	73.3 ± 1.7	0.79 ± 0.01
	Val	59.1 ± 9.5	80.9 ± 6.6	67.0 ± 10.1	75.1 ± 6.8	72.0 ± 5.0	0.78 ± 0.05
MSE$_{aud+prv}$+ISI1+ISI2+demos+ODI4$_b$	Train	66.4 ± 0.8	83.1 ± 2.0	72.0 ± 1.3	79.1 ± 1.4	76.5 ± 1.3	0.83 ± 0.01
	Val	63.8 ± 8.4	82.2 ± 3.7	70.8 ± 7.9	76.4 ± 7.5	74.3 ± 4.8	0.82 ± 0.03
MSE$_{aud+prv}$+ISI1+ISI2+ODI3$_{SQI}$	Train	62.5 ± 3.2	87.7 ± 4.4	77.5 ± 6.0	78.2 ± 1.2	77.8 ± 1.9	0.84 ± 0.01
	Val	60.2 ± 6.8	87.0 ± 4.6	75.4 ± 7.7	76.7 ± 6.2	76.2 ± 4.2	0.84 ± 0.03
MSE$_{aud+prv}$+ISI1+ISI2+demos+ODI3$_{SQI}$	Train	67.2 ± 2.4	88.3 ± 2.2	79.0 ± 2.1	80.5 ± 1.4	80.0 ± 1.4	0.87 ± 0.01
	Val	65.2 ± 6.0	88.3 ± 2.9	79.3 ± 5.8	78.3 ± 5.9	78.7 ± 2.9	0.87 ± 0.01
MSE$_{aud+prv}$+ISI1+ISI2+ODI4$_{SQI}$	Train	61.3 ± 3.8	85.0 ± 4.5	73.3 ± 4.9	77.1 ± 1.3	75.7 ± 1.4	0.81 ± 0.02
	Val	59.5 ± 8.0	84.0 ± 7.0	71.5 ± 10.2	75.9 ± 6.4	74.1 ± 4.7	0.81 ± 0.04
MSE$_{aud+prv}$+ISI1+ISI2+demos+ODI4$_{SQI}$	Train	66.4 ± 1.2	86.0 ± 3.2	75.7 ± 3.5	79.6 ± 1.2	78.2 ± 1.9	0.85 ± 0.01
	Val	65.1 ± 5.2	85.8 ± 4.6	75.9 ± 8.2	77.8 ± 5.9	77.1 ± 3.6	0.84 ± 0.02

Table 7.6: *The NRI and corresponding p value obtained when comparing the predictions across all five folds using the different ODI calculations. Audio-derived and PPG-derived features plus demographics have been classified using LDA. The ‡ denotes those NRI values that are significant.*

	ODI$_{VISI}$		**ODI3$_b$**		**ODI4$_b$**		**ODI3$_{SQI}$**		**ODI4$_{SQI}$**	
	NRI	*p*	**NRI**	*p*	**NRI**	*p*	**NRI**	*p*	**NRI**	*p*
ODI$_{VISI}$	-	-	-0.072	0.000‡	-0.100	0.000‡	-0.036	0.004	-0.060	0.000‡
ODI3$_b$	0.072	0.000	-	-	-0.028	0.002	0.036	0.029	0.011	0.394
ODI4$_b$	0.100	0.000	0.028	0.002	-	-	0.064	0.000	0.040	0.009
ODI3$_{SQI}$	0.036	0.004	-0.036	0.029	-0.064	0.000	-	-	-0.024	0.027
ODI4$_{SQI}$	0.060	0.000	-0.011	0.394	-0.040	0.009	0.024	0.027	-	-

Table 7.7: *The average performance statistics over all five folds (mean±σ) for different audio and PPG feature combinations on the validation data set when using a RF to classify subjects.*

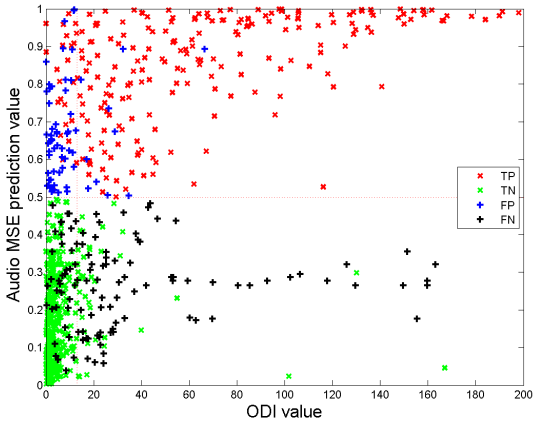
Features	Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{aud+prv}$	69.2 ± 5.3	88.0 ± 2.7	78.8 ± 5.4	81.4 ± 4.3	80.7 ± 3.3	0.88 ± 0.03
ISI1	47.2 ± 4.1	81.5 ± 6.3	62.9 ± 10.6	70.2 ± 5.2	67.7 ± 3.2	0.72 ± 0.04
ISI2	47.5 ± 3.6	82.1 ± 5.7	63.7 ± 10.1	70.4 ± 5.0	68.2 ± 2.4	0.72 ± 0.04
MSE$_{aud+prv}$**+ISI1**	71.0 ± 3.2	87.6 ± 2.9	78.4 ± 7.8	82.2 ± 3.4	81.0 ± 2.3	0.88 ± 0.03
MSE$_{aud+prv}$**+ISI2**	71.8 ± 4.9	87.2 ± 2.3	78.1 ± 6.3	82.4 ± 4.9	81.0 ± 3.0	0.88 ± 0.03
MSE$_{aud+prv}$**+ISI1+ISI2**	71.9 ± 4.7	87.5 ± 2.1	78.6 ± 5.6	82.6 ± 4.3	81.3 ± 3.2	0.88 ± 0.03
MSE$_{aud+prv}$**+ISI1+ISI2+demos**	73.3 ± 5.4	88.0 ± 3.1	79.9 ± 5.5	83.3 ± 4.9	82.2 ± 3.1	0.89 ± 0.03
MSE$_{aud+prv}$**+ISI1+ISI2+ODI**$_{VIS1}$	84.2 ± 2.7	90.8 ± 2.2	85.5 ± 3.8	89.6 ± 2.9	88.1 ± 0.9	0.96 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+demos+ODI**$_{VIS1}$	85.1 ± 4.2	90.1 ± 3.1	85.0 ± 3.9	89.8 ± 4.3	88.0 ± 2.9	0.96 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+ODI3**$_b$	82.5 ± 3.0	90.7 ± 1.8	85.2 ± 2.9	88.6 ± 3.5	87.4 ± 2.0	0.95 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+demos+ODI3**$_b$	83.4 ± 2.4	91.1 ± 2.1	85.9 ± 3.7	89.1 ± 3.3	88.0 ± 2.0	0.95 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+ODI4**$_b$	80.2 ± 3.6	90.2 ± 3.5	84.1 ± 5.6	87.1 ± 4.0	86.1 ± 3.6	0.94 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+demos+ODI4**$_b$	81.7 ± 4.6	90.6 ± 1.5	84.7 ± 3.9	88.1 ± 3.7	87.0 ± 2.4	0.94 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+ODI3**$_{SQI}$	83.4 ± 2.1	90.9 ± 2.8	85.7 ± 4.3	89.1 ± 3.1	87.9 ± 2.3	0.95 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+demos+ODI3**$_{SQI}$	82.8 ± 3.0	90.7 ± 1.5	85.3 ± 2.0	88.7 ± 3.3	87.5 ± 2.3	0.95 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+ODI4**$_{SQI}$	80.7 ± 3.2	91.1 ± 2.3	85.5 ± 4.0	87.6 ± 3.6	86.9 ± 2.9	0.94 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+demos+ODI4**$_{SQI}$	81.3 ± 2.7	91.3 ± 2.0	85.9 ± 2.9	87.9 ± 3.3	87.3 ± 2.5	0.94 ± 0.01

Table 7.8: *The IDI and corresponding p value obtained when comparing the predictions across all five folds using the different ODI calculations. Audio-derived and PPG-derived features plus demographics have been classified using the RF. The ‡ denotes those IDI values that are significant.*

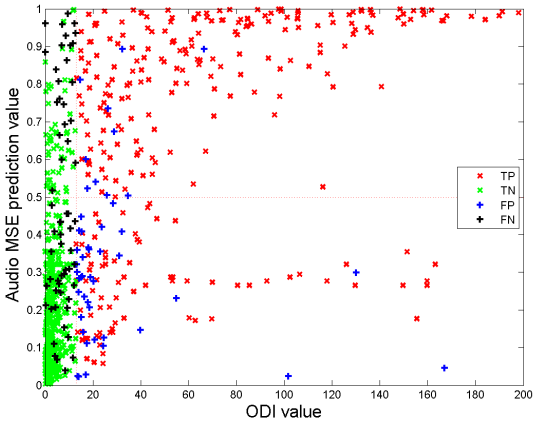
	ODI$_{VIS1}$		**ODI3**$_b$		**ODI4**$_b$		**ODI3**$_{SQI}$		**ODI4**$_{SQI}$	
	IDI	p	**IDI**	p	**IDI**	p	**IDI**	p	**IDI**	p
ODI$_{VIS1}$	-	-	-0.036	0.000‡	-0.055	0.000‡	-0.033	0.000‡	-0.047	0.000‡
ODI3$_b$	0.036	0.000	-	-	-0.019	0.000	0.003	0.365	-0.011	0.014
ODI4$_b$	0.055	0.000	0.019	0.000	-	-	0.022	0.000	0.008	0.033
ODI3$_{SQI}$	0.033	0.000	-0.003	0.365	-0.022	0.000	-	-	-0.014	0.001
ODI4$_{SQI}$	0.047	0.000	0.011	0.014	-0.008	0.033	0.014	0.001	-	-

Table 7.9: *The best performance for the validation data set for each set of features (audio, actigraphy, audio+actigraphy, PRV, audio+PRV) along with the performance obtained using ODI$_{VIS1}$.*

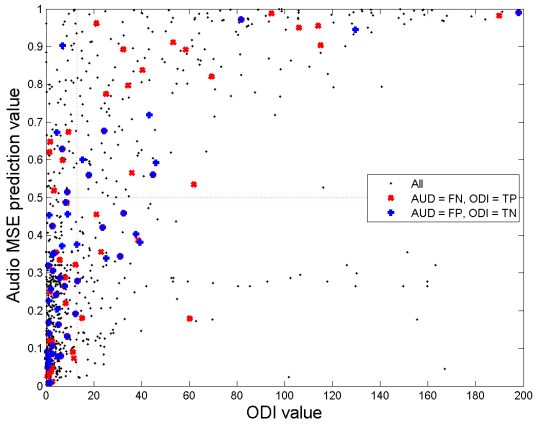
Features	Classifier	Validation Performance					
		Se (%)	Sp (%)	PPV (%)	NPV (%)	Ac (%)	AUC
MSE$_{aud}$**+ISI1+ISI2+demos+ODI**	RF	84.6 ± 4.1	91.3 ± 2.0	86.3 ± 3.1	89.9 ± 3.7	88.6 ± 1.8	0.96± 0.01
MSE$_{prv}$**+ODI**	RF	87.4 ± 4.6	90.5 ± 2.6	85.7 ± 4.0	91.3 ± 4.2	89.2 ± 3.0	0.95 ± 0.01
MSE$_{aud+prv}$**+ISI1+ISI2+demos+ODI3**$_b$	RF	83.4 ± 2.4	91.1 ± 2.1	85.9 ± 3.7	89.1 ± 3.3	88.0 ± 2.0	0.95 ± 0.01
ODI$_{VIS1}$ (15)	-	85.3	90.9	86.2	90.3	88.7	-
ODI3$_b$	-	82.9 ± 5.3	89.9 ± 3.6	84.5 ± 4.3	88.7 ± 4.5	87.1 ± 2.3	0.95 ± 0.02



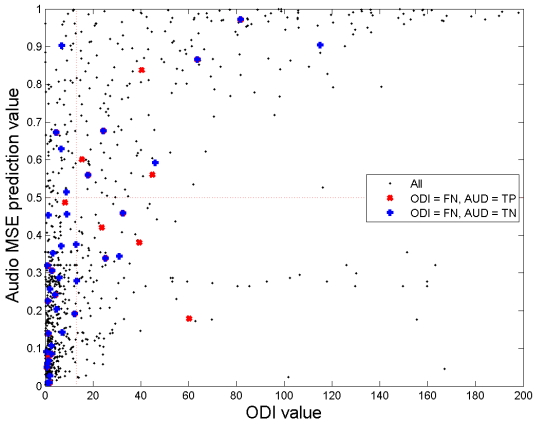
(a) *Audio MSE prediction*



(b) *ODI3$_b$ prediction*



(c) *Aud pred incorrect, ODI3$_b$ correct*



(d) *ODI3$_b$ incorrect, Aud pred correct*

Figure 7.10: *Scatter plots of audio MSE prediction value vs ODI3$_b$ value, highlighting where MSE predicts the outcome correctly and ODI does not, and vice versa.*

which can be seen in Tables 7.5 and 7.7.

7.6.2 Discussion of PRV Results

Figure 7.4 shows the MSE values for different combinations of m and r. As with Figures 5.7 and 6.4, the MSE values have been calculated for the same subjects. Although not as clear as in Figure 5.7, there are some differences between the two subjects. This indicates that MSE of PRV might be able to identify subjects requiring treatment from those who do not.

The search for the best m, r combination is consistent across all five folds which can be clearly seen in Figure 7.5, for both LDA and the RF. Based on the results, $m = 1$ and $r = 0.2$ or $r = 0.25$ are the optimal choice. It is clear that the m value is the most important parameter while the tolerance level (r) chosen is less important. This is similar to the audio MSE parameter search, where $m = 1$ was also chosen by the RF, while for LDA the tolerance level was more important. Both classifiers chose an m value of 1, which is not reflected in the differences between the snoring subject and the severe apnoeic in Figure 7.4. This indicates that the subjects chosen do not have MSE values that are indicative of the differences between the two classes, although they do emphasise the differences for the audio MSE values. Therefore, it is possible that there may be certain cases where audio MSE can identify subjects correctly that PRV MSE does not, and vice versa.

As can be seen from Tables 7.3 and 7.4, the best performance is obtained when using $MSE_{prv} + ODI$ in the analysis. LDA achieved Ac = 83.9% in training and Ac = 84.0% during validation. The RF achieved Ac = 89.2% during validation. Figures 7.6c and 7.8c shows that there is good separation between the two classes, which is confirmed in Figures 7.7c and 7.9c.

For LDA, MSE_{prv} achieves Se = 50.5% and Sp = 85.0% during validation. Using the RF, MSE is more specific (85.3%) than sensitive (54.9%). The accuracy of MSE when using LDA is 71.6% which is similar to that achieved when using the RF (73.1%). Figures 7.6a and 7.8a both show that, although there is good separation between the median values of the two classes, there is overlap between the 25^{th} percentile of the mild OSA cases and the 25^{th} percentile of the moderate OSA cases. There is even overlap between the 75^{th} percentile of mild OSA cases and the 25^{th} percentile of the severe OSA cases.

When combining features, demographics and ODI consistently improve performance for

both classifiers. Demographics improve performance only when combined with the PRV features, however, when all features are combined ($MSE + demo + ODI$), demographics actually worsen performance slightly when compared to $MSE+ODI$. This indicates that when the ODI is unavailable, demographics can be used, in combination with MSE_{prv} to give an accuracy of 75.9% when using a RF (improved from 73.1% when using MSE alone).

These results are promising, and have improved on the results obtained using MSE of actigraphy (Ac = 59.4%). However, they are not as good as those achieved using MSE of audio (Ac = 79.7%). When combined with ODI, the RF outperforms the ODI on its own (Ac = 88.7%) (see Table 5.2).

7.6.3 Discussion of Feature Combination

As shown in Table 6.2, including actigraphy features in the analysis actually worsened performance. Performance improvement by combining audio and PRV features is not as clear cut, as can been seen in Tables 7.5 and 7.7. For LDA, $MSE_{aud+prv}$ achieves 70.3% in training and 69.3% during validation. This is better than the accuracy achieved when using MSE_{aud} (66.3%), but slightly worse than that achieved when using MSE_{prv} alone (71.0%). As ISI1 and ISI2 are only based on audio; the results are the same as those achieved previously in Table 5.7. Including demographics improves performance (75.1%), the same as for audio features alone (69.5%) but not PRV features alone (71.0%). The best feature combination, regardless of the ODI used, is achieved using $MSE_{aud+prv} + ISI + demo + ODI$. Overall, the best accuracy achieved is 82.1% in training and 81.0% during validation using ODI_{VISI} plus $MSE_{aud+prv} + ISI + demos$. The performance drop is statistically significant, except when compared to the model using ODI3_{SQI}. Table 7.6 confirms this statement.

Table 7.7 shows that using a RF instead of LDA improves accuracy from 70.3% to 80.7% when using $MSE_{aud+prv}$ alone. This is better than the performance achieved by audio-derived features alone or PPG-derived features alone. Again, ISI1 and ISI2 are based solely on audio, and not a combined feature vector; the performance achieved is the same as that stated previously in Table 5.8 (Ac = 68.5%). Again, including demographics actually improves performance. Overall, the best accuracy achieved is 88.0% using $MSE_{aud+prv} + ISI + demos + ODI_{VISI}$. According to Table 7.8, the performance achieved by the four different ODI cal-

culations is worse in all cases when compared to the model using ODI_{VIS1}, indicating that ODI_{VIS1} is the best ODI calculation. It should be noted that there are only very minor differences between the performance metrics achieved by the different ODI calculations, for both classifiers.

Overall the best performance was achieved using a RF to classify MSE_{prv} + ODI_{VIS1}, which can be seen in Table 7.9. Using this combination of features achieved Ac = 89.2%, Se = 87.4% and Sp = 90.5% with an AUC of 0.95. When using ODI_{VIS1} alone to classify the subjects, the performance achieved was Ac = 88.7%, Se = 85.3% and Sp = 90.9%. Table 7.9 clearly shows that using audio features with demographics and ODI_{VIS1} (AUC = 0.96) or PRV MSE features with ODI_{VIS1} (AUC = 0.95) achieves slightly better performance than using ODI_{VIS1} alone. It is clear that, PPG and in particular ODI, should be used to diagnose subjects as requiring treatment or not. However, in cases where it is not possible to record PPG, audio features and demographics can be used to distinguish between treatment and non-treatment subjects with Ac = 81.5% and AUC = 0.88 (Table 5.8).

Figure 7.10 provides some insight into whether audio can correctly identify subjects that are misclassified by $ODI3_b$. Figures 7.10a and 7.10b plot the $ODI3_b$ value versus the RF prediction value for MSE_{aud}; in 7.10a the TP, TN, FP and FN of MSE_{aud} are shown, while in 7.10b the TP, TN, FP and FN of $ODI3_b$ are shown. It is clear that neither MSE_{aud} nor $ODI3_b$ correctly predict all subjects. Figure 7.10c shows in red those cases where MSE_{aud} incorrectly predicts that the subject does not require treatment but $ODI3_b$ correctly predicts this, and in blue those cases where MSE_{aud} incorrectly predicts that the subject requires treatment but $ODI3_b$ correctly predicts that they do not. It is clear that MSE_{aud} predicts more FPs (53) and FNs (78) than $ODI3_b$. However, Figure 7.10d shows the opposite: in red those cases where where $ODI3_b$ incorrectly predicts that the subject does not require treatment but MSE_{aud} correctly predicts this, and in blue those cases where $ODI3_b$ incorrectly predicts that the subject requires treatment but MSE_{aud} correctly predicts that they do not. Although $ODI3_b$ does not predict as many FPs (38) and FNs (25) as MSE_{aud}, there are certain cases where MSE_{aud} is correct and $ODI3_b$ incorrect. Identifying the commonality between the subjects that are correctly identified as requiring treatment by MSE_{aud} could allow for improvement in classification accuracy.

Chapter 8

Discussion & Conclusion

OSA is an under-diagnosed disorder with serious side effects affecting approximately 4% of adults globally. Due to its prevalence, home diagnostic kits have become popular in order to reduce wait times for in-hospital sleep studies. This work focuses on finding a parsimonious and easy to collect set of signals (from the superset of signals used in sleep clinics) and other related information (such as demographics), that can be used to reliably determine which subjects are suitable for standard treatments using a smartphone. It is clear from Chapter 4 that the classic speech analysis techniques are not accurate enough on their own. In addition, the preprocessing required, namely the labelling of individual events, is extremely labour intensive. For this approach to be used in a real application, a detector would be required to identify these events and thus introduce another source of error.

The novelty of this thesis comes from Chapters 5, 6 and 7, where the entire recording is classified using MSE, ISI, demographics and ODI. This removes the need for experts to label the data, and results in an overall diagnosis instead. In Chapter 5, MSE and ISI of the audio signal greatly improves classification accuracy over that achieved using classic speech analysis techniques (accuracy improved from 69.6% to 81.5%). The addition of demographics and ODI improves the accuracy even more, from 81.5% to 88.6%. Applying the same process to actigraphy data shows that, although classifiers trained on features extracted from the actigraphy signal are highly specific, the overall performance statistics are lower than those obtained using audio features. As stated previously, the lack of utility of actigraphy in classification could be due to the preprocessing steps carried out by the Grey Flash device, or that the signal itself is not entirely reliable because it is worn on the upper arm rather than the chest and the

recording device can move relative to the subject during the night.

The ODI used in Chapters 5 and 6 was the one calculated by the Visi-Download software (ODI_{VISI}) using a proprietary algorithm. Chapter 7 investigated methods to improve the ODI with two approaches: with and without the use of SQIs. SQIs were used to remove dips in the saturation signal that were due to noise. It is clear that using SQIs does not improve upon the classification accuracy of ODI_{VISI} (although it did improve upon the calculation using the unfiltered SpO_2 data). This indicates that there is some form of signal quality included in the calculation of ODI_{VISI} already. When ODI_{VISI} is used to classify subjects, it achieves Ac = 88.7%, while $ODI3_{SQI}$ achieves a comparable Ac = 87.1%.

OSA subjects have unusual HR dynamics over the course of the night, related to the sleep stages, so it is expected that PRV would be a useful feature. Applying MSE to PRV data reveals dynamics over multiple time scales as it picks up short term PRV changes as well as long term variations, while also being robust to noise. As Chapter 7 shows, MSE of PRV is quite a successful approach, achieving Ac = 73.1% during validation when used alone, and Ac = 89.2% when combined with ODI. Using audio, PRV and demographic features improves performance from that achieved using audio-derived features alone, but not of PRV-derived features alone. The best accuracy achieved was 89.2% for PRV and ODI, which is a slight improvement on the accuracy achieved when using ODI_{VISI} alone (88.7%). Since the ODI improves the classifier accuracy substantially, it should always be used if affordable and feasible. If this is not the case, then the audio signal is an acceptable alternative. The AUC for $ODI3_b$ and $ODI3_{SQI}$ is 0.93 and 0.94 respectively, while including audio features, PPG features or both in the analysis improves the AUC to 0.96, 0.95 and 0.95 respectively.

Overall, classifying subjects as requiring treatment for OSA, or not, is less time-consuming, and more accurate, than classifying individually annotated events. No event detection is required and all of the features can be generated automatically. The features which displayed the highest predictive accuracy were derived from the PPG (MSE of PRV) and ODI. These features are novel additions to the existing literature for OSA screening. The results reported here are on a testing and training set of 858 subjects which is larger than any that has been reported before, as most analyses carried out on audio data have less than 100 subjects (see section 3.5). The accuracies reported here are comparable with the best reported in the litera-

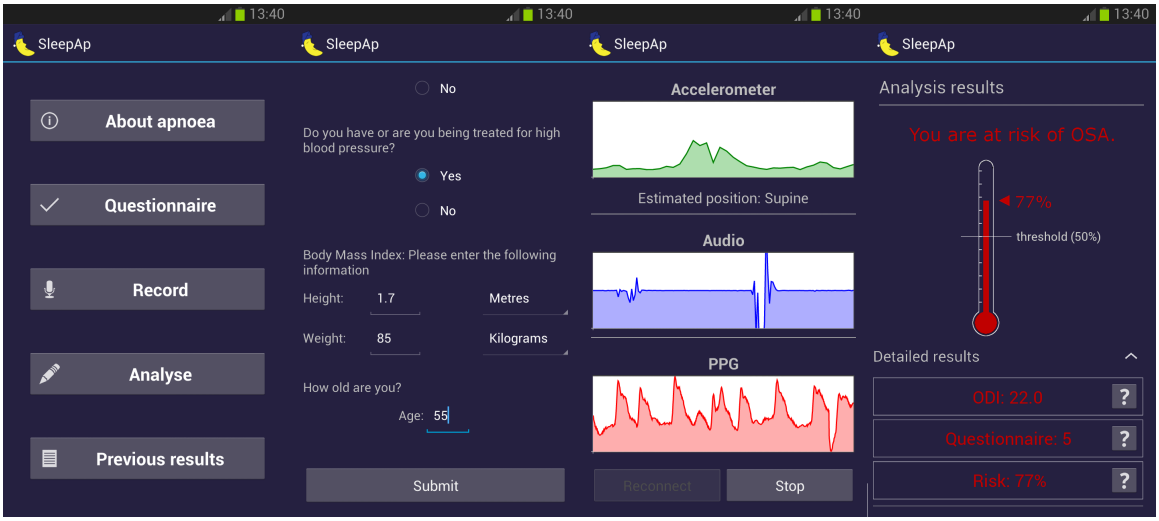
ture. It should be noted that using previously reported techniques on the data in this thesis led to substantially poorer results than for the novel techniques described herein. This indicates that the data used here presents a more difficult OSA patient population to the ones reported in the literature and possibly more representative than the smaller databases used previously.

In summary, the key contributions of this thesis are that it is possible to reliably classify subjects as requiring treatment from entire recordings rather than events, and although ODI improves the classifier accuracy substantially, if the PPG signal cannot be recorded, then the audio signal plus demographics is an acceptable alternative.

8.1 Future Work

Behar *et al.* [181] carried out a review of the smartphone applications (apps) currently available that use the phones' on-board sensors and found that none of them are clinically or scientifically validated. The work in this thesis has therefore been used to develop an Android phone app for screening OSA, called SleepAp. A mobile phone can record audio from the microphone, actigraphy from the on-board accelerometer and PPG from a BlueTooth or USB pulse oximeter. Details of the app can be found in [45, 182]. Figure 8.1 is a screen shot of the app using audio, actigraphy, PPG and the STOP-BANG questionnaire. Ethical approval has been granted for a clinical trial at the Churchill Hospital (Oxford, UK) to adapt thresholds to data collected on a mobile phone, rather than a professional PSG.

Another trial that has begun involves recruiting healthy volunteers with no known sleep problems. This second trial is necessary as the currently model does not include any such subjects. To be usable by the general population, these subjects, and their associated features, have to be included in the analysis. It is possible that an analysis done using these subjects as well as those referred to the hospital will have a different set of features that can be used to predict whether they would need treatment or not, *i.e.* some of the demographic information, such as neck size and BMI, could be more predictive of treatment than in this work.



(a) *Main Menu* **(b)** *Questionnaire* **(c)** *Recording* **(d)** *Analysis*

Figure 8.1: *Screenshots of SleepAp.*

Bibliography

[1] M. J. Thorpy, *The International Classification of Sleep Disorders: diagnostic and coding manual*. Westchester, IL, USA: American Sleep Disorders Association Rochester, MN, 1990.

[2] J. L. Hossain and C. M. Shapiro, "The prevalence, cost implications, and management of sleep disorders: an overview," *Sleep Breath*, vol. 6, no. 2, pp. 85–102, 2002.

[3] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of obstructive sleep apnea: a population health perspective," *Am J Resp Crit Care*, vol. 165, no. 9, pp. 1217–1239, 2002.

[4] A. L. Loomis, E. N. Harvey, and G. Hobart, "Electrical potentials of the human brain," *J Exp Psychol*, vol. 19, pp. 249–279, June 1936.

[5] A. L. Loomis, E. N. Harvey, and G. Hobart, "Cerebral states during sleep, as studied by human brain potentials," *J Exp Psychol*, vol. 21, no. 2, pp. 127–144, 1937.

[6] E. Aserinsky and N. Kleitman, "Regularly occurring periods of eye motility, and concomitant phenomena, during sleep," *Science*, vol. 118, pp. 273–274, Sep 1953.

[7] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Arch Gen Psychiat*, vol. 20, no. 2, p. 246, 1969.

[8] S. E. Martin, H. M. Engleman, R. N. Kingshott, and N. J. Douglas, "Microarousals in patients with sleep apnoea/hypopnoea syndrome," *J Sleep Res*, vol. 6, pp. 276–280, Dec 1997.

[9] Institute of Medicine (US) Committee on Sleep Medicine and Research, *Sleep disorders and sleep deprivation: an unmet public health problem*. National Academies Press, 2006.

[10] D. Pevernagie, R. M. Aarts, and M. De Meyer, "The acoustics of snoring," *Sleep Med Rev*, vol. 14, no. 2, pp. 131–144, 2010.

[11] P. D. Hill, B. W. V. Lee, J. E. Osborne, and E. Z. Osman, "Palatal snoring identified by acoustic crest factor analysis," *Physiol Meas*, vol. 20, no. 2, pp. 167–174, 1999.

[12] N. W. Dorland, *Dorland's Illustrated Medical Dictionary, Deluxe Edition*. Philadelphia, USA: W.B. Saunders Company, 2003.

[13] AASM, "International Classification of Sleep Disorders, 3rd edn.," 2014.

[14] T. Young, L. Evans, L. Finn, and M. Palta, "Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women," *Sleep*, vol. 20, no. 9, pp. 705–706, 1997.

[15] F. Dalmasso and R. Prota, "Snoring: analysis, measurement, clinical implications and applications," *Eur Respir J*, vol. 9, pp. 146–159, Jan 1996.

[16] C. Guilleminault, A. Tilkian, and W. C. Dement, "The sleep apnea syndromes," *Annu Rev Med*, vol. 27, no. 1, pp. 465–484, 1976.

[17] S. Thalhofer and P. Dorow, "Central sleep apnea," *Respiration*, vol. 64, no. 1, pp. 2–9, 1997.

[18] N. Collop, "The effect of obstructive sleep apnea on chronic medical disorders," *Clev Clin J Med*, vol. 74, no. 1, pp. 72–78, 2007.

[19] N. A. Antic, C. Buchan, A. Esterman, M. Hensley, M. T. Naughton, S. Rowland, B. Williamson, S. Windler, S. Eckermann, and R. D. McEvoy, "A randomized controlled trial of nurse-led care for symptomatic moderate-severe obstructive sleep apnea," *Am J Resp Crit Care*, vol. 179, no. 6, pp. 501–508, 2009.

[20] U. P. Ben-Bassey, A. O. Oduwole, and O. O. Ogundipe, "Prevalence of overweight and obesity in Eti-Osa LGA, Lagos, Nigeria," *Obes Rev*, vol. 8, no. 6, pp. 475–479, 2007.

[21] F. Taj, Z. Aly, M. Kassi, and M. Ahmed, "Identifying people at high risk for developing sleep apnea syndrome(SAS): a cross-sectional study in a Pakistani population," *BMC Neurol*, vol. 8, no. 1, pp. 50–59, 2008.

[22] H. Bearpark, L. Elliott, R. Grunstein, S. Cullen, H. Schneider, W. Althaus, and C. Sullivan, "Snoring and sleep apnea. A population study in Australian men," *Am J Resp Crit Care*, vol. 151, no. 5, pp. 1459–1465, 1995.

[23] M. S. Ip, B. Lam, I. J. Lauder, K. W. T. Tsang, K. F. Chung, Y. W. Mok, and W. K. Lam, "A community study of sleep-disordered breathing in middle-aged Chinese men in Hong Kong," *Chest*, vol. 119, no. 1, pp. 62–69, 2001.

[24] M. S. Ip, B. Lam, L. C. Tang, I. J. Lauder, T. Y. Ip, and W. K. Lam, "A community study of sleep-disordered breathing in middle-aged Chinese women in Hong Kong: prevalence and gender differences," *Chest*, vol. 125, no. 1, pp. 127–134, 2004.

[25] J. K. Kim, K. H. In, J. H. Kim, S. H. You, K. H. Kang, J. J. Shim, S. Y. Lee, J. B. Lee, S. G. Lee, C. Park, and C. Shin, "Prevalence of sleep-disordered breathing in middle-aged Korean men and women," *Am J Resp Crit Care*, vol. 170, no. 10, pp. 1108–1113, 2004.

[26] B. Lam, D. C. L. Lam, and M. S. M. Ip, "Obstructive sleep apnoea in Asia," *Int J Tuberc Lung D*, vol. 11, no. 1, pp. 2–11, 2007.

[27] S. K. Sharma, S. Kumpawat, A. Banga, and A. Goel, "Prevalence and risk factors of obstructive sleep apnea syndrome in a population of Delhi, India," *Chest*, vol. 130, no. 1, pp. 149–156, 2006.

[28] Z. F. Udwadia, A. V. Doshi, S. G. Lonkar, and C. I. Singh, "Prevalence of sleep-disordered breathing and sleep apnea in middle-aged urban Indian men," *Am J Resp Crit Care*, vol. 169, no. 2, pp. 168–173, 2004.

[29] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," *N Engl J Med*, vol. 328, no. 17, pp. 1230–1235, 1993.

[30] F. Portier, A. Portmann, P. Czernichow, L. Vascaut, E. Devin, D. Benhamou, A. Cuvelier, and J. Muir, "Evaluation of home versus laboratory polysomnography in the diagnosis of sleep apnea syndrome," *Am J Resp Crit Care*, vol. 162, no. 3, pp. 814–818, 2000.

[31] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to diagnosis and treatment of patients with suspected sleep apnea," *Am J Resp Crit Care*, vol. 169, no. 6, pp. 668–672, 2004.

[32] C. Guilleminault and V. C. Abad, "Obstructive Sleep Apnea," *Curr Treat Option N*, vol. 6, pp. 309–317, Jul 2004.

[33] D. A. Hanzel, N. G. Proia, and D. W. Hudgel, "Response of obstructive sleep apnea to fluoxetine and protriptyline," *Chest*, vol. 100, no. 2, pp. 416–421, 1991.

[34] I. E. Smith and T. G. Quinnell, "Pharmacotherapies for obstructive sleep apnoea: where are we now?," *Drugs*, vol. 64, no. 13, pp. 1385–1399, 2004.

[35] R. C. Heinzer, D. P. White, A. S. Jordan, Y. L. Lo, L. Dover, K. Stevenson, and A. Malhotra, "Trazodone increases arousal threshold in obstructive sleep apnoea," *Eur Respir J*, vol. 31, no. 6, pp. 1308–1312, 2008.

[36] K. A. Ferguson, R. Cartwright, R. Rogers, and W. Schmidt-Nowara, "Oral appliances for snoring and obstructive sleep apnea: a review," *Sleep*, vol. 29, no. 2, pp. 244–262, 2006.

[37] R. Cartwright, D. Stefoski, D. Caldarelli, H. Kravitz, S. Knight, S. Lloyd, and C. Samelson, "Toward a treatment logic for sleep apnea: the place of the tongue retaining device," *Behav Res Ther*, vol. 26, no. 2, pp. 121–126, 1988.

[38] A. D. McGown, H. K. Makker, J. M. Battagel, P. R. L'Estrange, H. R. Grant, and S. G. Spiro, "Long-term use of mandibular advancement splints for snoring and obstructive sleep apnoea: a questionnaire survey," *Eur Respir J*, vol. 17, no. 3, pp. 462–466, 2001.

[39] V. Hoffstein, "Review of oral appliances for treatment of sleep-disordered breathing," *Sleep Breath*, vol. 11, no. 1, pp. 1–22, 2007.

[40] A. Oliven, D. J. O'Hearn, A. Boudewyns, M. Odeh, W. De Backer, P. van de Heyning, P. L. Smith, D. W. Eisele, L. Allan, H. Schneider, *et al.*, "Upper airway response to electrical stimulation of the genioglossus in obstructive sleep apnea," *J Appl Physiol*, vol. 95, no. 5, pp. 2023–2029, 2003.

[41] A. R. Schwartz, D. W. Eisele, A. Hari, R. Testerman, D. Erickson, and P. L. Smith, "Electrical stimulation of the lingual musculature in obstructive sleep apnea," *J Appl Physiol*, vol. 81, no. 2, pp. 643–652, 1996.

[42] A. Oliven, R. P. Schnall, G. Pillar, N. Gavriely, and M. Odeh, "Sublingual electrical stimulation of the tongue during wakefulness and sleep," *Resp Physiol*, vol. 127, no. 2–3, pp. 217–226, 2001.

[43] S. L. Bolea, T. B. Hoegh, B. J. Persson, R. E. Atkinson, S. F. Hauschild, P. M. Kaplan, B. D. Kuhnley, K. E. Jasperson, W. Tesfayesus, and C. K. Thorp, "US patent: 8428727B2," 2013.

[44] A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. D. Clifford, "A review of signals used in sleep analysis," *Physiol Meas*, 2013.

[45] J. Behar, A. Roebuck, M. Shahid, J. Daly, A. H. M. Pureza, N. Palmius, J. Stradling, and G. D. Clifford, "An Evidence Based Android OSA Screening Application," in *Computers in Cardiology, 2013*, Sept 2013.

[46] D. G. Altman and J. M. Bland, "Diagnostic tests 1: Sensitivity and specificity," *Brit Med J*, vol. 308, no. 6943, p. 1552, 1994.

[47] D. G. Altman and J. M. Bland, "Diagnostic tests 2: Predictive values," *Brit Med J*, vol. 309, no. 6947, p. 102, 1994.

[48] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *J Biomed Inform*, vol. 38, no. 5, pp. 404–415, 2005.

[49] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area under a Receiver operating Characteristic (ROC) Curve," *Radiology*, vol. 743, pp. 29–36, 1982.

[50] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recogn*, vol. 30, no. 7, pp. 1145–1159, 1997.

[51] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.

[52] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[53] M. J. Pencina, R. B. D'Agostino, and R. S. Vasan, "Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond," *Stat Med*, vol. 27, no. 2, pp. 157–172, 2008.

[54] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, pp. 1137–1145, 1995.

[55] N. A. Collop, W. McDowell Anderson, B. Boehlecke, D. Claman, R. Goldberg, D. J. Gottlieb, D. Hudgel, M. Sateia, and R. Schwab, "Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients," *J Clin Sleep Med*, vol. 3, no. 7, pp. 737–747, 2007.

[56] D. Álvarez, R. Hornero, D. Abásolo, F. Del Campo, and C. Zamarrón, "Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection," *Physiol Meas*, vol. 27, no. 4, pp. 399–412, 2006.

[57] S. Ancoli-Israel, W. Mason, T. V. Coy, C. Stepnowsky, J. L. Clausen, and J. Dimsdale, "Evaluation of sleep disordered breathing with unattended recording: the Nightwatch System," *J Med Eng Technol*, vol. 21, no. 1, pp. 10–14, 1997.

[58] I. Ayappa, R. G. Norman, M. Suryadevara, and D. M. Rapoport, "Comparison of limited monitoring using a nasal-cannula flow signal to full polysomnography in sleep-disordered breathing," *Sleep*, vol. 27, no. 6, pp. 1171–1179, 2004.

[59] A. Bachour, J. Herrala, and P. Maasilta, "Is there a cost-effective way to diagnose mild sleep-disordered breathing?," *Respir Med*, vol. 96, no. 8, pp. 586–593, 2002.

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

[61] H. Nakano, T. Ikeda, M. Hayashi, E. Ohshima, M. Itoh, N. Nishikata, and T. Shinohara, "Effect of body mass index on overnight oximetry for the diagnosis of sleep apnea," *Respir Med*, vol. 98, no. 5, pp. 421–427, 2004.

[62] H. Schäfer, S. Ewig, E. Hasper, and B. Lüderitz, "Predictive diagnostic value of clinical assessment and nonlaboratory monitoring system recordings in patients with symptoms suggestive of obstructive sleep apnea syndrome," *Respiration*, vol. 64, no. 3, pp. 194–199, 1997.

[63] P. R. Westbrook, D. J. Levendowski, M. Cvetinovic, T. Zavora, V. Velimirovic, D. Henninger, and D. Nicholson, "Description and validation of the apnea risk evaluation system: a novel method to diagnose sleep apnea-hypopnea in the home," *Chest*, vol. 128, no. 4, pp. 2166–2175, 2005.

[64] W. A. Whitelaw, R. F. Brant, and W. W. Flemons, "Clinical usefulness of home oximetry compared with polysomnography for assessment of sleep apnea," *Am J Respir Crit Care Med*, vol. 171, no. 2, pp. 188–193, 2005.

[65] M. Yin, S. Miyazaki, and K. Ishikawa, "Evaluation of type 3 portable monitoring in unattended home setting for suspected sleep apnea: factors that may affect its accuracy," *Otolaryngol Head Neck Surg*, vol. 134, no. 2, pp. 204–209, 2006.

[66] N. Collop, "Portable monitoring in obstructive sleep apnea in adults." `http://www.uptodate.com`, October 2013.

[67] W. W. Flemons, M. R. Littner, J. A. Rowley, P. Gay, W. M. Anderson, D. W. Hudgel, and R. D. M. D. I. Loube, "Home diagnosis of sleep apnea: A systematic review of the literature," *Chest*, vol. 124, no. 4, pp. 1543–1479, 2003.

[68] L. J. Epstein, D. Kristo, J. Patrick J. Strollo, N. Friedman, A. Malhotra, S. P. Patil, K. Ramar, R. Rogers, R. J. Schwab, E. M. Weaver, and M. D. Weinstein, "Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults," *J Clin Sleep Med*, vol. 2009, no. 3, pp. 263–276, 5.

[69] M. W. Johns, "A new method for measuring daytime sleepiness: the Epworth sleepiness scale," *Sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[70] F. Chung, B. Yegneswaran, P. Liao, S. Chung, S. Vairavanathan, S. Islam, A. Khajehdehi, and C. M. Shapiro, "STOP questionnaire: a tool to screen patients for obstructive sleep apnea," *Anesthesiology*, vol. 108, no. 5, pp. 812–821, 2008.

[71] W. W. Flemons and M. A. Reimer, "Development of a disease-specific health-related quality of life questionnaire for sleep apnea," *Am J Resp Crit Care*, vol. 158, no. 2, pp. 494–503, 1998.

[72] N. C. Netzer, R. A. Stoohs, C. M. Netzer, K. Clark, and K. P. Strohl, "Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome," *Ann Intern Med*, vol. 131, no. 7, pp. 485–491, 1999.

[73] J. D. Parkes, S. Y. Chen, S. J. Clift, M. J. Dahlitz, and G. Dunn, "The clinical diagnosis of the narcoleptic syndrome," *J Sleep Res*, vol. 7, no. 1, pp. 41–52, 1998.

[74] R. N. Kingshott, H. M. Engleman, I. J. Deary, and N. J. Douglas, "Does arousal frequency predict daytime function?," *Eur Respir J*, vol. 12, no. 6, pp. 1264–1270, 1988.

[75] Scottish Intercollegiate Guidelines Network, "Management of Obstructive Sleep Apnoea/Hypopnoea Syndrome in Adults." A national clinical guideline. Available from: `http://www.sign.ac.uk/pdf/sign73.pdf`, June 2003.

[76] W. W. Flemons and M. A. Reimer, "Measurement Properties of the Calgary Sleep Apnea Quality of Life Index," *Am J Respir Crit Care Med*, vol. 165, no. 2, pp. 159–164, 2002.

[77] N. Ahmadi, S. A. Chung, A. Gibbs, and C. M. Shapiro, "The Berlin questionnaire for sleep apnea in a sleep clinic population: relationship to polysomnographic measurement of respiratory disturbance," *Sleep Breath*, vol. 12, no. 1, pp. 39–45, 2008.

[78] J. R. Stradling and J. H. Crosby, "Predictors and prevalence of obstuctive sleep apnoea and snoring in 1001 middle-aged men," *Thorax*, vol. 46, no. 2, pp. 85–90, 1991.

[79] O. Osborne, E. Z. Osman, P. D. Hill, B. V. Lee, and C. Sparkes, "A new acoustic method of differentiating palatal from non-palatal snoring," *Clin Otolaryngol*, vol. 24, no. 2, pp. 130–133, 1999.

[80] P. D. Hill, E. Z. Osman, J. E. Osborne, and B. W. V. Lee, "Changes in snoring during natural sleep identified by acoustic crest factor analysis at different times of night," *Clin Otolaryngol*, vol. 25, no. 6, pp. 507–510, 2000.

[81] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J Acoust Soc Am*, vol. 50, no. 2, pp. 637–655, 1971.

[82] J. Solá-Soler, R. Jane, J. A. Fiz, and J. Morera, "Spectral envelope analysis in snoring signals from simple snorers and patients with obstructive sleep apnea," in *25th Annual International Conference of the IEEE EMBS*, vol. 3, pp. 2527–2530, IEEE, 2003.

[83] A. K. Ng, T. S. Koh, E. Baey, T. H. Lee, U. R. Abeyratne, and K. Puvanendran, "Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea?," *Sleep Med*, vol. 9, no. 8, pp. 894–898, 2008.

[84] A. Yadollahi and Z. Moussavi, "Formant analysis of breath and snore sounds," in *31st Annual International Conference of the IEEE EMBS*, pp. 2563–2566, IEEE, 2009.

[85] J. A. Fiz, J. Abad, R. Jané, M. Riera, M. A. Mananas, P. Caminal, D. Rodenstein, and J. Morera, "Acoustic analysis of snoring sound in patients with simple snoring and obstructive sleep apnoea," *Eur Respir J*, vol. 9, no. 11, pp. 2365–2370, 1996.

[86] A. W. McCombe, V. Kwok, and W. M. Hawke, "An acoustic screening test for obstructive sleep apnoea," *Clin Otolaryngol*, vol. 20, no. 4, pp. 348–351, 1995.

[87] N. Meslier, Y. Auregan, A. Badatcheff, C. Depollier, and J. L. Racineux, "Spectral analysis of snores in patients with obstructive sleep apnoea syndrome (abst.)," *Am Rev Respir Dis*, vol. 141, p. A857, 1990.

[88] J. R. Perez-Padilla, E. Slawinski, L. M. Difrancesco, R. R. Feige, J. E. Remmers, and W. A. Whitelaw, "Characteristics of the snoring noise in patients with and without occlusive sleep apnea," *Am Rev Respir Dis*, vol. 147, no. 3, pp. 635–644, 1993.

[89] H. Hara, N. Murakami, Y. Miyauchi, and H. Yamashita, "Acoustic analysis of snoring sounds by a multidimensional voice program," *Laryngoscope*, vol. 116, no. 3, pp. 379–381, 2006.

[90] M. Herzog, A. Schmidt, T. Bremert, B. Herzog, W. Hosemann, and H. Kaftan, "Analysed snoring sounds correlate to obstructive sleep disordered breathing," *Eur Arch Oto-Rhino-L*, vol. 265, no. 1, pp. 105–113, 2008.

[91] T. H. Lee, U. R. Abeyratne, K. Puvanendran, and K. L. Goh, "Formant-structure and phase-coupling analysis of human snoring sounds for detection of obstructive sleep apnea," in *Comput Method Biome*, vol. 3, pp. 243–248, 2001.

[92] U. R. Abeyratne, A. S. Karunajeewa, and C. Hukins, "Mixed-phase modeling in snore sound analysis," *Med Biol Eng Comput*, vol. 45, no. 8, pp. 791–806, 2007.

[93] A. K. Ng, K. Y. Wong, C. H. Tan, and T. S. Koh, "Bispectral analysis of snore signals for obstructive sleep apnea detection," in *29th Annual International Conference of the IEEE EMBS*, pp. 6195–6198, 2007.

[94] A. K. Ng, T. San Koh, K. Puvanendran, and U. R. Abeyratne, "Snore signal enhancement and activity detection via translation-invariant wavelet transform," *IEEE Trans Biomed Eng*, vol. 55, no. 10, pp. 2332–2342, 2008.

[95] D. Matsiki, X. Deligianni, E. Vlachogianni-Daskalopoulou, and L. J. Hadjileontiadis, "Wavelet-based analysis of nocturnal snoring in apneic patients undergoing polysomnography," in *29th Annual International Conference of the IEEE EMBS*, pp. 1912–1915, 2007.

[96] W. D. Duckitt, S. K. Tuomi, and T. R. Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data," *Physiol Meas*, vol. 27, no. 10, pp. 1047–1056, 2006.

[97] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans on Audio and Electroacoustics*, vol. 28, no. 4, pp. 357–366, 1980.

[98] M. Cavusoglu, M. Kamasak, O. Erogul, T. Ciloglu, Y. Serinagaoglu, and T. Akcam, "An efficient method for snore/nonsnore classification of sleep sounds," *Physiol Meas*, vol. 28, no. 8, pp. 841–853, 2007.

[99] T. M. Jones, A. C. Swift, P. M. A. Calverley, M. S. Ho, and J. E. Earis, "Acoustic analysis of snoring before and after palatal surgery," *Eur Respir J*, vol. 25, no. 6, pp. 1044–1049, 2005.

[100] T. M. Jones, P. Walker, M. Ho, J. E. Earis, A. C. Swift, and P. Charters, "Acoustic parameters of snoring sound to assess the effectiveness of sleep nasendoscopy in predicting surgical outcome," *Otolaryng Head Neck*, vol. 135, pp. 269–275, Aug 2006.

[101] T. M. Jones, M. S. Ho, J. E. Earis, and A. C. Swift, "Acoustic parameters of snoring sound to assess the effectiveness of the Müller Manoeuvre in predicting surgical outcome," *Auris Nasus Larynx*, vol. 33, no. 4, pp. 409–416, 2006.

[102] M. B. Pringle and C. B. Croft, "A grading system for patients with obstructive sleep apnoea-based on sleep nasendoscopy," *Clin Otolaryngol*, vol. 18, no. 6, pp. 480–484, 1993.

[103] A. E. Camilleri, L. Ramamurthy, and P. H. Jones, "Sleep nasendoscopy: what benefit to the management of snorers?," *J Laryngol Otol*, vol. 109, no. 12, pp. 1163–1165, 1995.

[104] U. R. Abeyratne, A. S. Wakwella, and C. Hukins, "Pitch jump probability measures for the analysis of snoring sounds in apnea," *Physiol Meas*, vol. 26, no. 5, pp. 779–798, 2005.

[105] U. R. Abeyratne, C. K. Patabandi, and K. Puvanendran, "Pitch-jitter analysis of snoring sounds for the diagnosis of sleep apnea," in *23rd Annual International Conference of the IEEE EMBS*, vol. 2, pp. 2072–2075, IEEE, 2001.

[106] J. M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications," *P IEEE*, vol. 79, no. 3, pp. 278–305, 1991.

[107] J. W. Shepard Jr, W. B. Gefter, C. Guilleminault, E. A. Hoffman, V. Hoffstein, D. W. Hudgel, P. M. Suratt, and D. P. White, "Evaluation of the upper airway in patients with obstructive sleep apnea," *Sleep*, vol. 14, no. 4, pp. 361–371, 1991.

[108] T. N. Liesching, C. Carlisle, A. Marte, A. Bonitati, and R. P. Millman, "Evaluation of the Accuracy of SNAP Technology Sleep Sonography in Detecting Obstructive Sleep Apnea in Adults Compared to Standard Polysomnography," *Chest*, vol. 125, no. 3, pp. 886–891, 2004.

[109] P. G. Michaelson, P. Allan, J. Chaney, and E. A. Mair, "Validations of a portable home sleep study with twelve-lead polysomnography: comparisons and insights into a variable gold standard," *Ann Oto Rhinol Laryn*, vol. 115, no. 11, pp. 802–809, 2006.

[110] S. Su, F. M. Baroody, M. Kohrman, and D. Suskind, "A comparison of polysomnography and a portable home sleep study in the diagnosis of obstructive sleep apnea syndrome," *Otolaryng Head Neck*, vol. 131, no. 6, pp. 844–850, 2004.

[111] C. Galer, A. Yonkers, W. Duff, and B. Heywood, "Clinical significance of SNAP somnography test acoustic recording," *Otolaryng Head Neck*, vol. 136, no. 2, pp. 241–245, 2007.

[112] J. Solá-Soler, J. A. Fiz, J. Morera, and R. Jané, "Multiclass classification of subjects with sleep apnoea–hypopnoea syndrome through snoring analysis," *Med Eng Phys*, vol. 34, no. 9, pp. 1213–1220, 2012.

[113] N. Ben-Israel, A. Tarasiuk, and Y. Zigel, "Obstructive apnea hypopnea index estimation by analysis of nocturnal snoring signals in adults," *Sleep*, vol. 35, no. 9, p. 1299, 2012.

[114] J. A. Fiz, R. Jané, J. Solá-Soler, J. Abad, M. García, and J. Morera, "Continuous analysis and monitoring of snores and their relationship to the apnea-hypopnea index," *Laryngoscope*, vol. 120, no. 4, pp. 854–862, 2010.

[115] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms. American Academy of Sleep Medicine Review Paper," *Sleep*, vol. 26, no. 3, pp. 342–392, 2003.

[116] J. B. Webster, D. F. Kripke, S. Messin, D. J. Mullaney, and G. Wyborney, "An activity-based sleep monitor system for ambulatory use," *Sleep*, vol. 5, no. 4, pp. 389–399, 1982.

[117] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, no. 5, pp. 461–469, 1992.

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

[119] L. de Souza, A. A. Benedito-Silva, M. N. Pires, D. Poyares, S. Tufik, H. M. Calil, *et al.*, "Further validation of actigraphy for sleep studies," *Sleep*, vol. 26, no. 1, pp. 81–85, 2003.

[120] J. Paquet, A. Kawinska, and J. Carrier, "Wake detection capacity of actigraphy during sleep," *Sleep*, vol. 30, no. 10, pp. 1362–1369, 2007.

[121] J. Lötjönen, I. Korhonen, K. Hirvonen, S. Eskelinen, M. Myllymäki, and M. Partinen, "Automatic sleep-wake and nap analysis with a new wrist worn online activity monitoring device Vivago Wristcare®," *Sleep*, vol. 26, no. 1, pp. 86–90, 2003.

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

[123] M. Drinnan, A. Murray, J. White, A. Smithson, G. Gibson, and C. Griffiths, "Evaluation of activity-based techniques to identify transient arousal in respiratory sleep disorders," *J Sleep Res*, vol. 5, no. 3, pp. 173–180, 1996.

[124] A. Sadeh, "The role and validity of actigraphy in sleep medicine: an update," *Sleep Med Rev*, vol. 15, no. 4, pp. 259–267, 2011.

[125] H. A. Middelkoop, A. Knuistingh Neven, J. J. Van Hilten, C. W. Ruwhof, and H. A. Kamphuisen, "Wrist actigraphic assessment of sleep in 116 community based subjects suspected of obstructive sleep apnoea syndrome," *Brit Med J*, vol. 50, no. 3, p. 284, 1995.

[126] M. Elbaz, G. M. Roue, F. Lofaso, and M. A. Salva, "Utility of actigraphy in the diagnosis of obstructive sleep apnea," *Sleep*, vol. 25, no. 5, pp. 527–531, 2002.

[127] N. T. Ayas, S. Pittman, M. MacDonald, and D. P. White, "Assessment of a wrist-worn device in the detection of obstructive sleep apnea," *Sleep Med*, vol. 4, no. 5, pp. 435–442, 2003.

[128] J. Hedner, G. Pillar, S. D. Pittman, D. Zou, L. Grote, and D. P. White, "A novel adaptive wrist actigraphy algorithm for sleep-wake assessment in sleep apnea patients," *Sleep*, vol. 27, no. 8, pp. 1560–1566, 2004.

[129] M. J. Kim, G.-H. Lee, C.-S. Kim, W. S. Kim, Y.-S. Chung, S. Chung, and S.-A. Lee, "Comparison of three actigraphic algorithms used to evaluate sleep in patients with obstructive sleep apnea," *Sleep Breath*, vol. 17, no. 1, pp. 297–304, 2013.

[130] M. Littner, C. A. Kushida, W. M. Anderson, D. Bailey, R. B. Berry, D. G. Davila, M. Hirshkowitz, S. Kapen, M. Kramer, D. Loube, M. Wise, and S. F. Johnson, "Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002," *Sleep*, vol. 26, no. 3, pp. 337–341, 2003.

[131] A. Sadeh and C. Acebo, "The role of actigraphy in sleep medicine," *Sleep Med Rev*, vol. 6, no. 2, pp. 113–124, 2002.

[132] K. Chin, T. Oga, K.-i. Takahashi, M. Takegami, Y. Nakayama-Ashida, T. Wakamura, K. Sumi, T. Nakamura, S. Horita, Y. Oka, *et al.*, "Associations between obstructive sleep apnea, metabolic syndrome, and sleep duration, as measured with an actigraph, in an urban male working population in Japan," *Sleep*, vol. 33, no. 1, pp. 89–95, 2010.

[133] T. Morgenthaler, C. Alessi, L. Friedman, J. Owens, V. Kapur, B. Boehlecke, T. Brown, A. C. Jr., J. Coleman, T. Lee-Chiong, J. Pancer, and T. J. Swick, "Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: An update for 2007," *Sleep*, vol. 30, no. 4, pp. 519–529, 2007.

[134] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol Meas*, vol. 28, no. 3, pp. R1–R39, 2007.

[135] A. H. Khandoker, C. K. Karmakar, and M. Palaniswami, "Comparison of pulse rate variability with heart rate variability during obstructive sleep apnea," *Med Eng Phys*, vol. 33, no. 2, pp. 204–209, 2011.

[136] S. G. Fleming and L. Tarassenko, "A comparison of signal processing techniques for the extraction of breathing rate from the photoplethysmogram," *International Journal of Biological and Medical Sciences*, vol. 2, no. 4, pp. 232–236, 2007.

[137] M. O. Mendez, E. Gil, J. M. Vergara, S. Cerutti, A. M. Bianchi, and P. Laguna, "Relationship among envelope fluctuations in PPG, HRV and apnea," in *5th Conference of the European Study Group on Cardiovascular Oscillations*, 2008.

[138] C. G. Scully, J. Lee, J. Meyer, A. M. Gorbach, D. Granquist-Fraser, Y. Mendelson, and K. H. Chon, "Physiological parameter monitoring from optical recordings with a mobile phone," *IEEE Trans Biomed Eng*, vol. 59, no. 2, pp. 303–306, 2012.

[139] J. A. Bennett and W. J. M. Kinnear, "Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome," *Thorax*, vol. 54, no. 11, pp. 958–959, 1999.

[140] N. C. Netzer, A. H. Eliasson, C. Netzer, and D. A. Kristo, "Overnight pulse oximetry for sleep-disordered breathing in adults," *Chest*, vol. 120, no. 2, pp. 625–633, 2001.

[141] E. Chiner, J. Signes-Costa, J. M. Arriero, J. Marco, I. Fuentes, and A. Sergado, "Nocturnal oximetry for the diagnosis of the sleep apnoea hypopnoea syndrome: a method to reduce the number of polysomnographies?," *Thorax*, vol. 54, no. 11, pp. 968–971, 1999.

[142] I. Fietze, K. Dingli, K. Diefenbach, N. J. Douglas, M. Glos, M. Tallafuss, W. Terhalle, and C. Witt, "Night-to-night variation of the oxygen desaturation index in sleep apnoea syndrome," *Eur Respir J*, vol. 24, no. 6, pp. 987–993, 2004.

[143] J. C. Vázquez, W. H. Tsai, W. W. Flemons, A. Masuda, R. Brant, E. Hajduk, W. A. Whitelaw, and J. E. Remmers, "Automated analysis of digital oximetry in the diagnosis of obstructive sleep apnoea," *Thorax*, vol. 55, no. 4, pp. 302–307, 2000.

[144] H. M. Al-Angari and A. V. Sahakian, "Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier," *IEEE T Inf Technol B*, vol. 16, no. 3, pp. 463–468, 2012.

[145] D. Álvarez, R. Hornero, J. V. Marcos, and F. del Campo, "Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis," *IEEE Trans Biomed Eng*, vol. 57, no. 12, pp. 2816–2824, 2010.

[146] R. Hornero, D. Álvarez, D. Abásolo, F. del Campo, and C. Zamarrón, "Utility of approximate entropy from overnight pulse oximetry data in the diagnosis of the obstructive sleep apnea syndrome," *IEEE Trans Biomed Eng*, vol. 54, no. 1, pp. 107–113, 2007.

[147] D. S. Morillo, J. L. Rojas, L. F. Crespo, A. León, and N. Gross, "Poincaré analysis of an overnight arterial oxygen saturation signal applied to the diagnosis of sleep apnea hypopnea syndrome," *Physiol Meas*, vol. 30, no. 4, pp. 405–420, 2009.

[148] L. M. Sepúlveda-Cano, E. Gil, P. Laguna, and G. Castellanos-Dominguez, "Sleep apnoea detection in children using PPG envelope-based dynamic features," in *33rd Annual International Conference of the IEEE EMBS*, pp. 1483–1486, IEEE, 2011.

[149] E. Gil, V. Monasterio, P. Laguna, and J. M. Vergara, "Pulse photopletismography amplitude decrease detector for sleep apnea evaluation in children," in *27th Annual International Conference of the IEEE EMBS*, pp. 2743–2746, 2006.

[150] E. Gil, M. Mendez, J. M. Vergara, S. Cerutti, A. M. Bianchi, and P. Laguna, "Discrimination of sleep-apnea-related decreases in the amplitude fluctuations of PPG signal in children by HRV analysis," *IEEE Trans Biomed Eng*, vol. 56, no. 4, pp. 1005–1014, 2009.

[151] E. Gil, R. Bailón, J. M. Vergara, and P. Laguna, "PTT variability for discrimination of sleep apnea related decreases in the amplitude fluctuations of PPG signal in children," *IEEE Trans Biomed Eng*, vol. 57, no. 5, pp. 1079–1088, 2010.

[152] V. Monasterio, F. Burgess, and G. D. Clifford, "Robust classification of neonatal apnoea-related desaturations," *Physiol Meas*, vol. 33, no. 9, pp. 1503–1516, 2012.

[153] Stowood Scientific Instruments, "Private communication," 2014.

[154] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and Trends in Signal Processing*, vol. 1, no. 1, pp. 1–194, 2007.

[155] S. Saito and F. Itakura, "The theoretical consideration of statistically optimum methods for speech spectral density," *Electrical Communication Laboratory, NTT, Tokyo, Rep*, vol. 3107, 1966.

[156] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, pp. 209–243, 1963.

[157] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," *Adv Neur In*, vol. 17, pp. 1569–1576, 2004.

[158] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, 1995.

[159] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[160] M. H. Kryger, T. Roth, and W. C. Dement, *Principles and practice of sleep medicine: 5th edition*. W.B. Saunders Company, 2011.

[161] P. Grassberger, "Information and complexity measures in dynamical systems," in *Information Dynamics*, Springer, 1991.

[162] J. R. Stradling, "Private communication," September 2010.

[163] AASM, "Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research," *Sleep*, vol. 22, no. 5, pp. 667–689, 1999.

[164] J. L. Engstrom, S. A. Paterson, A. Doherty, M. Trabulsi, and K. L. Speer, "Accuracy of self-reported height and weight in women: An integrative review of the literature," *J Midwifery Wom Heal*, vol. 48, no. 5, pp. 338–345, 2003.

[165] J. Vrhovec, "Evaluating the progress of the labour with sample entropy calculated from the uterine EMG activity," *Electrotechnical Review*, vol. 79, pp. 165–170, 2009.

[166] M. Costa, C. K. Peng, A. L. Goldberger, and J. M. Hausdorff, "Multiscale entropy analysis of human gait dynamics," *Physica A*, vol. 330, no. 1-2, pp. 53–60, 2003.

[167] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[168] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Phys Rev Lett*, vol. 89, no. 6, pp. 068102–1–068102–4, 2002.

[169] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy to distinguish physiologic and synthetic RR time series," in *Computers in Cardiology, 2002*, pp. 137–140, 2002.

[170] A. Kales, R. J. Cadieux, E. O. Bixler, C. R. Soldatos, A. Vela-Bueno, C. A. Misoul, and T. W. Locke, "Severe obstructive sleep apnea-I: onset, clinical course, and characteristics," *J Chronic Dis*, vol. 38, no. 5, pp. 419–425, 1985.

[171] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.

[172] A. E. W. Johnson, N. Dunkley, L. Mayaud, A. Tsanas, A. A. Kramer, and G. D. Clifford, "Patient specific predictions in the intensive care unit using a bayesian ensemble," in *Computing in Cardiology (CinC), 2012*, pp. 249–252, IEEE, 2012.

[173] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J Chem Phys*, vol. 21, p. 1087, 1953.

[174] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[175] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans Biomed Eng*, vol. BME-32, no. 3, pp. 230–236, 1985.

[176] J. A. Sukor, S. J. Redmond, and N. H. Lovell, "Signal quality measures for pulse oximetry through waveform morphology analysis," *Physiol Meas*, vol. 32, no. 3, pp. 369–384, 2011.

[177] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of biological signals," *Physical Review E*, vol. 71, no. 2, pp. 021906–1 – 021906–18, 2005.

[178] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiol Meas*, vol. 33, no. 9, pp. 1491–1501, 2012.

[179] W. Zong, T. Heldt, G. B. Moody, and R. G. Mark, "An open-source algorithm to detect onset of arterial blood pressure pulses," in *Computers in Cardiology, 2003*, pp. 259–262, IEEE, 2003.

[180] G. D. Clifford, P. E. McSharry, and L. Tarassenko, "Characterizing artefact in the normal human 24-hour RR time series to aid identification and artificial replication of circadian variations in human beat to beat heart rate using a simple threshold," in *Computers in Cardiology, 2002*, pp. 129–132, 2002.

[181] J. Behar, A. Roebuck, J. S. Domingos, E. Gederi, and G. D. Clifford, "A review of current sleep screening applications for smartphones," *Physiol Meas*, vol. 34, no. 7, pp. R29–R46, 2013.

[182] J. Behar, A. Roebuck, M. Shahid, J. Daly, A. H. M. Pureza, N. Palmius, J. Stradling, and G. D. Clifford, "SleepAp: An Automated Obstructive Sleep Apnoea Screening Application for Smartphones," *submitted to IEEE J Biomed Health Inform*, 2013.

Appendix A

Glossary of Terms

AASM : American academy of sleep medicine
Ac: accuracy
AHI: apnoea-hypopnoea index
AI: apnoea index
APAP: autopositive airway pressure
app: application
AR: autoregressive
AS: apnoeic snores
AUC: area under the curve
BiPAP: bilevel positive airway pressure
BMI: body mass index
BQ: Berlin questionnaire
BS: benign snores
CPAP: continuous positive airway pressure
CSA: central sleep apnoea
CSAQLI: Calgary sleep apnoea quality of life index
DAP: decreases in the amplitude of the PPG signal
DFT: discrete Fourier transform
DTFT: discrete time Fourier transform
DTW: dynamic time warping
ECG: electrocardiogram
EEG: electroencephalogram
EMG: electromyogram
EOG: electrooculogram
ESS: Epworth sleepiness score
FFT: fast Fourier transform
FN: false negative
FP: false positive
$\mathcal{H}_A$: approximate entropy
$\mathcal{H}_S$: sample entropy
HMM: hidden Markov model
HR: heart rate
HRV: heart rate variability
ICSD: international classification of sleep disorders
IDI: integrated discriminative improvement
ISI: inter-snore interval
LDA: linear discriminant analysis
LPC: linear predictive coding
LR: likelihood ratio
MAD: mandibular advancement devices
MCMC: Markov chain Monte Carlo
MFCCs: mel-frequency cepstral coefficients
MSE: multiscale entropy
NHR: noise to harmonics ratio
NPV: negative predictive value/negative predictivity
NREM: non rapid eye movement
NRI: net reclassification index
OA: oral appliance
ODI: oxygen desaturation index
OSA: obstructive sleep apnoea
PAT: peripheral arterial tonometry
PM: portable monitor/portable monitoring
PP: intervals between pulses in the PPG sig-

nal
PPG: photoplethysmogram/ photoplethysmography
PPV: positive predictive value/positive predictivity
PHR: PPG derived heart rate
PR: pulse rate
PSG: polysomnogram/polysomnography
PTTV: pulse transit time variability
R&K: Rechtschaffen & Kales
RDI: respiratory disturbance index
REM: rapid eye movement
RF: random forest
RMS: root mean square
ROC: receiver operating curve
RR: intervals between R peaks in the ECG signal
SAS: sleep apnoea syndrome
SDB: sleep disordered breathing
SE: sleep efficiency
Se: sensitivity
se: standard error
SNR: signal to noise ratio
Sp: specificity
SPI: soft phonation index
SPL: sound pressure level
SpO_2: non-invasive measure of blood oxygen saturation
SQI: signal quality index
SVM: support vector machine
TN: true negative
TP: true positive
TRD: tongue retaining device
TST: total sleep time
UA: upper airway
UCSD: University of California San Diego

Appendix B

Journal Articles

The journal and conference papers associated with this work are detailed below. A number of them are in preparation, and have journal targets indicated, while the remainder have been published or are in submission.

- A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel and G.D. Clifford, "A review of signals used in sleep analysis", *Physiol Meas*, vol. 35, no. 1, pp. R1-R57, 2014. Associated with Chapter 1.

- J. Behar, A. Roebuck, J. Domingos, E. Gederi and G.D. Clifford, "A review of current sleep screening applications for smartphones", *Physiol Meas*, vol. 34, no. 7, pp. R29-R46, 2013. Associated with Chapter 8.

- J. Behar, A. Roebuck, M. Shahid, J. Daly, A. Hallack Miranda Pureza, N. Palmius, J. Stradling and G.D. Clifford, "An Evidence Based Android OSA Screening Application", *presented at and appears in the proceedings of Computers in Cardiology, 2013*, Sept 2013. Associated with Chapter 8.

- J. Behar, A. Roebuck, M. Shahid, J. Daly, A. Hallack Miranda Pureza, N. Palmius, J. Stradling and G.D. Clifford, "SleepAp: An Automated Obstructive Sleep Apnoea Screening Application for Smartphones", *in press: IEEE J Health Inform*, 2013. Associated with Chapter 8.

- A. Roebuck and G.D. Clifford, "Optimisation of AHI and ODI thresholds for identifying treatable OSA", *in preparation for Physiol Meas*. Associated with Chapter 5.

- A. Roebuck and G.D. Clifford, "Classifying OSA subjects using entire audio recordings", *in preparation for IEEE Trans Biomed Eng*. Associated with Chapter 5.

- A. Roebuck and G.D. Clifford, "Fusing audio, actigraphy and PPG to improve OSA classification performance", *in preparation for IEEE Trans Biomed Eng*. Associated with Chapters 6 and 7.

Appendix C

Questionnaires

C.1 Epworth Sleepiness Score

The following questionnaire will help you measure your general level of daytime sleepiness. You are to rate the chance that you would *doze off or fall asleep* during different routine daytime situations. Answers to the questions are rated on a reliable scale called the Epworth Sleepiness Scale (ESS). Each item is rated from 0 to 3, with 0 meaning you would never doze or fall asleep in a given situation and 3 meaning that there is a very high chance that you would doze or fall asleep in that situation.

How likely are you to *doze off or fall asleep* in the following situations, in contrast to just feeling tired? Even if you haven't done some of the activities recently, think about how they would have affected you. Use the following scale to choose the most appropriate number for each situation:
0 = Would never doze
1 = Slight chance of dozing
2 = Moderate chance of dozing
3 = High chance of dozing

Situation	Chance of Dozing
Sitting and reading Watching TV Sitting, inactive in a public place (*e.g.* in a theatre or a meeting) As a passenger in a car for an hour without a break Lying down to rest in the afternoon when circumstances permit Sitting and talking to someone Sitting quietly after a lunch without alcohol In a car, while stopped for a few minutes in the traffic	
Total	

Score:
0-10 Normal range
10-12 Borderline
12-24 Abnormal

C.2 STOP BANG

Answer the following questions to find out if you are at risk of Obstructive Sleep Apnoea.

1. **S**noring
   Do you snore loudly (louder than talking or loud enough to be heard through closed doors)?
   Yes No

2. **T**ired
   Do you often feel tired, fatigued, or sleepy during daytime?
   Yes No

3. **O**bserved
   Has anyone observed you stop breathing during your sleep?
   Yes No

4. Blood **P**ressure
   Do you have or are you being treated for high blood pressure?
   Yes No

5. **B**MI
   BMI more than 35 kg/m^2?
   Yes No

6. **A**ge
   Age over 50 yr old?
   Yes No

7. **N**eck circumference
   Neck circumference greater than 40 cm?
   Yes No

8. **G**ender
   Gender male?
   Yes No

High risk of OSA: answering yes to three or more items
Low risk of OSA: answering yes to less than three items

C.3 CSAQLI

This questionnaire has been designed to find out how you have been doing and feeling over the last 4 weeks. You will be questioned about the impact that sleep apnea and/or snoring may have had on your daily activities, your emotional functioning, and your social interactions, and about any symptoms they might have caused.

1. Daily Functioning

   (a) Most important daily activity. With regard to performing your most important, usual daily activity (*e.g.* work, school, child care, housework, etc.) during the previous 4 weeks:

      i. How much have you had to force yourself to do this activity? [yellow card]

      ii. How much of the time have you had to push yourself to remain alert while performing this activity? [yellow card]

      iii. How often have you adjusted your schedule to avoid this activity because you felt that you would be unable to remain alert while doing it? [yellow card]

      iv. How often do you use all of your energy to accomplish only this activity? [yellow card]

   (b) Secondary activities. With regard to activities other than your most important daily activity during the previous 4 weeks:

i. How much difficulty have you had finding the energy to exercise and/or do activities that you find relaxing (leisure activities)? [green card]
ii. How much difficulty have you had finding the time for activities that you find relaxing? [green card]
iii. How much difficulty have you had with your ability to do exercise and/or activities that you find relaxing? [green card]
iv. How much difficulty have you had getting chores done around the place where you live? [green card]

(c) General functioning. During the previous 4 weeks:
i. How much difficulty have you had with trying to remember things? [green card]
ii. How much difficulty have you had with trying to concentrate? [green card]
iii. How much of a problem have you had with having to fight to stay awake? [red card]

(d) Social Interactions
The following questions pertain to how your relationship with your partner, other household members, relatives, and/or close friends have been during the previous 4 weeks. If you have not interacted with a partner, etc. in the previous 4 weeks, please try to work out how your relationship might have been with these people.
i. How upset have you been about being told that your snoring was bothersome or irritating? [green card]
ii. How upset have you been about having to (or possibly having to) sleep in separate bedrooms from your partner? [green card]
iii. How upset have you been as a result of frequent conflicts or arguments? [green card]
iv. How aware have you been of not wanting to talk to other people? [green card]
v. How much concern have you had about the need to make special sleeping arrangements if you were travelling and/or staying with someone? [green card]
vi. How guilty have you felt about your relationship with family members or close personal friends? [green card]
vii. How often have you looked for excuses for being tired? [yellow card]
viii. How often have you experienced wanting to be left alone?
ix. How often have you felt like not wanting to do things together with your partner, children, and/or friends? [yellow card]
x. How much of a problem have you felt there is with your relationship to the person who is closest to you? [red card]
xi. How much of a problem have you had from not being involved in family activities? [red card]
xii. How much of a problem have you had with inadequate and/or infrequent sexual intimacy? [red card]
xiii. How much of a problem have you had with a lack of interest in being around other people? [red card]

(e) Emotional Functioning
With respect to how you have been feeling inside during the previous 4 weeks:
i. How often have you been feeling depressed, down, and/or hopeless? [yellow card]
ii. How often have you been feeling anxious or fearful about what was wrong? [yellow card]
iii. How often have you been feeling frustrated? [yellow card]
iv. How often have you been feeling irritable and/or moody? [yellow card]
v. How often have you been feeling impatient? [yellow card]

vi. How often have you been feeling that you are being unreasonable? [yellow card]

vii. How often have you been getting easily upset? [yellow card]

viii. How often have you experienced a tendency to become angry? [yellow card]

ix. How often have you been feeling like you were unable to cope with everyday issues? [yellow card]

x. How concerned have you been about your weight? [green card]

xi. How concerned have you been about heart problems (heart attacks or heart failure) and/or premature death? [green card]

(f) Symptoms

Below is a list of symptoms that some people with sleep apnea and/or who snore may experience. As each symptom is read please indicate whether it has been a problem or not (answer yes or no). Circle those symptoms that you have experienced during the previous 4 weeks. Once the list is finished please write down additional symptoms in the blank spaces you may have had that are not included in the list below. Next select the five most important symptoms you have experienced. For each of the five symptoms please identify how much of a problem it has been. [red card]

i. Decreased energy

ii. Excessive fatigue

iii. Feeling that ordinary activities require an extra effort to perform or complete

iv. Falling asleep at inappropriate times or places

v. Falling asleep if not stimulated or active

vi. Difficulty with a dry or sore mouth/throat upon awakening

vii. Waking up often (more than twice) during the night

viii. Difficulty returning to sleep if you wake up in the night

ix. Concern about the times you stop breathing at night

x. Waking up at night feeling like you were choking

xi. Waking up in the morning with a headache

xii. Waking up in the morning feeling unrefreshed and/or tired

xiii. Waking up more than once per night to urinate

xiv. A feeling that your sleep is restless

xv. Difficulty staying awake while reading

xvi. Difficulty staying awake while trying to carry on a conversation

xvii. Difficulty staying awake while trying to watch something (concert, movie, TV)

xviii. Fighting the urge to fall asleep while driving

xix. A reluctance or inability to drive for > 1 h

xx. Concern regarding close calls while driving due to your inability to remain alert

xxi. Concern regarding your or other's safety when you're operating a motor vehicle or machinery

xxii.

xxiii.

(g) Treatment-related Symptoms

If you haven't had some type of therapy for sleep apnea and/or snoring leave this section blank. Below is a list of symptoms that some people who have been treated for sleep apnea and/or snoring may experience. As each symptom is read please indicate whether it has been a problem or not (answer yes or no). Circle those symptoms that you have experienced during the previous 4 weeks. Once the list is finished please write down any symptoms in the blank spaces you may have had that are not included in the list below. Next select the five most important symptoms you have experienced. For each of the five symptoms please identify

how much of a problem it has been. [red card]

i. Runny nose
ii. Stuffed or congested or blocked nose
iii. Excessive dryness of the nose or throat passages, especially upon awakening
iv. Soreness in the nose or throat passages
v. Headaches
vi. Eye irritation
vii. Ear pain
viii. Waking up frequently during the night
ix. Difficulty returning to sleep if you awaken
x. Air leakage from the nasal mask
xi. Discomfort from the nasal mask
xii. Marks or rash on your face
xiii. Complaints from your partner about the noise of the CPAP machine
xiv. Having fluid/food pass into your nose when you swallow
xv. A change in how your voice sounds
xvi. Pain in the throat when swallowing
xvii. Pain or aching in your jaw joint or jaw muscles
xviii. Feeling self conscious
xix. Aching in your teeth that lasts at least an hour
xx. Discomfort, aching, or tenderness of your gums
xxi. Hardship in being able to pay for the treatment
xxii. A sense of suffocation
xxiii. Excessive salivation
xxiv. Difficulty chewing in the morning
xxv. Difficulty chewing with your back teeth that persists most of the day
xxvi. Movement of the teeth so that the upper and lower teeth no longer meet properly
xxvii.
xxviii.

(h) Impact
Complete this section only if you have completed section E above.

i. Please think of the questions in Sections A, B, C, and D. Having been treated for your sleep apnea and/or snoring do you believe that overall there has been an improvement in your quality of life since you started treatment? If yes, how much of an impact on your quality of life has there been as reflected by the questions asked in Sections A, B, C, and D. Place a mark on the line.
Scale:
0-10
(no impact) (extremely large impact)

ii. Please think of the symptoms that developed as a result of being treated for sleep apnea and/or snoring that you highlighted in Section E. How much of an impact on your quality of life have these symptoms had?
Scale:
0-10
(no impact) (extremely large impact)

Response Options

Yellow card

1. All the time
2. A large amount of the time
3. A moderate to large amount of the time
4. A moderate amount of the time
5. A small to moderate amount of the time
6. A small amount of the time
7. Not at all

Green card

1. A very large amount
2. A large amount
3. A moderate to large amount

4. A moderate amount
5. A small to moderate amount
6. A small amount
7. None

Red card

1. A very large problem
2. A large problem
3. A moderate to large problem
4. A moderate problem
5. A small to moderate problem
6. A small problem
7. No problem

A note about scoring: To obtain mean scores for Domains A through D the total score of each domain should be divided by the total number of questions answered. When the SAQLI is administered after a therapeutic intervention, allowance has been made for the possibility that the treatment, even if it is "successful", may have some independent negative consequences on a patient's quality of life. The scores from Domain E (Treatment-related Symptoms), are dealt with in a manner different from that of the other four domains. First the scores require recoding (7 to 0, 6 to 1, 5 to 2, 4 to 3, 3 to 4, 2 to 5, and 1 to 6). For Domain E the mean recoded score is obtained by dividing the total score by 5 (regardless of how many symptoms were identified). Next, the mean value of the recoded scores needs to be weighted according to the impact of the treatment-related symptoms on quality of life in comparison with the impact of the improvement of Domains A through D. Weighting is accomplished by dividing the impact score for Domain E (a number from 0 to 10) by the impact score for Domains A through D (Section F of the SAQLI). If this quotient exceeds 1, the result should be reduced so that the weighting factor never exceeds 1. The mean recoded score from Domain E is multiplied by the weighting factor, and it is this product that should be subtracted from the sum of the mean scores from Domains A, B, C, and D.
To obtain the final SAQLI score the sum of the mean domain scores A, B, C, and D is divided by 4. If Domain E has been used after a therapeutic intervention, the SAQLI score is obtained by summing the mean domain scores A, B, C, and D, subtracting the mean recoded Domain E score (that has been adjusted by the weighting factor described above) and dividing by 4.

C.4 Berlin Questionnaire

Adapted from [72]. The questionnaire consists of three categories related to the risk of having sleep apnoea.
Patients can be classified into High Risk or Low Risk based on their responses to the individual items and their overall scores in the symptom categories.

Categories and scoring:
Category 1: items 1, 2, 3, 4, 5
Item 1: if'**Yes**' assign **1 point**
Item 2: if '**c**' or '**d**' is the response, assign **1 point**
Item 3: if '**a**' or '**b**' is the response, assign **1 point**
Item 4: if '**a**' is the response, assign **1 point**
Item 5: if '**a**' or '**b**' is the response, assign **2 points**
Add points. Category 1 is positive if the total score is 2 or more points

Category 2: items 6, 7, 8 (item 9 should be noted separately).
Item 6: if '**a**' or '**b**' is the response, assign **1 point**
Item 7: if '**a**' or '**b**' is the response, assign **1 point**
Item 8: if '**a**' is the response, assign **1 point**
Add points. Category 2 is positive if the total score is 2 or more points. Category 3 is positive if the answer to item 10 is 'Yes' or if the BMI of the patient is greater than 30 kg/m^2**.**
(BMI must be calculated. BMI is defined as weight (kg) divided by height (m) squared, *i.e.* kg/m^2.)

High Risk: if there are 2 or more Categories where the score is positive

Low Risk: if there is only 1 or no Categories where the score is positive

1. Complete the following:
   Height
   Weight
   Age
   Gender

Category 1

2. Do you snore?
   (a) Yes
   (b) No
   (c) Don't know

If you snore:

3. Your snoring is:
   (a) Slightly louder than breathing
   (b) As loud as talking
   (c) Louder than talking
   (d) Very loud - can be heard in adjacent rooms

4. How often do you snore?
   (a) Nearly every day
   (b) 3-4 times a week
   (c) 1-2 times a week
   (d) 1-2 times a month
   (e) Never or nearly never

5. Has your snoring ever bothered other people?
   (a) Yes
   (b) No
   (c) Don't know

6. Has anyone noticed that you quit breathing during your sleep?
   (a) Nearly every day
   (b) 3-4 times a week
   (c) 1-2 times a week
   (d) 1-2 times a month
   (e) Never or nearly never

Category 2

7. How often do you feel tired or fatigued after your sleep?
   (a) Nearly every day
   (b) 3-4 times a week
   (c) 1-2 times a week
   (d) 1-2 times a month
   (e) Never or nearly never

8. During your waking time, do you feel tired, fatigued or not up to par?
   (a) Nearly every day
   (b) 3-4 times a week
   (c) 1-2 times a week
   (d) 1-2 times a month
   (e) Never or nearly never

9. Have you ever nodded off or fallen asleep while driving a vehicle?
   (a) Yes
   (b) No

If yes:

10. How often does this occur?
    (a) Nearly every day
    (b) 3-4 times a week
    (c) 1-2 times a week
    (d) 1-2 times a month
    (e) Never or nearly never

Category 3

11. Do you have high blood pressure?
    (a) Yes
    (b) No
    (c) Don't know

Appendix D

Statistical Differences in Data

Table D.1, taken from Chapter 4, shows the demographics for each sub-group (normal, snorer, mild OSA, moderate OSA, severe OSA) used in the analysis. The distributions for each demographic can be found in Figure D.1.

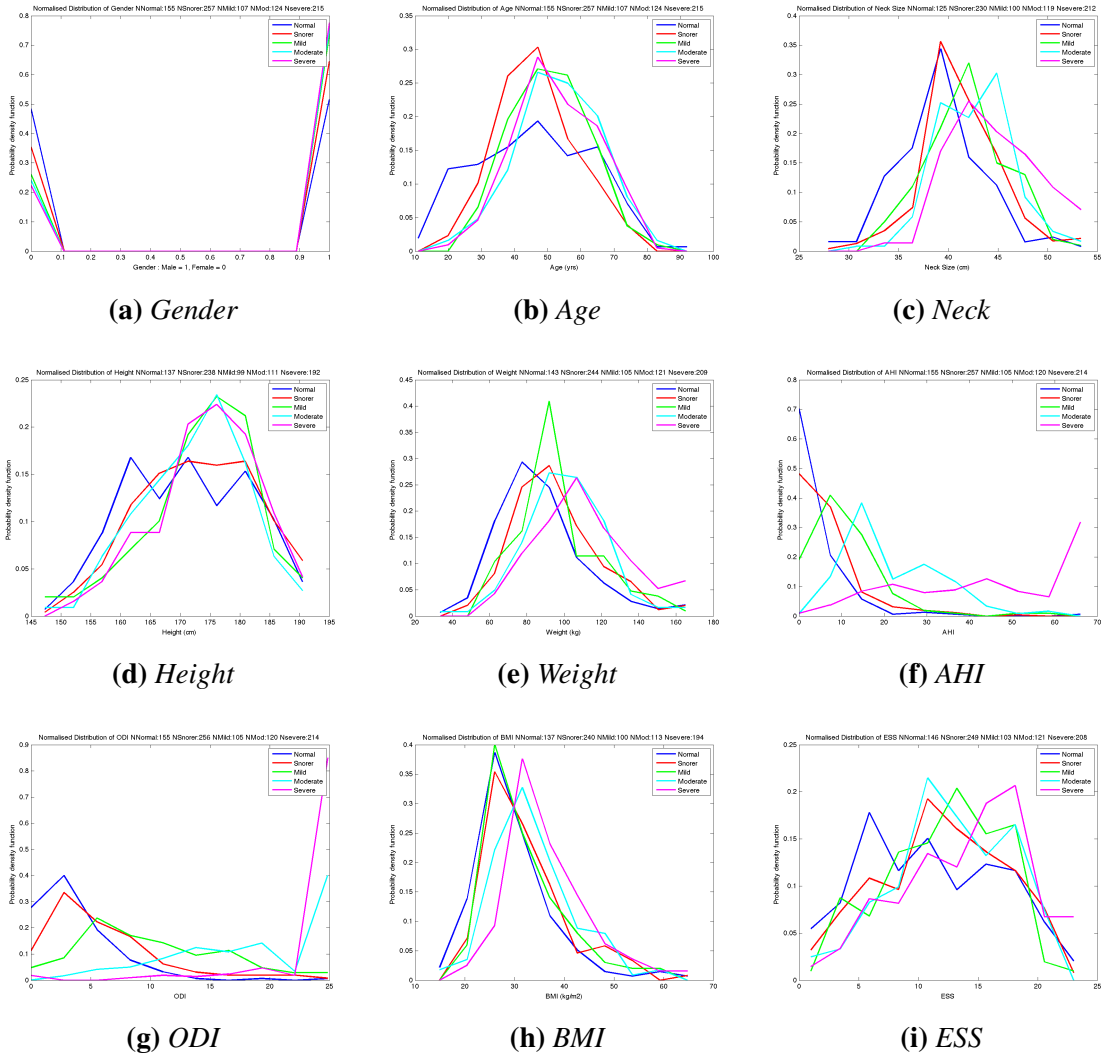
The Kruskal-Wallis test makes the following assumptions about the data in x:

- All samples come from populations having the same continuous distribution, apart from possibly different locations due to group effects.
- All observations are mutually independent.

The classical one-way ANOVA test replaces the first assumption with the stronger assumption that the populations have normal distributions. The p values for the one-way ANOVA test and the Kruskal-Wallis test can be found in Table D.2. The results indicate that both tests find that all demographics except for height are significantly different, although the results as less significant according to the Kruskal-Wallis test. It is typical that when a data set has a reasonable fit to the normal distribution, the classical ANOVA test is more sensitive to differences between groups.

Table D.1: *Subject demographics for each sub-group: normal, snorer, mild OSA, moderate OSA and severe OSA (mean±σ). neck = neck circumference, m = male, f = female.*

Group	Normal	Snorer	Mild	Moderate	Severe
Gender	80 m, 75 f	166 m, 91 f	79 m, 28 f	94 m, 30 f	167 m, 48 f
Age (yrs)	45.9 ± 17.1	46.5 ± 12.0	50.5 ± 11.4	53.1 ± 12.4	52.5 ± 12.6
Neck (cm)	39.4 ± 4.6	41.4 ± 4.3	41.9 ± 4.1	42.9 ± 3.8	45.0 ± 4.8
Height (cm)	171.2 ± 10.7	173.5 ± 10.4	174.2 ± 9.9	173.0 ± 9.7	175.0 ± 9.1
Weight (kg)	77.7 ± 23.0	96.0 ± 24.2	212.0 ± 48.8	221.2 ± 49.5	247.3 ± 74.4
AHI ($events/h$)	4.4 ± 7.5	6.4 ± 7.4	10.6 ± 9.0	21.5 ± 11.6	47.5 ± 24.5
ODI ($events/h$)	3.7 ± 3.5	6.0 ± 5.2	10.3 ± 7.0	22.0 ± 11.6	56.8 ± 32.4
BMI (kg/m^2)	29.6 ± 7.9	32.0 ± 8.4	31.9 ± 7.9	33.8 ± 8.5	36.9 ± 11.2
ESS	11.0 ± 5.6	12.0 ± 5.2	12.2 ± 4.7	12.7 ± 4.7	14.1 ± 5.3



(a) *Gender* **(b)** *Age* **(c)** *Neck*

(d) *Height* **(e)** *Weight* **(f)** *AHI*

(g) *ODI* **(h)** *BMI* **(i)** *ESS*

Figure D.1: *Probability density functions for all five sub-groups for the normalised demographics. Normal is shown in blue, snorer in red, mild in green, moderate in cyan and severe in magenta.*

Table D.2: *p value for the five sub-group (normal, snorer, mild OSA, moderate OSA, and severe OSA) demographics.*

Demographic	ANOVA	Kruskal-Wallis
Gender	3.5×10^{-7}	4.7×10^{-7}
Age	4.5×10^{-9}	1.2×10^{-8}
Neck	1.4×10^{-28}	3.2×10^{-26}
Height	0.0171	0.0211
Weight	9.7×10^{-17}	3.2×10^{-19}
AHI	9.7×10^{-164}	3.3×10^{-112}
ODI	9.2×10^{-170}	7.8×10^{-122}
BMI	8.7×10^{-12}	4.4×10^{-17}
ESS	1.1×10^{-6}	3.5×10^{-6}

Appendix E

Annotation Protocol

Please mark the snoring events, apnoea events and the first breath after an apnoea. If you see any breathing or noise events *e.g.* speech, TV, radio, etc. and wish to annotate them, feel free, but it is not necessary.

For each event type there is a corresponding single letter label with an optional secondary label/qualifier as shown in Table E.1:

Table E.1: *The labels to be used in annotating the data*

Event	**Label** (first letter)	**Optional sub-label**
Apnoea	A	U
First breath after apnoea	F	U
Snoring	S	U, C, S
Noise	N	U, V, T, R
Breathing	B	U, L, H

where U = uncertain, C = crescendo snoring, S = simple snoring, V = voice (live human not TV), T = TV, R = radio, L = light and H = heavy.

The uncertain label (U) denotes that you think the event is in this category, but are not entirely sure. Upper or lower cases may be used, or a mix. If you wish to add a free-text qualification/description after the annotations, please feel free. The letters A, F, S, N or B **must** be the first letter in the label. The other letters may be used if required. If there are any other subclasses that you wish to include please make a note of the label you are using.

To annotate the data:

1. Open the required file.
2. In order to mark an event, click the 'Insert All Channels Marker' button - marked in a red box in Figure E.1.
3. Click and drag the marker across the section of data that you wish to annotate - it will be highlighted as shown in Figure E.2.
4. When the relevant section of data is highlighted, markers will appear across all of the channels as in Figure E.3. By clicking the 'Insert All channels Marker' button again, it is possible to add a label to the annotation.
5. Right-click anywhere in the highlighted section and two options appear: *Edit Marker*, and *Delete*. If the marker is in the wrong section it can be removed. By using the *Edit Marker* option a label can be added as in Figure E.4.
6. When the *Edit Marker* option is chosen, a dialogue box appears as in Figure E.5. The label for the event is entered in the 'Annotation' box. Once the appropriate label has been entered, click 'OK'.

Figure E.1: *Where to find the 'Insert All Channels Marker' button.*

Figure E.2: *Click and drag the marker across the section of data that you wish to annotate.*

7. The label appears on each channel as in Figure E.6. Repeat this process until the required events have been labelled.

Pressing the 'Save' button while annotating the data saves the start time of each event, the duration of each event and the label associated with that event in a WorkPad (WP) file. If the patient file is closed, and you wish to add more annotations they will be saved in a different WP file. If, however, you have been saving the annotations throughout the process, each time you click 'Save' a dialogue box opens as in Figure E.7. By overwriting the WorkPad file, all of the annotations, including the latest annotations, are saved in the same file.

The WP files are saved in chronological order i.e. WP0, WP1, etc. and appear in the folder containing the patient data. For example: if the patient data has been saved in C:\data where data is a folder containing a number of patient folders, the WP files for patient X will be saved in C:\data\X.

The WP files can be opened using WordPad, NotePad, etc. There are no patient identifiers contained in the file, and so the files can be sent via email.

Figure E.3: *Relevant section has markers on all channels.*

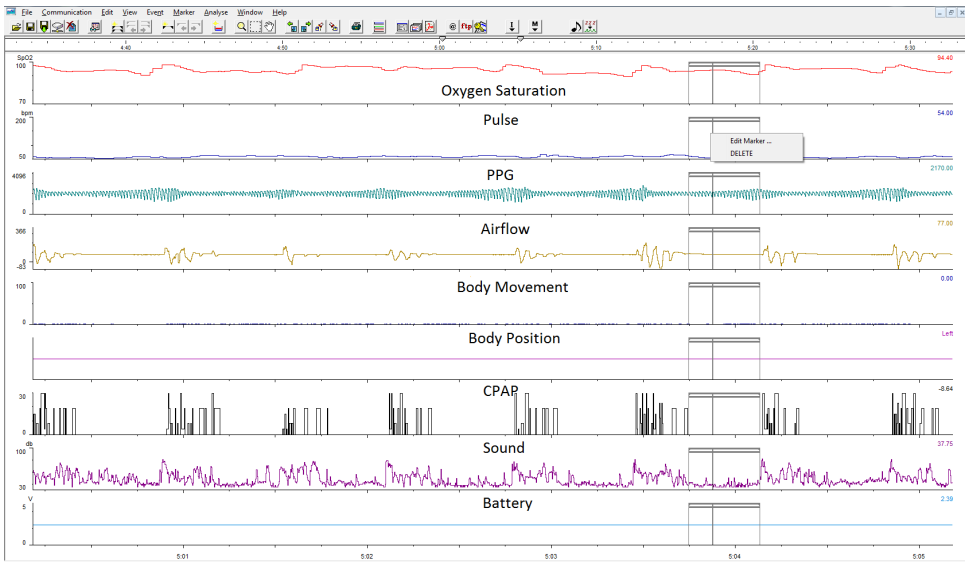


Figure E.4: *Right-click to get the options: Edit Marker and Delete.*

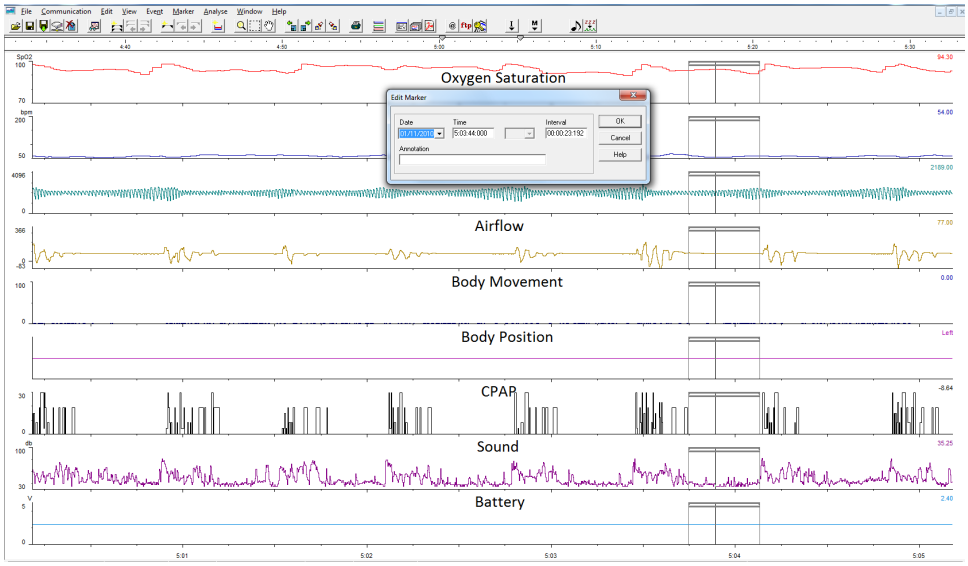


Figure E.5: *The label for the event is entered in the box under 'Annotation'.*

Figure E.6: *Channel markers with labels.*

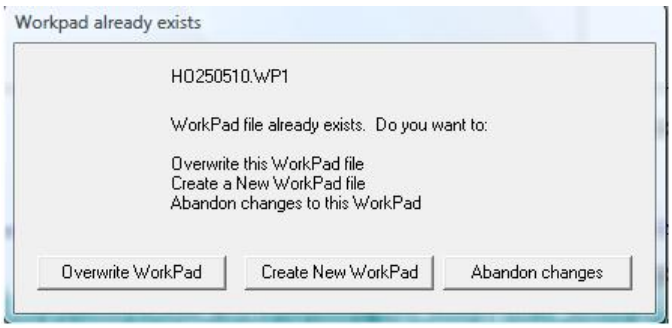


Figure E.7: *Options given when saving the annotations.*