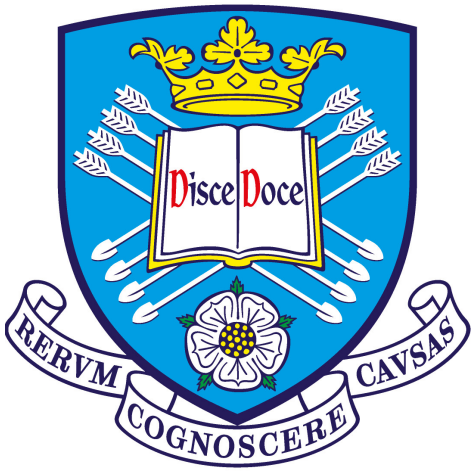
Deep Sleep

Deep Learning Methods for the Acoustic Analysis of Sleep-disordered Breathing



Hector E. Romero

Supervisors: Prof. Guy J. Brown
Dr. Ning Ma

Department of Computer Science
University of Sheffield

This thesis is submitted for the degree of
Doctor of Philosophy

August 2021

I would like to dedicate this thesis to my loving family...

Acknowledgements

First and foremost, I would like to thank Guy Brown and Ning Ma for their encouraging and careful supervision throughout this great 4-year white-knuckle ride. I am deeply grateful to Passion for Life Healthcare and the Department of Computer Science of the University of Sheffield for having funded my PhD, without their support I would not have been able to do it.

Special thanks to Jon Barker and Mari-Cruz Villa-Uriol not only for their insightful comments in our PhD panel meetings but also for having encouraged me to pursue a PhD. I would like to express my gratitude to Richard Wiffen, Iain Spray, Stelios Giannoulis, Phil Gillespie, Sam Johnson, Will Webb and Amna Rahim of Passion for Life Healthcare for their support.

I am profoundly thankful to my examiners, Heidi Christensen and Nick Cummins, for their invaluable feedback and for having made the viva voce an enjoyable experience. Thanks are also due to Amy Beeston and Madina Hasan for having paved the way for my PhD. Finally, I thank my family for their constant love and support.

Abstract

Sleep-disordered breathing (SDB) is a serious and prevalent condition that results from the collapse of the upper airway during sleep, which leads to oxygen desaturations, unphysiological variations in intrathoracic pressure, and sleep fragmentation. Its most common form is obstructive sleep apnoea (OSA). This has a big impact on quality of life, and is associated with cardiovascular morbidity. Polysomnography, the gold standard for diagnosing SDB, is obtrusive, time-consuming and expensive. Alternative diagnostic approaches have been proposed to overcome its limitations. In particular, acoustic analysis of sleep breathing sounds offers an unobtrusive and inexpensive means to screen for SDB, since it displays symptoms with unique acoustic characteristics. These include snoring, loud gasps, chokes, and absence of breathing.

This thesis investigates deep learning methods, which have revolutionised speech and audio technology, to robustly screen for SDB in typical sleep conditions using acoustics. To begin with, the desirable characteristics for an acoustic corpus of SDB, and the acoustic definition of snoring are considered to create corpora for this study. Then three approaches are developed to tackle increasingly complex scenarios. Firstly, with the aim of leveraging a large amount of unlabelled SDB data, unsupervised learning is applied to learn novel feature representations with deep neural networks for the classification of SDB events such as snoring. The incorporation of contextual information to assist the classifier in producing realistic event durations is investigated. Secondly, the temporal pattern of sleep breathing sounds is exploited using convolutional neural networks to screen participants sleeping by themselves for OSA. The integration of acoustic features with physiological data for screening is examined. Thirdly, for the purpose of achieving robustness to bed partner breathing sounds, recurrent neural networks are used to screen a subject and their bed partner for SDB in the same session. Experiments conducted on the constructed corpora show that the developed systems accurately classify SDB events, screen for OSA with high sensitivity and specificity, and screen a subject and their bed partner for SDB with encouraging performance. In conclusion, this thesis makes promising progress in improving access to SDB diagnosis through low-cost and non-invasive methods.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AASM American Academy of Sleep Medicine

ACF autocorrelation function

AHI apnoea-hypopnea index

API application programming interface

ASR automatic speech recognition

AUC area under the ROC curve

BLSTM bidirectional long short-term memory

CASA computational auditory scene analysis

CNN convolutional neural network

ComParE Computational Paralinguistics Challenge

CPAP continuous positive airway pressure

CSA central sleep apnoea

DER diarisation error rate

DISE drug-induced sleep endoscopy

DNN deep neural network

ECG electrocardiogram

EEG electroencephalogram

EMG electromyogram

EOG electrooculogram

ERB equivalent rectangular bandwidth

FFT fast Fourier transform

FN false negatives

FP false positives

GCC-PHAT generalised crosscorrelation with phase transform

GMM Gaussian mixture model

HMM hidden Markov model

HSAT home sleep apnoea testing

HTK Hidden Markov Model Toolkit

ICA independent component analysis

IID interaural intensity difference

ISD interaural spectral difference

ITD interaural time difference

logSTFT log-scaled short-time Fourier transform

LSTM long short-term memory

MFCC mel-frequency cepstral coefficient

MPM McLeod pitch method

MSE mean squared error

MSLT multiple sleep latency test

MWT maintenance of wakefulness test

NHS National Health Service

NLP Natural Language Processing

NREM non-rapid eye movement (sleep)

ODI oxygen desaturation index

OSA obstructive sleep apnoea

PCA principal component analysis

PPG photoplethysmography

PSD power spectral density

PSG polysomnography

ReLU rectified linear unit

REM rapid eye movement (sleep)

RI respiratory index

RM (auditory nerve firing) rate map

RMS root mean square

ROC receiver operating characteristic

SDB sleep-disordered breathing

SER snore-event error rate

SIFT simplified inverse filtering technique

SNR signal-to-noise ratio

STFT short-time Fourier transform

SVM support vector machine

tanh hyperbolic tangent

TDOA time difference of arrival

TN true negatives

TP true positives

Chapter 1

Introduction

"Sleep that knits up the ravell'd sleave of care,
The death of each day's life, sore labor's bath,
Balm of hurt minds, great nature's second course,
Chief nourisher in life's feast."
– W. Shakespeare

How did you sleep last night? Did you sleep deeply and soundly like a log? Or were you perhaps disturbed by someone snoring? Regardless, it is likely that you spend about one third of your life sleeping or at least trying to. Although we still do not fully understand sleep, we do know that it is critical for our survival, health, and normal cognitive functioning (Colten and Altevogt, 2006a). Sleep-related disorders and sleep deprivation have serious consequences. Poor attention, short-term memory and vigilance resulting from sleep deprivation (Mantua and Simonelli, 2019) have been contributing factors to many catastrophic events like the Chernobyl and Three Mile Island nuclear accidents (Shokoueinejad et al., 2017), the Challenger explosion (Durning et al., 2014), and the Exxon Valdez grounding (Hossain and Shapiro, 2002).

Several physiological changes take place during sleep: brain activity, heart rate and body temperature decrease, muscles relax, breathing becomes erratic, and airway resistance increases (Colten and Altevogt, 2006a). If your sleep last night was disrupted by someone snoring, you have evidenced first-hand an unhealthy deviation from those physiological changes that occur during sleep in the respiratory system: muscle tone decreases, airway resistance increases, and turbulent airflow causes the vibration of structures in the upper airway. The result is a loud sound that is disruptive, but also may be indicative of more serious conditions such as obstructive sleep apnoea (OSA): muscle tone decreases, upper airway collapses, airflow ceases, blood oxygen levels drop, the sufferer arouses gasping for breath, and sleep is fragmented.

OSA can be life-threatening, since it is associated with vascular morbidity and related causes of mortality, for instance, hypertension, stroke and myocardial infarction. Furthermore, it causes daytime hypersomnolence, which commonly results in car accidents (Cho et al., 2013; Lee et al., 2016; Young et al., 1993). Sleep disorders have also been associated with metabolic problems such as diabetes and insulin resistance (Castaneda et al., 2018), and neurocognitive conditions like dementia (Lal et al., 2012). OSA affects approximately 24% of men and 9% of women in Europe and the United States (Shokoueinejad et al., 2017). Its prevalence is increasing: in the last ten years the number of tests carried out by the National Health Service (NHS) in the United Kingdom to diagnose sleep disorders has doubled (Rhodes, 2017). It is estimated that sleep disorders cause the loss of 200,000 working days a year in the United Kingdom, which cost £40 billion (Hope, 2016), and 75% of severe cases remain undiagnosed (Motamedi et al., 2009) due to poor accessibility to diagnosis (Young et al., 2008). Not recognising sleep disorders precludes treatment, and the possibility of preventing their grave consequences (Colten and Altevogt, 2006b).

This thesis is concerned with the relationship between sleep, breathing and sound. Its central argument is that breathing sounds during sleep can be analysed to unobtrusively obtain relevant information about respiratory conditions associated with it. The analysis of breathing sounds to screen for sleep-disordered breathing might contribute to improving accessibility to diagnosis, as it would potentially allow more people to conveniently assess and monitor their condition at home using readily available hardware, for example, a smartphone. This would also allow a better use of the limited diagnostic resources, as the high prevalence of these conditions and the COVID-19 pandemic have put clinical respiratory services under pressure (Voulgaris et al., 2020; Williamson, 2020). The tools of speech technology, which have been extensively developed and researched, provide a sensible starting point for that analysis, since both speech and breathing sounds are time-series signals, have a similar frequency range, and are generated in the vocal tract. Specifically, deep learning methods — machine learning techniques based on deep neural networks (DNNs) — will be important tools in our work here, as they have revolutionised audio and speech technology in the last decades.

This introductory chapter consists of three parts. First, sleep-disordered breathing is introduced and its acoustic analysis is motivated. The second section presents an account of the use of sound and machine learning in healthcare from a historical perspective. We will see that the diagnostic usefulness of the sounds of the human body has been recognised for millennia, and machine learning can be used in healthcare for diagnosis. Lastly, the third part provides an overview of this thesis and indicates its structure.

1.1 Sleep-disordered Breathing

Although sleep-disordered breathing was not recognised as a well-defined clinical condition until the 1960s (Petersen-George, 1999), the relationship between sleep and breathing has been noted for centuries. For instance, in the history play 'Henry IV' written by the English playwright William Shakespeare (1564–1616) one can read Prince Harry, future King Henry V, and Peto, a highwayman, talking about Sir John Falstaff:

> PETO: Falstaff! Fast asleep behind the arras, and snorting like a horse.
> PRINCE: Hark how hard he fetches breath. Search his pockets. What hast thou found? (Shakespeare, 2014)

Sleep-disordered breathing is a condition characterised by frequent breathing pauses during sleep that result in blood oxygen desaturations, unphysiological variations in intrathoracic pressure, and sleep fragmentation. Its physiological spectrum ranges from increased upper airway resistance and partial airway collapse, evidenced as snoring and hypopnea events (i.e., shallow breathing), to complete airway collapse, manifested as apnoea events (i.e., absence of breathing) that can last more than 60 seconds. OSA, the most severe form of sleep-disordered breathing, is clinically defined by recurrent apnoea and hypopnea events along with symptoms of functional impairment.

Various factors can increase the risk of OSA. These include excess weight, craniofacial and upper airway structure (e.g., soft tissue or skeletal abnormalities, enlarged tonsils, and maxillary position or size), age (i.e., higher prevalence of OSA among older subjects), sex (e.g., men, and postmenopausal women), alcohol consumption before sleep, genetics (e.g., ethnic differences), smoking, and nasal congestion (e.g., anatomy, and allergic rhinitis) (Yaggi and Strohl, 2010; Young et al., 2004). It is very likely that Shakespeare's character Sir John Falstaff suffers from OSA, since he is described as an overweight and old man who is a heavy drinker (Shakespeare, 2014).

During sleep we regularly cycle through two sleep states: non-rapid eye movement (NREM), and rapid eye movement (REM). NREM is additionally divided into three stages imaginatively called N1, N2, and N3. Each stage has particular physiological characteristics. Sleep begins with N1, which serves as the transition from wakefulness to sleep. Muscle activity decreases slightly, slow eye movements are usually present, and it can be easily interrupted by noise. In stage N2, heart rate, blood pressure, gastrointestinal activity, and brain metabolism are attenuated. Memory consolidation takes place during this stage. About 50% of the night is spent in N2. In N3 sleep, also known as deep sleep, muscle tone decreases, eye movement is absent, and the threshold for arousal is the highest of all stages. 60–90 minutes after sleep onset, the first episode of REM sleep occurs. It is characterised by atonia or absence of muscle tone to prevent you from acting out your dreams (or nightmares), bursts of rapid eye movements, fluctuations in heart rate

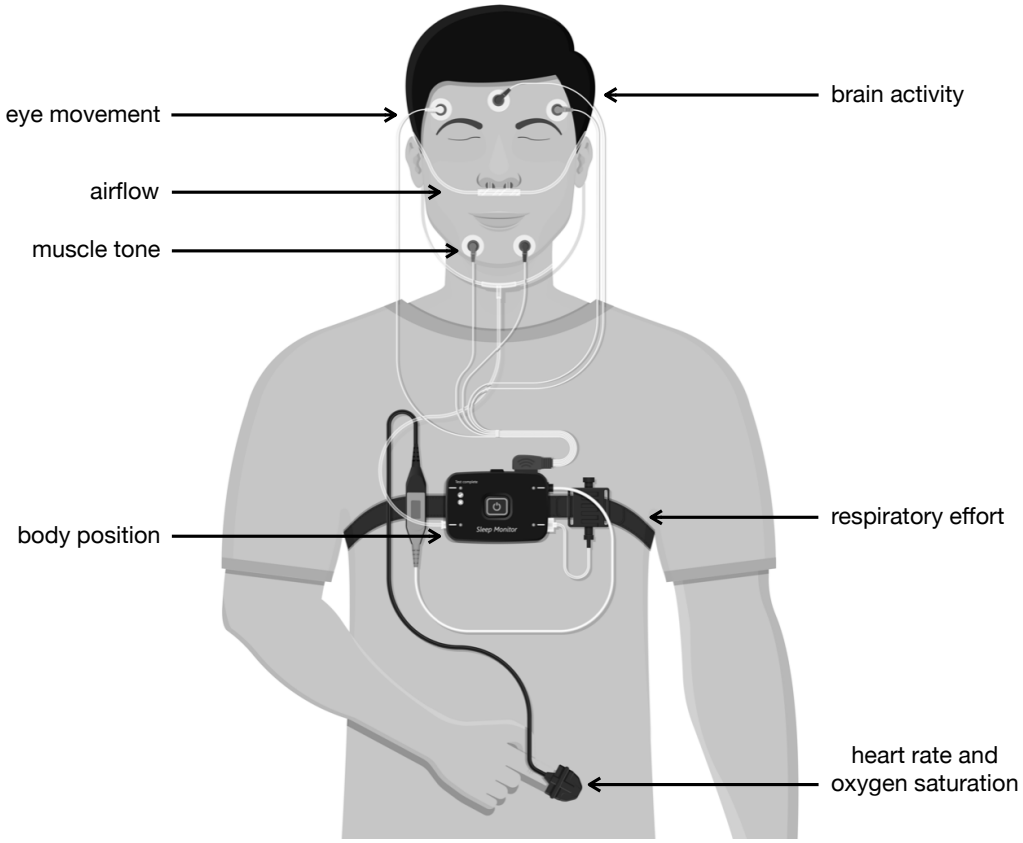


Fig. 1.1 Illustration of a typical PSG test. Heart rate and oxygen saturation are measured with a pulse oximeter on the index finger. Respiratory effort is measured with a thoracic belt. Body position is recorded with an accelerometer in the device on the chest. Airflow is measured with a nasal cannula. Muscle tone is recorded with electrodes on the chin. Brain activity is recorded with electrodes on the scalp and forehead. Eye movement is recorded with electrodes placed next to the eyes.

and blood pressure, erratic respiration, and brain activity similar to wakefulness. 25% of the night is spent in this stage. Apnoeas and hypopneas are more likely to happen during REM sleep due to the absence of muscle tone (Chokroverty and Avidan, 2016; Colten and Altevogt, 2006a).

Polysomnography (PSG) is the gold standard for assessing sleep-disordered breathing. It evaluates sleep, breathing and movement, and is performed overnight in a sleep laboratory. At least 4 hours of PSG recordings are needed for a reliable assessment. Several physiological parameters are measured during PSG: brain activity with electroencephalography (EEG), muscle activity and tone with electromyography (EMG), eye movement with electrooculography (EOG), cardiac function with electrocardiography (ECG), heart rate and oxygen saturation with photoplethysmography (PPG) or piezoelectric crystals, airflow with nasal pressure sensors or thermistors, respiratory effort with thoracic

Table 1.1 AHI and OSA Severity

AHI (events/hour)	**OSA Severity**
<5	Normal
5 – 15	Mild
15 – 30	Moderate
≥30	Severe

and abdominal belts, and body position with video or accelerometers (Mansukhani et al., 2020). Sleep stages are derived from EEG and EOG. This is illustrated in Figure 1.1. A PSG test costs around £360 per night in the United Kingdom (NHS, 2020) and £700 in the United States (Kim et al., 2015).

PSG is manually scored by a polysomnographic technologist, who annotates sleep stages and respiratory events in the recorded data according to well-defined guidelines such as those of the American Academy of Sleep Medicine (AASM). A polysomnographic technologist spends up to two hours to manually score it (Malhotra et al., 2013). Sleep stages are derived from the brain activity, muscle activity and eye movement. Respiratory events are marked considering the airflow, respiratory effort, oxygen saturation and sleep stage. Apnoeas are defined as a >90% reduction in airflow for more than 10 seconds, and there are two types. Obstructive apnoeas are those caused by a collapse of the upper airway, and occur with respiratory effort. Whereas central apnoeas are the result of unstable ventilatory drive by the central nervous system and, therefore, take place with no respiratory effort. Hypopneas are defined as a >30% reduction in airflow for more than 10 seconds associated with either a >3% oxygen desaturation or an arousal (i.e., lightening of sleep stage). After scoring the recorded data, OSA severity is rated by the apnoea-hypopnea index (AHI) as shown in Table 1.1. The AHI is defined as the average number of apnoeas and hypopneas per hour of sleep (American Academy of Sleep Medicine, 2020; Mansukhani et al., 2020):

$$\text{AHI} = \frac{\text{number of apnoea-hypopnea events}}{\text{time spent sleeping (in hours)}} \tag{1.1}$$

PSG is inconvenient, time-consuming and expensive. Additionally, the 'first-night effect' limits the reliability of it: longer sleep latency (i.e., the time required to fall asleep), longer periods of N1 sleep, shorter periods of REM sleep, sleep fragmentation, decreased sleep efficiency (i.e., the proportion of time actually spent sleeping while in bed), and less total sleep time are observed during the first night of sleep in a laboratory in comparison with following nights. This results from the limitation of movement and inconvenience from the cables and sensors attached to the body, the psychological effects of being un-

der observation (Lu et al., 2014), and the unfamiliar sleep environment (Mansukhani et al., 2020).

It is worth noting that, due to the increasing prevalence of sleep disorders and limited diagnostic resources, home sleep apnoea testing (HSAT) is employed in most sleep studies rather than PSG. Although PSG and HSAT are very similar, they exhibit some differences. The latter is performed at home instead of a sleep clinic. Sleep stage and muscle tone data are not available on HSAT, since brain activity is not recorded with EEG, and eye movements and muscle activity are not monitored. Whereas PSG is carried out by a Registered Polysomnographic Technologist, participants place the sensors and operate the HSAT device themselves, which sometimes results in unreliable measurements (Rosen et al., 2018).

As discussed in the introduction of this chapter, relevant information about respiratory conditions associated with sleep can be obtained unobtrusively by analysing breathing sounds, since sleep-disordered breathing displays symptoms with unique acoustic characteristics, for example, snoring, loud gasps, chokes and absence of breathing. We also mentioned that deep learning — a form of machine learning — could be used for that analysis, given the revolution that it has brought about to audio and speech technology. In the upcoming section, we will consider how sound and machine learning have been used in healthcare from a historical viewpoint.

1.2 Sound and Machine Learning in Healthcare

The diagnostic usefulness of the sounds of the human body has been recognised for millennia. The civilisation of ancient Egypt listened to the sounds of the human body for diagnosis more than 5,000 years ago (Breasted, 1980; Hughes, 1988; Middendorp et al., 2010; Stiefel et al., 2006). Ancient Greek physicians auscultated heart and lung sounds by placing their ear on the patient's chest (Hajar, 2012). Hippocrates (460–377 BC), historically regarded as the father of medicine, described the pleural friction rub due to pleuritis as "a creak like new leather" (Montinari and Minelli, 2019, p. 184). He also proposed listening to the sounds in the chest after shaking the patient by their shoulders to detect the presence of fluid in the chest, which is currently known as hydropneumothorax (Girithari et al., 2018). The Italian polymath Leonardo Da Vinci (1452–1519) studied cardiac auscultation (Montinari and Minelli, 2019). The French surgeon Ambroise Paré (1510–1590) observed that "if there is matter or other humours in the thorax, one can hear a noise like that of a half-filled gurgling bottle" (Hajar, 2012, p. 24).

Until the start of the 19th century, physicians performed immediate auscultation, that is, they placed their ear directly on the patient's chest. In 1816, the French physician and musician René Laënnec (1781–1826), embarrassed to place his ear on a young woman's

chest for examination, rolled sheets of paper into a cylinder, placed one end to his ear and the other to her chest. Laënnec "was not a little surprised and pleased to find that... [he] could thereby perceive the action of the heart in a manner much more clear and distinct than... [he] had ever been able to do by the immediate application of... [his] ear" (Laënnec, 1821). After this, the French physician experimented with different materials and came up with a hollow wooden cylinder, which he called stethoscope (Montinari and Minelli, 2019). This device has evolved over the years, but its principle is the same: transmission and amplification of sound by resonance (Andres et al., 2018). The stethoscope is still indispensable in current clinical practice.

In the second half of the 20th century, physicians started using sounds that they could not hear for monitoring and diagnosis. In the 1950s, Ian Donald, John MacVicar and Tom Brown developed the first 2D ultrasound scanner, and called it diasonograph (Campbell, 2013). It recorded and mapped the reflections produced by pulsed ultrasound in the human body. They proposed its use for the diagnosis of early pregnancy complications and detection of abdominal masses (Donald et al., 1958). This technique has been significantly improved over the years and is essential in obstetrics and gynaecology.

At about the same period that the diasonograph was being developed, the English mathematician Alan Turing (1912–1954) published a paper in which he considered the question 'Can machines think?', and posed the idea of a learning machine:

> Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Turing, 1950)

Turing ended his paper stating that "we can only see a short distance ahead, but we can see plenty there that needs to be done." As electronic computers came into use in the second half of the 20th century, the algorithms that allow analysing and modelling data started being developed, and the field of machine learning appeared (Kononenko, 2001).

Machine learning applications in healthcare can be dated back to the 1970s. In 1973, the SUMEX-AIM (Stanford University Medical EXperimental computer for Artificial Intelligence in Medicine) project started in the United States and ran for almost 20 years (Kulikowski, 2019; Shortliffe, 2019). Projects in SUMEX-AIM focused on the design of knowledge-based systems, which mainly used heuristic decision-making strategies expressed symbolically (Freiherr, 1980). One of the successful applications was INTERNIST, an experimental diagnostic system that was intended to aid physicians in solving complex cases in internal medicine. For this purpose, essential internal medicine knowledge

was translated into symbolic data structures and stored in the computer. The system took as input descriptions of disease manifestation, and asked for more information if needed (e.g., laboratory data). It then compared the input data with the stored disease profiles in a hierarchical manner, and output a diagnosis. For instance, given a particular set of descriptions, INTERNIST started suggesting a liver disease and, after requiring more information, it specified a disease like hepatitis (i.e., inflammation of the liver). However, the programme was not sufficiently robust for clinical use, as it was not able to consider temporal information (e.g., the progression of symptoms) and did not construct differential diagnoses (Miller et al., 1985). Considering temporal information is critically important for diagnosis, as the way symptoms and physiological parameters vary over time provides relevant cues that aid in understanding the dynamics of medical conditions (Zhou and Hripcsak, 2007) whether the diagnostic process is carried out by a physician or a computer. This is a fundamental concept that will be exploited throughout this thesis.

Many of the projects in SUMEX-AIM were based on rules crafted by human experts rather than patterns or models automatically learned by the computer from data. Rule-based approaches are unlikely to be robust enough in practice, since the great variability of (medical) data might not be fully considered by human experts. Then these approaches are usually fragile in the presence of noise, and their performance tends to degrade on unseen data. Machine learning approaches are better suited to deal with that variability, as they can find and leverage complex patterns and relationships in data. However, large amounts of data are required to effectively learn models or patterns from it, which is a challenge in data-scarce fields like healthcare (Meijerink et al., 2020), and highlights the need for data collection.

As computers became more powerful and affordable at the end of the 20th century, and much more data started being gathered, more complex machine learning techniques, in particular those of deep learning, could be developed and applied to a wider range of tasks (Kononenko, 2001). As mentioned before, deep learning — a form of machine learning based on DNNs that will be considered in detail in the next chapter — has revolutionised audio and speech technology in the 21st century (Dean, 2020). Based on the diagnostic usefulness of the sounds of the human body recognised more than five millennia ago, and the revolution in audio and speech technology in the last decades, deep learning approaches that make use of sound have been extrapolated to healthcare applications (Cummins et al., 2018; Latif et al., 2020).

Deep learning has been used to detect heart diseases from cardiac sounds as an alternative to diagnostic tests that require the person to attend hospital facilities, for example, ECG or coronary angiography. Brunese et al. (2020) recorded cardiac sounds with a smartphone, and classified them as either healthy or unhealthy using acoustic features

as input to a DNN. This study evidences the potential of using deep learning for analysing sounds of the human body to conveniently screen for prevalent conditions at home with a smartphone. Deep learning has also been employed to detect lung sounds indicative of pulmonary disease. Using a bespoke device, Messner et al. (2020) collected multichannel lung sound recordings, since several recording positions over the chest are needed to cover all the organ. Similar to the previous study, acoustic features were provided to a DNN to classify lung sounds as either healthy or pathological. Although the use of a specialised hardware might limit the wide application of the proposed method, multichannel sound recordings provide additional information, in comparison with single-channel recordings, that can be exploited in addition to temporal and spectral features, for instance, spatial cues.

Screening for COVID-19 infection from coughs has been another application of the tools of speech technology and deep learning to healthcare. To overcome the inability to test at scale for COVID-19, Imran et al. (2020) exploited the pathomorphological alterations in the respiratory system produced by COVID-19 infection to predict its presence from coughs recorded with a smartphone. Acoustic features were used as input to DNNs, and the system produced one of the following outcomes: 'COVID-19 likely', 'COVID-19 not likely' and 'test inconclusive'. Unlike previous studies, which have used passive sound recordings, Chan et al. (2019) used active sound recordings and machine learning to screen for ear infection in children by playing chirps close to the ear and analysing the reflected waves from the eardrum. Leveraging the fact that the presence of middle ear fluid is a key diagnostic marker for paediatric ear infections, they employed the speaker and microphone of smartphones to detect middle ear fluid by evaluating eardrum mobility. A paper funnel, which can be easily assembled using a printed paper template, was attached to the smartphone and placed at the child's ear canal entrance to direct sound into it. The studies by Imran et al. (2020) and Chan et al. (2019) demonstrate once more the potential of applying deep learning to conveniently screen for health conditions with smartphones by analysing sounds of the human body.

1.3 Thesis Overview

We have seen in this introductory chapter that there is a clear need for alternative diagnostic approaches for assessing sleep-disordered breathing, since PSG is obtrusive, expensive and time-consuming. Furthermore, one night of PSG might not provide a reliable assessment of the condition (Lu et al., 2014; Mansukhani et al., 2020). The prevalence of sleep disorders is increasing (Rhodes, 2017), they have serious consequences along with well-established comorbidities (Castaneda et al., 2018; Cho et al., 2013; Lal et al., 2012; Lee

et al., 2016; Young et al., 1993), and most of the severe cases remain undiagnosed (Motamedi et al., 2009) due to poor accessibility to diagnosis (Young et al., 2008).

Ideally, alternative diagnostic approaches should be robust enough to be applied in typical sleep conditions — a bedroom at home — and make use of unobtrusive and readily available hardware — for example, a smartphone. Exploiting the unique acoustic characteristics of sleep-disordered breathing symptoms and building on the success of deep learning methods, this thesis investigates the analysis of breathing sounds during sleep as an unobtrusive means to robustly screen for sleep-disordered breathing in typical sleep conditions.

1.3.1 Research Questions

This thesis addresses the following research questions:

RQ1 Healthcare research is a data-scarce field, and the performance of deep learning methods greatly depends on the quantity and quality of available data. Although a large amount of healthcare data is collected everyday, it is seldom made available for research. Since there are no standard corpora of sleep-disordered breathing collected in realistic conditions with suitable annotation for acoustic analysis tasks, collection and annotation of data is needed in the present study. Then, **what are the desirable characteristics for an acoustic corpus of sleep-disordered breathing?**

RQ2 We have spoken so far about snoring, a common form of sleep-disordered breathing, and agreed that it is a loud sound caused by the vibration of structures in the upper airway resulting from the partial collapse of it during sleep. However, in order to build effective sound classifiers, it is important to understand what characterises snoring in an acoustic sense. In fact, there is no widely recognised definition of snoring, and it would be helpful to have one, as it might inform decisions for the acoustic analysis of breathing sounds. We therefore ask **what is the acoustic definition of snoring?**

RQ3 Highly prevalent, high-risk and treatment-available diseases normally have a diagnostic test or gold standard, and a screening test. Usually, diagnostic tests are invasive and expensive, whereas screening tests are non-invasive and inexpensive. So, the latter can be applied massively unlike diagnostic tests. If the screening test provides a positive result, the diagnostic test is then carried out to establish a definitive diagnosis. For instance, mammographies — a non-invasive test — are routinely performed to screen for breast cancer — a highly prevalent disease. If the presence of cancer is indicated, a biopsy — an invasive test — is done to confirm the diagnosis. There is no widely accepted screening test for sleep-disordered breathing.

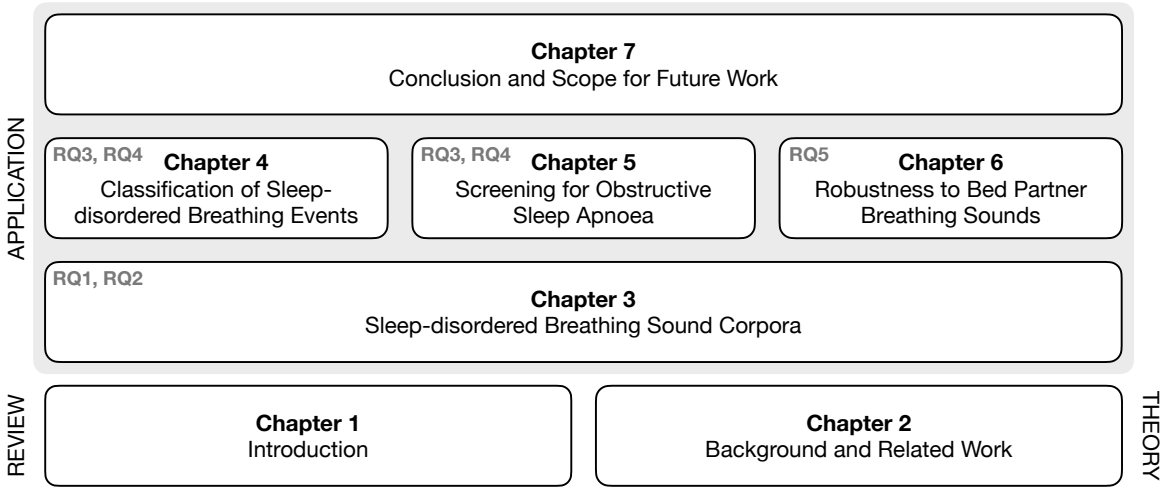


Fig. 1.2 Organisation of the thesis. The research questions addressed by each chapter are indicated.

Furthermore, the consistency and reliability of PSG — the gold standard for sleep-disordered breathing diagnosis — have been questioned, since the limited movement and discomfort from wearing sensors, the unfamiliar sleep environment, and the psychological effects of being under observation affect sleep. For this reason, a screening test for sleep-disordered breathing should ideally be applied unobtrusively in typical sleep conditions and make use of readily available hardware — in a bedroom at home using a smartphone, for example. Consequently, **to what extent is it possible to robustly screen for sleep-disordered breathing in typical sleep conditions using acoustics?**

RQ4 After collecting sleep-disordered breathing data, large amounts of unlabelled data would be available. Since manually annotating data is expensive and time-consuming, it would not be possible to annotate all the collected data. We therefore ask **how can a large amount of unlabelled sleep-disordered breathing data be leveraged to achieve robustness to background noise in typical sleep conditions?**

RQ5 As noted in RQ3, a screening test for sleep-disordered breathing should ideally be applied unobtrusively in typical sleep conditions, that is, a bedroom at home, which might include a bed partner who can negatively impact on the performance of the screening test. Then, **to what extent is it possible to achieve robustness to bed partner breathing sounds?**

1.3.2 Organisation of the Thesis

A diagram of the organisation of this thesis is shown in Figure 1.2. In the current chapter we have introduced the fundamentals of sleep-disordered breathing, and the use of sound and machine learning in healthcare. We have also motivated the acoustic analysis of sleep-disordered breathing. Chapter 2 will review machine learning techniques that can be employed in the analysis of sound events with special attention to deep learning methods. It will then review some approaches that have been proposed to classify sleep-disordered breathing events, and screen for OSA. These two chapters will provide the foundation for the present study.

Having recognised that healthcare is a data-scarce field, and discussed how the performance of deep learning methods greatly depends on the quantity and quality of data, Chapter 3 will introduce the *Snoring Sound Corpus* and the *OSA Sound Corpus*. Their collection and manual annotation will be described. Next, Chapter 3 will characterise the main forms of sleep-disordered breathing — snoring and OSA — from an acoustic perspective using the introduced corpora. By the end of this chapter, research questions RQ1 and RQ2 will have been addressed.

Making use of the sleep-disordered breathing corpora presented in Chapter 3 and informed by the acoustic characterisation carried out, Chapters 4 and 5 will attempt to address research questions RQ3 and RQ4. Chapter 4 will focus on the classification of sleep-disordered breathing events — mainly snoring — and Chapter 5 will concentrate on screening for OSA. For both tasks, the temporal pattern of breathing sounds during sleep will be exploited, novel feature representations will be proposed and compared with conventional acoustic features, and different machine learning techniques will be considered. Chapter 5 will also investigate the integration of acoustic features with physiological data, and will conclude by examining the feasibility of deploying the screening system on mobile devices.

Chapter 6 will look into a more challenging task. While Chapters 4 and 5 will work on data from participants sleeping on their own, Chapter 6 will consider the problem of detecting snoring when a bed partner is present and snores as well. It will start by exploring a multichannel approach with the idea of potentially using source separation techniques to segregate each subject's breathing sounds. Following this, it will propose and evaluate a single-channel method that draws inspiration from speaker diarisation work. Research question RQ5 will have been tackled by the end of this chapter.

Lastly, Chapter 7 will revisit the research questions that were put forward in this introductory chapter to reflect on the contributions and limitations of the present study. It will end by contemplating future research directions.

Chapter 2

Background and Related Work

"You understand sleep when you are awake, not when you are sleeping. You can see mistakes in arithmetic when your mind is working properly, while you are making them you cannot see them. Good people know about both good and evil, bad people do not know about either."
– C. S. Lewis

This study is concerned with the application of deep learning to the acoustic analysis of sleep-disordered breathing. Deep learning is regarded as the state-of-the-art in machine learning (Alom et al., 2019). In the present chapter, the principles of sound analysis and machine learning are introduced, and related work is reviewed. It starts with an account of the general sound analysis workflow, and covers the fundamentals of feature extraction, classification and evaluation. Following this, generative and discriminative machine learning methods that can be used for sound analysis are presented. Their advantages and limitations are discussed. In the final part of the chapter, previous attempts to classify sleep-disordered breathing events, mainly snoring, and screen for OSA, the most severe form of sleep-disordered breathing, are reviewed with the aim of identifying what has been achieved in the field, and what needs further improvement.

2.1 Sound Analysis

As mentioned in Chapter 1, the physiological spectrum of sleep-disordered breathing ranges from increased upper airway resistance and partial airway collapse, manifested as snoring and hypopnea events, to complete airway collapse, evidenced as apnoea events. Detecting and classifying these events based on their particular acoustic characteristics has been proposed as an alternative to PSG or HSAT for obtaining relevant information about respiratory conditions associated with sleep. Drawing inspiration from speech technology tools, different approaches to the classification of sleep-disordered breathing

Fig. 2.1 Analysis of sleep-disordered breathing sounds workflow

events have been previously investigated. These have exploited the similarities between speech and breathing sounds discussed in the introductory chapter: both are time-series signals, have a similar frequency range, and are produced in the vocal tract. Although the approaches vary in terms of data, annotation scheme, features, and classifier architecture used, all follow a general workflow to some extent. This is shown in Figure 2.1. Breathing sounds are first recorded and manually annotated. Then, the audio waveform is converted into a temporal sequence of acoustic features, and provided as input to a classifier, which is trained in order to learn to map the features to the expected output (e.g., labels). Lastly, after classification, the performance of the proposed approach is evaluated. In Chapter 3, breathing sounds capture, and manual annotation will be considered. Feature extraction, classifier training, and classifier evaluation will be explained in detail next.

2.1.1 Feature Extraction

Feature extraction consists of transforming the input data (i.e., the audio waveform) into a compressed representation that encodes the important information for the classification task. This can be performed at different levels. Usually, the audio waveform is split into *frames*, since signal processing techniques assume that the statistical properties of a signal slowly vary with time or are stationary. Signals such as speech and breathing sounds are quasi-stationary, that is, they are stationary over very short periods of time or frames (Kwong and He, 2001). For example, 25- or 30-ms frames are typically used in automatic speech recognition (ASR) (Jurafsky and Martin, 2014a), and in the acoustic analysis of sleep-disordered breathing (Duckitt et al., 2016). A frame rate shorter than the frame size is commonly used — for instance, 10 ms — to obtain a better temporal resolution (Rao and Vuppala, 2014), and to preserve information at the frame boundaries that would be lost otherwise. In ASR, frames make up words, whereas in the context of the sleep-disordered breathing, frames make up *events*, for example, a snore or an apnoea, which can last for some seconds. Given that apnoeas are periods of silence that extend over a long duration, a large analysis *segment* is commonly used for either exploiting the temporal pattern of breathing sounds or obtaining an overview of the acoustic characteristics of the signal. For instance, 1-minute segments or even features computed from whole-night audio recordings have been provided as input to classifiers for OSA screen-

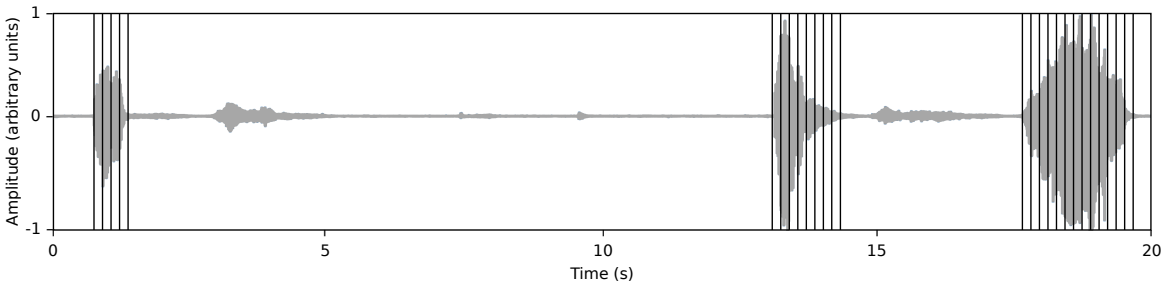


Fig. 2.2 A 20-second *segment* from a sleep audio recording with snore *events* split into *frames* (narrow rectangles). Frames are not to scale.

ing (Kim et al., 2018; Nakano et al., 2019). The relationship between frames, events and segments is illustrated in Figure 2.2.

2.1.2 Classifier Training

The aim of training a classification algorithm is to obtain a mapping from a set of examples to their corresponding labels while also demonstrating good generalisation, that is, being able to accurately predict the labels for unseen examples (Japkowicz and Shah, 2011a). There are two broad categories of machine learning approaches for classification: generative and discriminative methods. Generative methods learn a probability density model over the data and manipulate this model to compute a classification function, whereas discriminative methods directly learn to map the input features to the output labels without modelling the underlying distributions. In this way, only the classification boundary is adjusted in discriminative approaches, and the intermediate step of obtaining a generator that models the data is omitted. This concentrates model and computational resources on the classification task, and commonly results in better performance (Lasserre et al., 2006). However, discriminative techniques commonly require larger amounts of data in comparison with generative techniques, and the former do not make the relationships between variables as explicit as in the latter (Jebara, 2002). Specific generative and discriminative machine learning methods for classification will be considered in detail in the next section.

Machine learning approaches can be also divided into supervised and unsupervised methods. In supervised approaches the learning algorithm has access to the labels of the training examples, whereas in unsupervised approaches it does not. The availability of labels determines the objective of the learning methods. Supervised learning aims to learn a model that provides an output based on the observed input, for example, classifying sleep-disordered breathing events based on acoustic features. Unsupervised learning, on the other hand, attempts to model the input itself, for instance, clustering sleep-disordered breathing events that are similar to each other (Japkowicz and Shah, 2011a).

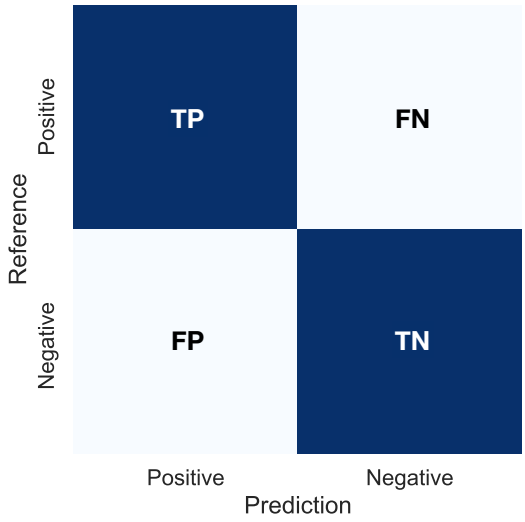


Fig. 2.3 Confusion matrix

2.1.3 Classifier Evaluation

Classifiers are learned on a finite set of data — training data — and evaluated on data different from that used for training — test data. The objective of evaluating the performance of a classifier is to examine its generalisation capability or how well it performs on unseen data. Training and testing a classifier on the same set of data would result in an optimistically biased evaluation, as overfitting the data, for example, commonly leads to poor performance on new data. Since the amount of data to learn and evaluate a classifier is finite, it is typically partitioned using a hold out or a cross-validation approach. The former (randomly) divides the available data into two independent datasets, for instance, 4/5 of the data for training, and 1/5 of it for testing. The cross-validation approach (randomly) partitions the available data into K complementary training and test datasets, and learns and evaluates the classifier K times using different training and test datasets each time.

Scalar Performance Measures

In the context of the acoustic analysis of sleep-disordered breathing, binary classification is the most common setting in which the performance of proposed approaches is evaluated, for example, snore vs. non-snore or OSA vs. healthy. One class is regarded as 'positive' and the other, as 'negative'. Four characteristic values are considered for evaluation: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). True positives and true negatives represent, respectively, the number of samples from the test set that were correctly classified as positive and negative. False positives and false negatives denote, respectively, the number of negative and positive samples that

were incorrectly classified as positive and negative. These values are typically presented in a confusion matrix (Japkowicz and Shah, 2011b). This is illustrated in Figure 2.3.

The true-positive rate of a classifier is also known as sensitivity or recall. In the healthcare domain, this metric reflects the effectiveness of a test in detecting a condition, that is, how sensitive the test is to the presence of the condition or how many of the positive instances are detected by the test. The complement to this metric, the true-negative rate or specificity, focus on the proportion of negative instances or healthy subjects that the test can successfully detect. Another aspect of evaluation is the proportion of instances that really belong to the positive class from all the instances classified or predicted as positive. The positive predictive value or precision measures this. Although these metrics are informative, they can make comparisons between different approaches more difficult. For example, if test X had a higher sensitivity than test Y, but test Y had a higher precision than test X, it would be hard to conclude which test performed better. For this reason, combination metrics such as the F-measure and accuracy are used. The F-measure is the harmonic mean of precision and sensitivity. The accuracy summarises the overall performance by measuring the proportion of correctly classified instances regardless of whether they are positive or negative. However, the accuracy can be misleading when the distribution of instances in each class is skewed (Japkowicz and Shah, 2011b). These metrics are defined as:

$$\text{sensitivity or recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.1}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{2.2}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.3}$$

$$\text{F-measure} = \frac{2 \cdot \text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}} \tag{2.4}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.5}$$

Graphical Performance Measures

In addition to the described scalar metrics, evaluation can be also carried out using graphical performance measures. One of the most commonly used is the receiver operating characteristic (ROC) curve, which allows the visualisation of the classifier performance over all its operating range and, therefore, under different class distribution ratios. The ROC curve plots the relationship between the sensitivity and specificity of

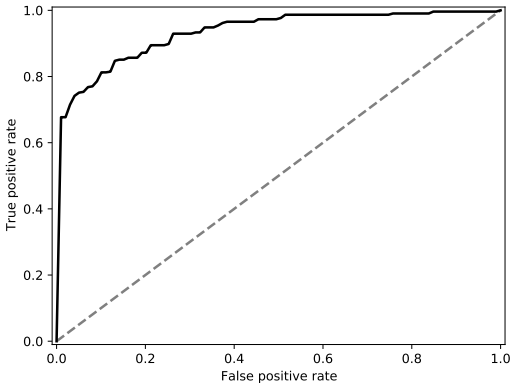


Fig. 2.4 Example of a ROC curve

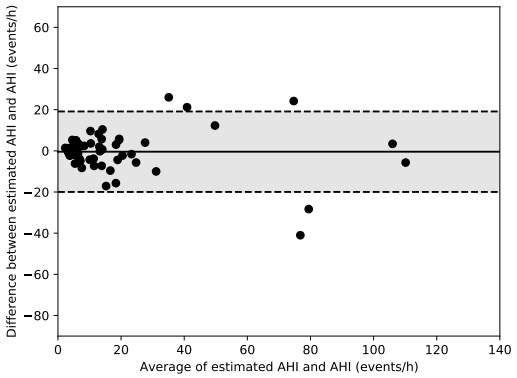


Fig. 2.5 Example of a Bland-Altman plot

the classifier. Specifically, it plots the false-positive rate or 1 – specificity on the horizontal axis and the true-positive rate or sensitivity on the vertical axis at different decision thresholds for classification. Each decision threshold is used to label the instances in the test set. Those instances with a score above the threshold are labeled as positive, whereas the ones below it are labeled as negative. An example is presented in Figure 2.4. The point $(0, 0)$ indicates the trivial classifier that labels all instances as negative, whereas the point $(1, 1)$ denotes the trivial classifier that labels all instances as positive. The point $(0, 1)$ indicates the ideal classifier. The dashed diagonal corresponds to random performance. The ROC curves lying below this diagonal indicate worse than random performance, whereas those above it — such as the one in the example — denote better than random performance.

Although the ROC curve offers a very informative format for evaluation, it is complex to handle in practice. For instance, visual comparisons between different classifiers might be difficult. For this reason, similar to the F-measure, which summarises precision and sensitivity into a single value, the area under the ROC curve (AUC) is widely used for summarising the ROC graph (Scott et al., 1998). The random classifier has an AUC of 0.5, whereas the ideal classifier has an AUC of 1.0. A classifier with better performance than random guessing would have an AUC greater than 0.5. In the example in Figure 2.4, the AUC is 0.95, which indicates a reasonably good performance. Unlike accuracy, which is biased towards the dominant class, the AUC can be a more reliable measure of performance, specially in the case of unbalanced data, since it summarises the classifier performance under different class distribution ratios (Japkowicz and Shah, 2011c).

Another graphical performance measure is the Bland-Altman plot, which is commonly used in medical literature to evaluate the agreement between two different measurement techniques. It plots for each instance in the test set the average of the estimated and reference measurement on the horizontal axis against the difference between the estimated and reference measurement on the vertical axis as a scatter diagram. The mean

difference between all reference and estimated measurements is graphed as a solid line, and the agreement limits, ±1.96 times the standard deviation of the difference between all reference and estimated measurements, are plotted as dashed lines (Bland and Altman, 1986). For instance, Bland-Altman plots are used to evaluate the agreement between the AHI estimated by a classifier from sleep audio recordings and the reference AHI from PSG or HSAT. An example is depicted in Figure 2.5. 92% of the test samples are within the agreement limits, and the mean difference between all reference and estimated measurements is 0 events/h, which indicates a reasonably good agreement between the estimated and reference AHIs.

2.2 Machine Learning Methods for Sound Analysis

In the previous section, we introduced the general idea of generative and discriminative machine learning methods for classification, and discussed their advantages and limitations. In this section, we will consider specific generative and discriminative approaches that have been used for sound analysis with special attention to DNNs. In Section 2.3, studies that have made use of these methods for classifying sleep-disordered breathing events and screening for OSA will be reviewed.

2.2.1 Generative Methods

To reiterate, generative methods model the underlying distributions of data, and compute a classification function based on this modelling. Examples of these include Gaussian mixture model, and hidden Markov model.

Gaussian Mixture Model

The Gaussian or normal distribution is widely used to model the distribution of continuous variables. It is defined by the mean and variance in the case of a single variable, and by the mean vector and the covariance matrix in the multidimensional case:

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{2.6}$$

where x is a D-dimensional vector, μ is the D-dimensional mean vector, and Σ is the $D\times D$ covariance matrix. Since a single Gaussian is usually unable to properly model real data, a linear combination or superposition of Gaussians is used for modelling complex probability distributions. Given a sufficient number of Gaussians with appropriate means, covariances, and coefficients for their linear combination, nearly any probability distri-

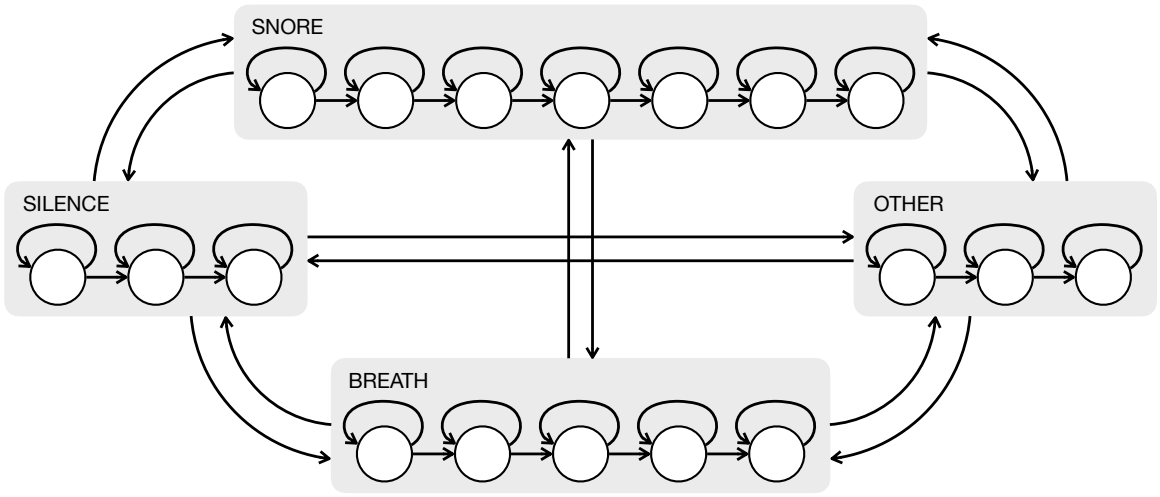


Fig. 2.6 Example of an HMM. Snore, breath, silence and 'other' events are shown as grey rectangles; hidden states, as white circles; and transitions, as arrows.

bution can be approximated. This is known as Gaussian mixture model (GMM), and can be defined as:

$$P(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{2.7}$$

where $N(x|\mu_k, \Sigma_k)$ is the Gaussian distribution k of the K-components mixture $P(x)$ with its mean vector μ_k, covariance matrix Σ_k, and mixing coefficient π_k (Bishop, 2006a).

Hidden Markov Model

A hidden Markov model (HMM) allows the representation of probability distributions over a sequence of observations. That is, an HMM model infers a sequence of hidden states from a sequence of observations. An HMM models a sleep-disordered breathing event as a sequence of states. Each state corresponds to a small portion of the event, and is represented by a GMM, which models the probability of a given observation — a vector of acoustic features — being generated from a particular state. Hence HMMs are a generative method. The probability of moving from one state to another is modelled with a transition probability matrix (Jurafsky and Martin, 2014a). This is depicted in Figure 2.6.

2.2.2 Discriminative Methods

As pointed out before, discriminative methods do not model data explicitly — as generative methods do — but directly map the input features to the output. Logistic re-

gression, support-vector machine, AdaBoost, and DNNs are examples of discriminative approaches.

Logistic Regression

Logistic regression is a discriminative approach with linear parameters used for binary classification, which makes it useful to classify between snore and non-snore events, for example. It can be defined as:

$$P(l|x, w) = \text{Ber}(l|\text{sigmoid}(w^T x)) \tag{2.8}$$

where $l \in \{0, 1\}$ is the label or class, x is the vector of features, w is the vector of regression weights, Ber is the Bernoulli distribution defined in Equation 2.9, and sigmoid is the logistic function defined in Equation 2.10.

$$P(l = 1) + P(l = 0) = 1 \tag{2.9}$$

$$\text{sigmoid}(x) = \frac{1}{1 + \exp(-x)} \tag{2.10}$$

A linear decision boundary is obtained with a threshold of 0.5. That is, if the output of the classifier is below 0.5, the acoustic features are assigned to the class $l = 0$. Otherwise, the acoustic features are assigned to the class $l = 1$. An advantage of logistic regression is that it does not require large amounts of training data, as only the vector of regression weights has to be learned, and its dimensionality is equivalent to the number of features (Murphy, 2012a).

Support-vector Machine

Similar to logistic regression, support-vector machine (SVM) is a discriminative approach used for binary classification, and requires a small amount of training data. It maps the input vectors to a high dimensional feature space to construct a hyperplane that ensures the generalisation of the classifier. The optimal hyperplane is the linear decision function with the maximum margin between the vectors of the two classes. A very small amount of data, the support vectors, is required to find the hyperplane. This is illustrated in Figure 2.7 (Cortes and Vapnik, 1995).

AdaBoost

AdaBoost, which stands for Adaptive Boosting, is a method for improving the performance of any learning algorithm. It leverages 'weak' learners — learning algorithms that

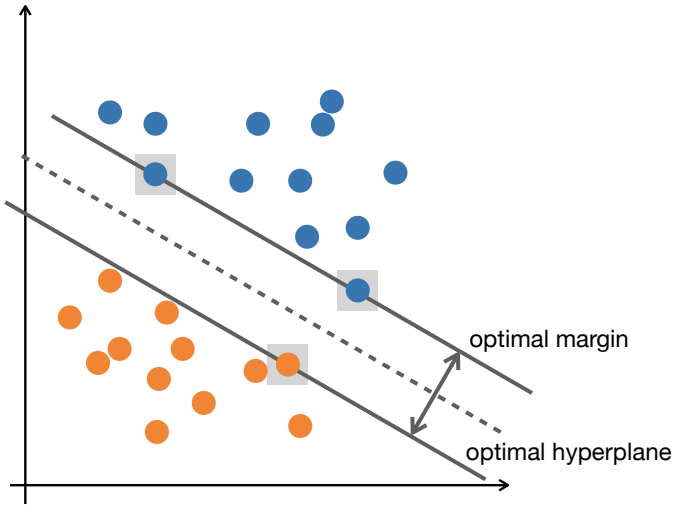


Fig. 2.7 Visualisation of a SVM in the 2-dimensional space. The support vectors are inside grey squares, and define the maximum margin between the two classes, shown as blue and orange dots, respectively. The dashed line corresponds to the hyperplane.

do slightly better than random guessing — for binary classification, similar to logistic regression and SVM. In this method, the classification decision is calculated as a weighted summation of such weak learners, and the output is either -1 or 1. It can be defined as:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{2.11}$$

where α_t is the weighting parameter of weak learner t, $h_t(x)$ is the hypothesis or output of weak learner t given the vector of acoustic features x, T is the number of weak learners, and $H(x) \in \{-1, 1\}$ is the final hypothesis or output of the classifier given the vector of acoustic features x (Freund and Schapire, 1999).

Deep Neural Networks (DNNs)

Deep learning draws inspiration from the human brain, which has many levels of processing that learn representations at increasing degrees of abstraction. For instance, the brain performs vision by first extracting "edges, then patches, then surfaces, then objects, etc." (Murphy, 2012b, p. 995). Deep learning aims to reproduce this architecture in a computer by using DNNs: artificial neural networks with multiple layers of artificial neurons.

An artificial neural network is a mathematical model inspired in the human brain. Our brain consists of around 100 billion interconnected units, known as neurons, forming a network (Lent et al., 2012). A neuron is a specialised nervous cell that receives, processes, and transmits information in the form of electrical pulses. Its dendrites receive

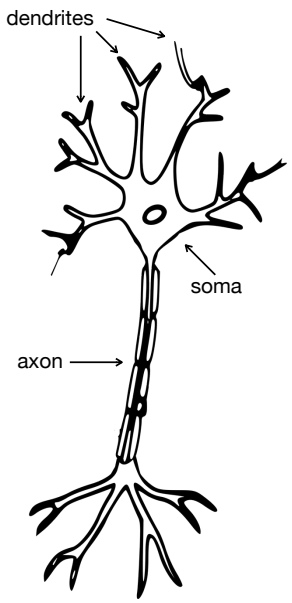


Fig. 2.8 Biological neuron

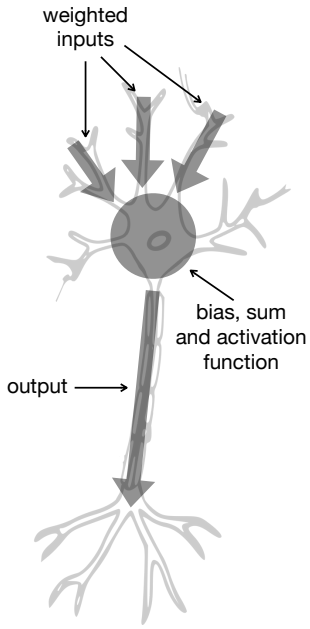


Fig. 2.9 Artificial neuron

information from other neurons, its soma processes it, and its axon transmits the result to other neurons. This is illustrated in Figure 2.8. Likewise, an artificial neuron, the basic building block of an artificial neural network, receives weighted information (i.e., input values) from other artificial neurons, processes them by applying an activation function, and outputs the processed information, which may become the input to other artificial neurons. This is shown in Figure 2.9. An artificial neuron can be defined as:

$$y = F\left(\sum_{i=1}^{M} w_i \cdot x_i + b\right) \tag{2.12}$$

where x_i is the input value i of M inputs, w_i is the weight value for input i, b is the bias value, F is the activation function, and y is the output. The activation function can be a linear or non-linear function, and is selected depending on the specific application of the artificial neural network (Krenker et al., 2003).

An artificial neural network is obtained by combining several artificial neurons. These are interconnected in layers. Similar to the neurons in our brains, the way artificial neurons are interconnected can be of two forms: feed-forward or recurrent (Murphy, 2012b). In feed-forward neural networks the information flows in one direction: from inputs to outputs, whereas in recurrent neural networks the information also runs in the opposite direction: from outputs to inputs (Krenker et al., 2003). Like logistic regression, SVM, and AdaBoost, artificial neural networks are a discriminative machine learning approach. When an artificial neural network consists of multiple layers, it is called a DNN.

The layers between the input and output layers are referred to as hidden layers (Schmidhüber, 2015).

In the same way that our brains need to learn responses to inputs or stimuli from the environment, artificial neural networks need to learn them from data. This is achieved by applying supervised or unsupervised learning. As previously discussed, supervised learning consists in providing labelled data to the network, whereas for unsupervised learning no labels are provided to it. The aim of learning is to find the appropriate network parameters — weights and biases — for mapping the input into the expected output (Krenker et al., 2003). Specifically, the network parameters are found using the gradient descent algorithm to minimise a loss function (also known as cost or error function). The loss function computes the error between the output of the network and the expected output for a particular input, and is selected based on the task of the neural network. For instance, for classification problems, categorical crossentropy is commonly used:

$$\text{cce}(y, \hat{y}) = -\sum_{i=1}^{n} y_i \cdot \log \hat{y}_i \tag{2.13}$$

whereas for regression problems, mean squared error is typically employed:

$$\text{mse}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.14}$$

where y and $\hat{y}$ are n-dimensional vectors, y is the expected output, and $\hat{y}$ is the output of the network (Goodfellow et al., 2016). The network parameters are randomly initialised, and updated through multiple iterations on the training data — otherwise known as epochs — based on the gradient of the loss function with respect to the network parameters:

$$\theta_{t+1} = \theta_t - \alpha \frac{\partial L(X, \theta_t)}{\partial \theta} \tag{2.15}$$

where θ_t denotes the network parameters at iteration t, L is the loss function, X is an input-output pair, α is the learning rate, and θ_{t+1} indicates the updated network parameters. The learning rate is a hyper-parameter that controls the step size towards a minimum of the loss function. If it is too large, the loss function can fail to converge, whereas if it is too small, convergence will be very slow (Bengio, 2012). Successive updates to the network parameters are performed until the loss function converges, and the gradient is efficiently computed making use of the backpropagation algorithm (Rumelhart et al., 1986). It computes the gradient with respect to each network parameter using the chain rule (of calculus) to reuse computations. The backpropagation algorithm starts the computation from the final layer and propagates the losses backwards — hence the name. The

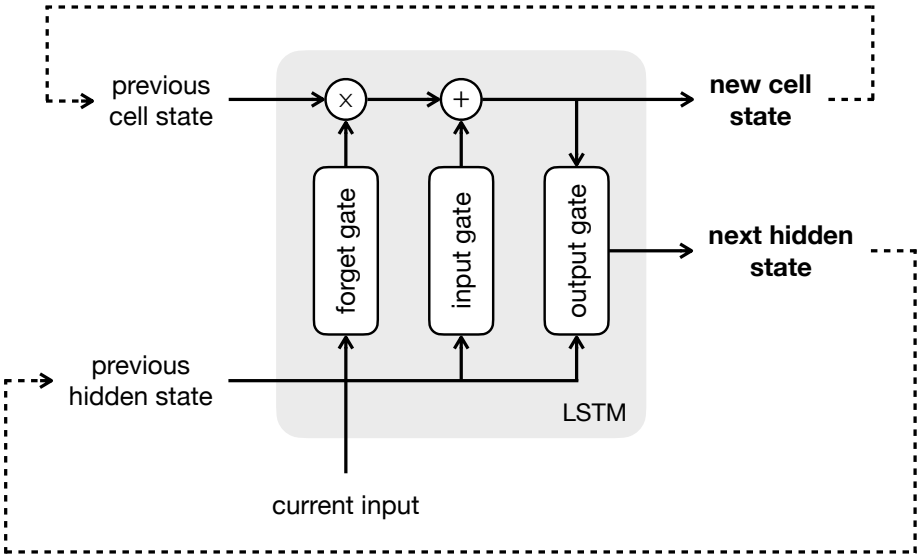


Fig. 2.10 LSTM cell

reader is referred to (Goodfellow et al., 2016) or (LeCun et al., 2012) for a deeper treatment of backpropagation.

Throughout this thesis, DNNs will be employed for a variety of tasks such as learning novel feature representations, recognising snorers, and screening for OSA. We will now look at two widely used DNN architectures: long short-term memory neural networks, and convolutional neural networks. A long short-term memory (LSTM) neural network is a recurrent DNN architecture proposed by Hochreiter and Schmidhüber (1997) with the aim of allowing artificial neural networks to store information over extended time intervals. For this reason, LSTMs are particularly useful when dealing with sequential or time-series data, like breathing sounds, for example. As illustrated in Figure 2.10, an LSTM cell has a state and three different gates. Its state is the 'long-term memory' of the network, as it carries previous relevant information throughout the processing of an input sequence. The cell state is updated by the gates, which are neural networks that learn what information should be kept or forgotten. These gates use the hyperbolic tangent (tanh) and sigmoid activation (Equation 2.10) functions.

The first gate, known as the 'forget' gate, determines what previous information should be kept or forgotten. It passes the current input and the previous hidden state (i.e., previous output or the 'short-term memory' of the network) through the sigmoid function, which outputs values between 0 and 1. If its output is close to 1, the information is kept, whereas if its output is close to 0, the information is forgotten. The second gate, known as the 'input' gate, updates the cell state based on the current input. In addition to passing the current input and the previous hidden state through the sigmoid function, it passes them through the tanh function to define a new cell state candidate,

and multiplies the output of both activation functions. The previous cell state is multiplied by the output of the forget gate, after which the output of the input gate is added to it. This updates the cell state to the new values that the network has found relevant. Lastly, the third gate, known as the 'output' gate, determines the next hidden state (i.e., current output). It passes the current input and the previous hidden state through the sigmoid function, and the new cell state through the tanh function. Lastly, the output of both functions is multiplied. The result of this multiplication is the next hidden state, which along with the new cell state is passed to the next time step (Kumar, 2020). Later in this thesis, we will use LSTMs for identifying and predicting the number of active snorers in 2-snorer sleep audio recordings.

A convolutional neural network (CNN) is a network architecture inspired by the organisation of the human visual cortex. So, it is mainly used for image-driven pattern recognition tasks (O'Shea and Nash, 2015). CNNs usually consist of three types of layers: convolutional, pooling, and fully connected layers. Convolutional and pooling layers perform feature extraction, whereas fully connected layers map the extracted features into the expected output. A convolution is a linear operation in which a kernel or filter — a matrix — is applied across the input tensor to extract features. Specifically, the element-wise product of the kernel and the input tensor is computed at every position of the input, and summed. This is illustrated in Figure 2.11 for a 2-dimensional input tensor, and a 3×3 kernel with a stride of 1×1. During training the kernels that work best for the task at hand are learned from data. For example, some kernels highlight vertical or horizontal lines, other kernels, circular shapes, etc. The result of the convolution operation is then passed through an activation function to model non-linearities between the input features and the expected output. The most commonly used function is the rectified linear unit (ReLU) activation:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{2.16}$$

Convolutional layers are followed by pooling or downsampling layers, which reduce the dimensionality of the features extracted by the convolutional layers with the aim of lessening the number of learnable parameters, and introducing translation invariance to small distortions and shifts. Two pooling operations are commonly used: max-pooling, and global average pooling. Max-pooling takes out patches from the features extracted by the convolutional layers, and outputs the maximum value in each patch. This is depicted in Figure 2.12. Global average pooling downsamples the output of the convolutional layers to a 1×1 tensor by averaging out all the elements in it. This is illustrated in Figure 2.13. This operation allows the network to accept inputs of different size, and decreases the number of learnable parameters in subsequent layers. Global average pooling is typically performed before the fully connected layers. It is worth noting that there are no learn-

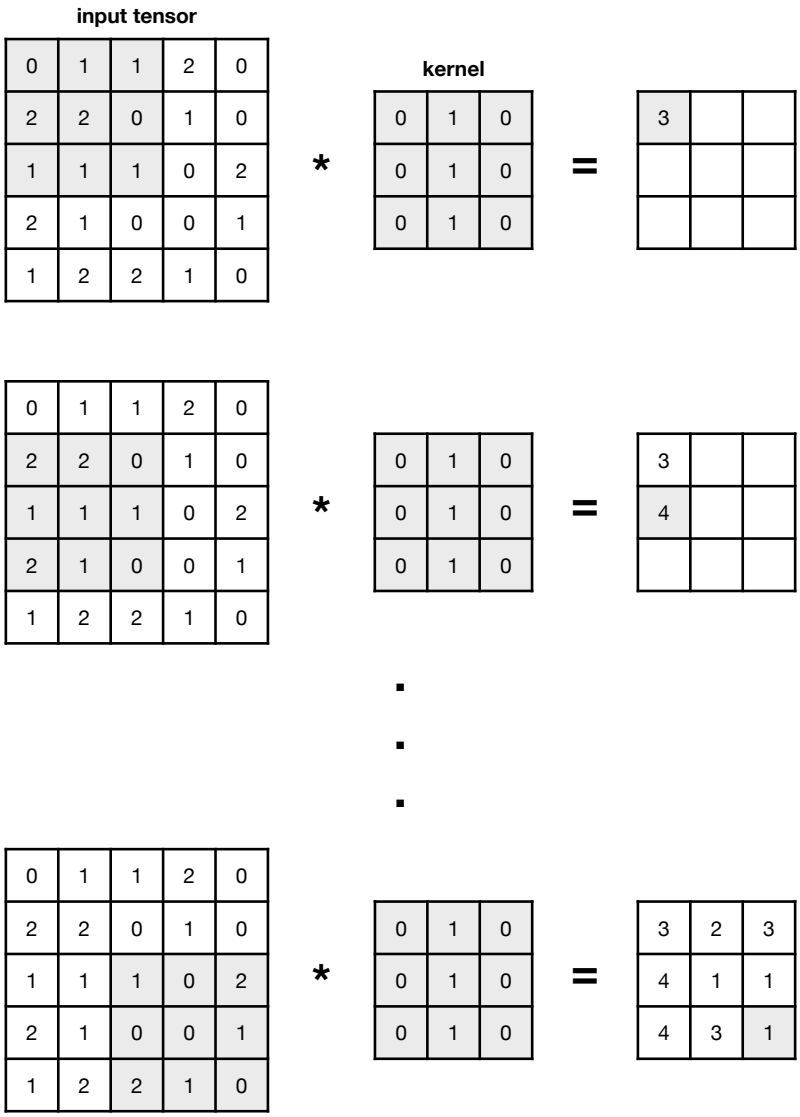


Fig. 2.11 Example of a 3×3 convolution operation with a stride of 1×1

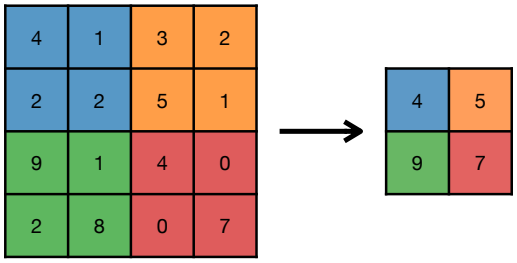


Fig. 2.12 Example of a 2×2 max-pooling operation with a stride of 2×2

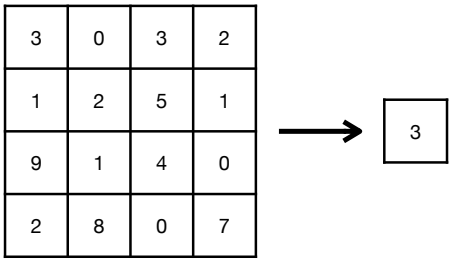


Fig. 2.13 Example of a global average pooling operation

able parameters in the pooling layers. Lastly, the result of the convolutional and pooling layers is provided as input to the fully connected layers, which map the extracted features into the expected output (Yamashita et al., 2018). The softmax activation function is commonly used in the last layer of a DNN that performs classification, as it normalises the output of the network to a probability distribution. In this way, the output of the network can be interpreted as the probability of the input belonging to each class in consideration. The softmax activation, also known as the normalised exponential function, can be defined as:

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{D} \exp(x_j)} \tag{2.17}$$

where x_i is the element i of the D-dimensional vector x (Bishop, 2006b).

2.3 Related Work

Having introduced machine learning methods that can be used for sound analysis, we will now review related studies that have employed these methods for classifying sleep-disordered breathing events — mainly snoring — and screening for OSA. The reviewed studies will be related to the previously presented machine learning methods based on the notation used to introduce them.

2.3.1 Classification of Sleep-disordered Breathing Events

Classical Machine Learning Methods

Nonaka et al. (2016) developed a logistic regression classifier to detect snoring in sleep audio recordings. They collected data from 40 participants with a microphone placed 50 cm above the subject's head during PSG. Snoring was defined as "loud breathing during sleep", and marked in the audio recordings by three annotators. Having to provide their own description of snoring evidences the lack of a widely agreed definition of it, which this thesis will attempt to address from an acoustic point of view in Chapter 3. Also, their definition might lead to a very subjective annotation as it does not specify how loud breathing has to be for being regarded as snoring.

Statistical features were obtained from an auditory-motivated time-frequency representation of the audio signal, and provided as input to the classifier. These included, among others, average kurtosis, average spectral centroid, average spectral bandwidth, and average spectral contrast. Considering Equation 2.8, in the study by Nonaka et al., $l = 1$ was the snore class, $l = 0$ was the non-snore class, x were the statistical features, and w was the vector of regression weights learned from the collected data. A sensitivity

of 97% with a specificity of 96% were reported. However, it is important to bear in mind that the audio recordings were made in controlled conditions: a sleep laboratory. This might limit the performance of the developed system in realistic sleep conditions, as the data used was collected in the same acoustic environment, and background noise levels and room acoustics might differ in a bedroom at home.

In another study by Sun et al. (2015), a SVM classifier was implemented for the same task. In this case, the sleep audio recordings were collected from a smaller number of participants, but made in realistic sleep conditions. They collected data from four subjects with a portable recorder in a bedroom at home. Also, a more detailed annotation scheme was used. Snoring on inspiration, snoring on expiration, silence, and 'other' events were manually annotated. Means and standard deviations of the spectral energy of 3 sub-bands at 50–300 Hz, 50–550 Hz, and 50–800 Hz calculated on 128-ms windows with 50% overlap were provided as input to the classifier.

Although SVM was originally proposed as a binary classifier, Sun et al. used an SVM for multi-class classification. They implemented a one-versus-all approach (Murphy, 2012c), in which four different classifiers — one for each class considered — were trained. The data from one class was regarded as positive, and the data from all the other classes was regarded as negative. This was done for each class. Given the limited amount of data available, leave-one-participant-out cross-validation was carried out: the data from one participant was used for testing the classifier, and the data from the remaining three participants to train it. This was done for each participant. A recall of 100% was reported. However, approximately 23 minutes of sleep audio recordings were used — less than 6 minutes from each of four participants collected during a single night. Therefore, the developed system might not generalise well due to the small amount of data used and the proposed handcrafted features being based on it. This study reflects one of the challenges of working on the acoustic analysis of sleep-disordered breathing: labelled data is scarce, and manually annotating sleep audio recordings is time-consuming and expensive. Even though 32 hours of data were recorded, only 23 minutes were manually annotated. However, as we previously discussed, unsupervised learning is one way of making use of unlabelled data. In Chapters 4 and 5, we will leverage large amounts of unlabelled sleep audio recordings to learn novel feature representations by applying unsupervised learning.

In the same way as Sun et al. (2015), Demir et al. (2018) used an SVM for multi-class classification in a slightly different task. Using the Munich-Passau corpus (Janott et al., 2018), they developed a system to predict the obstruction site in the upper airway that causes snoring. Data from 219 participants who underwent drug-induced sleep endoscopy[1] was collected to create the Munich-Passau corpus. It consists of over 800 snore

[1]Drug-induced sleep endoscopy is a technique for visually examining the upper airway during sleep. Examination is performed with an endoscope introduced through the nose of the subject under sedation (Kezirian et al., 2011).

events from four obstruction sites annotated by otolaryngologists: oropharyngeal lateral walls, velum, epiglottis, and tongue base. Features computed from spectrogram images were provided as input to the classifier. Later in this thesis we will also use spectrogram images as input to a classifier that predicts the presence or absence of apnoea-hypopnea events. Demir et al. reported an unweighted average recall of 73%. They implemented a SVM because of its robustness to limited amounts of training data, since — as we saw before — only the support vectors are required to find the hyperplane. One of the limitations of this study is that the data used might not epitomise natural sleep physiology, as the participants were under sedation during drug-induced sleep endoscopy (Kezirian et al., 2011). Then the performance on snoring occurring during natural sleep remains to be investigated.

The Munich-Passau corpus was used in the Snoring Sub-challenge of the INTERSPEECH 2017 Computational Paralinguistics Challenge (ComParE). As the previous study, it aimed to predict the obstruction site in the upper airway that causes snoring. The winners of the challenge, Kaya and Karpov (2017), implemented a weighted partial least squares regression classifier. In the fashion of SVM, this classifier projects the input vectors to a high dimensional feature space, but — unlike SVM — it finds a linear regression model instead of constructing a hyperplane (Wold, 1973). Acoustic features including mel-frequency cepstral coefficients (MFCCs) along with deltas and accelerations[2] were provided as input to the classifier. 2-fold cross-validation was performed: the available data was partitioned into 2 datasets — training and development — and the classifier was learned and evaluated twice using different training and development datasets each time. Contestants did not have access to the labels of the test dataset. In Chapter 6, we will consider a 2-fold cross-validation approach too with the aim of validating the performance of a classifier trained using a limited amount of data. An unweighted average recall of 64% was reported by Kaya and Karpov. This study has the same limitation as that by Demir et al. (2018): natural sleep physiology might not be properly represented by the data used.

An AdaBoost classifier was developed by Dafna et al. (2013) for the same task as the first two studies. Similar to Nonaka et al. (2016), they collected sleep audio recordings from 67 participants with a microphone placed 1 m above the subject's head during PSG. It was recognised that there is no widely accepted definition of snoring. So, for annotation purposes, they defined snoring as a "breathing sound during inspiration". In Chapter 3, we will see that snoring also occurs during expiration, although less frequently. In the same way as Nonaka et al., the sleep audio recordings were annotated by three observers. 34 features from the time and frequency domains were considered. These included, among

[2]The extraction of MFCCs, and the computation of deltas and accelerations are considered in detail in Appendix A.

others, periodicity, total energy, relative energy before an event, duration, signal spectrum, vocal tract parameterisation, and dynamics of frequencies. Unlike the first two studies, an event detection step was implemented before extracting features, since sleep audio recordings mainly consist of silence or low background noise, as will be seen in the next chapter. It detected sound events with a duration of 0.2 to 3.5 seconds with an intensity of 25 to 75 dB. Features were only extracted for these events, and provided as input to the classifier.

Considering Equation 2.11, in the study by Dafna et al., $H(x) = 1$ was the snore class, and $H(x) = -1$ was the non-snore class, x were the features from the time and frequency domains, and α_t was the weighting parameter for weak learner t learned from the collected data. Using 100 weak learners, $T = 100$, was found to be a reliable and efficient approach. A sensitivity and a specificity of 98% were reported. However, similar to the study by Nonaka et al., it is important to take into account that the sleep audio recordings were collected in controlled conditions (i.e., a sleep laboratory), which potentially limits the performance of the proposed system in typical sleep conditions (i.e., a bedroom at home), where room acoustics and background noise levels might be different.

Up to this point, we have looked at studies that have implemented discriminative machine learning methods for the classification of sleep-disordered breathing events. We will now consider a study that implemented a generative machine learning method. Duckitt et al. (2016) developed an HMM classifier. Unlike the previous study, they collected sleep audio recordings from six subjects in typical sleep conditions with a microphone placed on the bedside table. Similar to the study by Sun et al. (2015), a detailed annotation scheme was used: snoring, breathing, duvet noise, silence, and 'other' noise were marked in the collected data. Energy, MFCCs along with deltas and accelerations were used as features. These were computed on 30-ms frames with a frame rate of 10 ms, and provided as input to the HMM classifier. Snoring was modelled by a sequence of 3 states, given the great variability of snore events, whereas all other classes were modelled by 1 state. The observation distribution of each HMM state was modelled by an 8-component GMM. A recall of 82% was reported. It is worth noting that, like previous studies, specialised hardware — high-quality microphones — was employed to collect the sleep audio recordings, which might hinder the wide use of the proposed system. Later in this thesis we will implement an HMM as a baseline system for the same task using data collected with readily available hardware — smartphones.

Deep Learning Methods

Thus far, we have reviewed studies that have used classical machine learning techniques for classifying sleep-disordered breathing events. We will now consider what is regarded as the state-of-the-art in machine learning, and the focus of this thesis: deep learning.

Emoto et al. (2018) developed a DNN for the classification of sleep-disordered breathing events. As the studies by Nonaka et al. (2016) and Dafna et al. (2013), sleep audio recordings were collected from 20 subjects during PSG with a microphone placed 50 cm above the participant's head, and annotated by three human observers. 400-sample frames were directly provided as input to the DNN. A feed-forward DNN with three layers was trained by applying supervised learning. Hyperbolic tangent was used as the activation function in the hidden layers to model the non-linear relationship between the input and output. Linear activation was employed in the output layer, since the expected output was either 0: background noise, or 1: sound activity. The frames classified by the DNN as sound activity were further classified as either snore or non-snore frames based on energy. If the energy of a given frame was above a threshold, it was classified as snore. Otherwise, the frame was classified as non-snore. Such a threshold was optimised on the ROC curve. A sensitivity of 76% with a specificity of 78% were achieved. The performance reported is lower with respect to the previously reviewed studies, which have used traditional machine learning techniques. This evinces one of the limitations of deep learning. Although deep learning methods can model complex relationships between the input and output, they require considerably more training data than traditional machine learning techniques to properly generalise (Schröder et al., 2016).

As the previous study, Xie et al. (2021) also implemented a DNN to detect snore events. For this purpose, they collected sleep audio recordings from 38 participants during PSG, and snore events were annotated by a sleep technician. Sound events were first detected, and split into 3.5-second segments. Then, similar to Demir et al. (2018), spectrogram images were computed for each segment, and provided as input to the DNN. Xie et al. trained a recurrent neural network with convolutional and LSTM layers. The LSTM layer consisted of 64 units or cells, and softmax activation (Equation 2.17) was used in the output layer. The number of artificial neurons in the output layer was two — one neuron for every class considered: snore, and non-snore. A sensitivity of 92% and a specificity of 98% were obtained. Even though the DNN architecture implemented differs from the one used by Emoto et al. (2018), Xie et al. had nearly twice as much data as Emoto et al., which facilitated the proper generalisation of the network, since deep learning methods require a great amount of data. However, similar to previous studies, the sleep audio recordings were performed in a sleep laboratory, which would potentially limit the performance of the proposed system in a bedroom at home, as has been discussed before.

All of the studies covered here — with the exception of those by Demir et al. (2018) and Kaya and Karpov (2017), which employed the Munich-Passau corpus — have used their own data, and it has not been made available to others for research. On one hand, it evidences the need for data collection for the study encompassed in the present thesis. The available dataset — the Munich-Passau corpus — consists of isolated snore events

recorded with specialised hardware in a sleep clinic. However, we require whole-night audio recordings made in typical sleep conditions (e.g., a bedroom at home) with readily available hardware (e.g., a smartphone), as data collected in these settings would be representative of the intended real-world use of the systems to be developed in our study. This will be addressed in the next chapter. On the other hand, direct comparison of the performance of different studies is not entirely appropriate.

Finally, a summary of the studies reviewed in this section is presented in Table 2.1.

Table 2.1 Classification of sleep-disordered breathing events

Study	Dafna et al. (2013)	Sun et al. (2015)	Duckitt et al. (2016)	Nonaka et al. (2016)	Kaya et al. (2017)	Demir et al. (2018)	Emoto et al. (2018)	Xie et al. (2021)
Task	Detection of snore events	Classification of sleep-disordered breathing events	Classification of sleep-disordered breathing events	Detection of snore events	Prediction of obstruction site	Prediction of obstruction site	Detection of snore events	Detection of snore events
Audio recordings	Ambient sound at clinic	Ambient sound at home	Ambient sound at home	Ambient sound at clinic	Drug-induced sleep endoscopy (video recordings)	Drug-induced sleep endoscopy (video recordings)	Ambient sound at clinic	Ambient sound at clinic
Data	67 subjects	4 subjects	6 subjects	40 subjects	219 subjects	219 subjects	20 subjects	38 subjects
Features	34 features from the time and frequency domains	Means and standard deviations from 3 frequency sub-bands	MFCCs, deltas and accelerations	Statistics from an auditory-motivated time-frequency representation	MFCCs, deltas and accelerations	From spectrogram images	Raw waveform and energy	Spectrogram
Classifier	AdaBoost	SVM	HMM	Logistic regression	Weighted partial least squares regression	SVM	DNN	LSTM
Results	Sensitivity: 98% Specificity: 98%	Recall: 100%	Recall: 82%	Sensitivity: 97% Specificity: 96%	Unweighted average recall: 64%	Unweighted average recall: 73%	Sensitivity: 76% Specificity: 78%	Sensitivity: 92% Specificity: 98%

2.3.2 Screening for Obstructive Sleep Apnoea

In the previous section we looked at studies that have used machine learning for classifying sleep-disordered breathing events — mainly snoring — in sleep audio recordings. In the present section we will consider studies that have developed systems to screen for OSA, the most severe form of sleep-disordered breathing, from sleep audio recordings or speech. Unlike snoring, apnoeas are (nearly) silent events with a relatively long duration. By definition, they last for at least 10 seconds (American Academy of Sleep Medicine, 2020). However, the temporal pattern in which apnoeas take place provides salient cues that indicate their occurrence, for instance, the collapse of the upper airway leading to an apnoea evidenced as snoring before the event, and the resumption of breathing after it with a loud gasp or recovery breath (Maas et al., 2011). As we will see, many studies have proposed rule-based methods to screen for OSA rather than machine learning approaches. The latter might be more robust than the former, since handcrafted rules are unlikely to effectively capture the great variability of breathing sounds during sleep, and deal with different room acoustics and levels of background noise. For this reason, we will pay special attention to studies that have used machine learning to screen for OSA. But first, for the sake of completeness, we will succinctly consider those studies that have proposed rule-based methods.

Rule-based Methods

Almazaydeh et al. (2013) used voice activity detection (VAD) for detecting silent segments of audio. VAD uses acoustic features such as energy, zero-crossing rate, and spectral entropy to differentiate between voice-active and silent segments in speech recordings. They recorded breathing sounds from 50 awake subjects who held their breath to simulate apnoeas. If a detected silent segment had a duration of more than 15 seconds, it was marked as an apnoea. An accuracy of 97% was reported. However, this study did not actually work on *sleep* apnoea, and detecting apnoeas exclusively based on the duration of silent segments is a fragile approach, since a long period of quiet respiration, for example, would be incorrectly regarded as an apnoea.

Unlike the previous study, Alakuijala and Salmi (2016) did analyse breathing sounds during sleep. They collected audio recordings from 211 participants during HSAT, and used the percentage of snoring in a night with respect to the total time in bed to screen for OSA. That is, predicting whether a subject has clinically relevant OSA — an AHI $\geq$ 15 events/hour — or not. If the percentage of snoring was above 15%, the participant was screened as positive. A sensitivity of 93% with a specificity of 35% were reported, which evidences the limitations of the proposed approach. While it is true that most subjects

with OSA snore, not all snorers have OSA (Maimon and Hanly, 2010). Therefore, using snoring on its own to screen for OSA has high sensitivity, but poor specificity.

Acoustic evidence is commonly integrated with additional (physiological) parameters to screen for OSA, as the long-duration and silent nature of apnoea events makes screening for OSA using acoustic evidence alone a very challenging task. Al-Mardini et al. (2014) used tracheal respiratory sounds, oxygen saturation, and body movement to screen for OSA. They collected data from 15 participants with sensors paired with a smartphone during PSG. The smartphone was placed on the participant's arm to record body movement with the built-in accelerometer, and the microphone was attached to their throat. Apnoea events were detected as audio recording segments in which the energy was below 90% of the average for at least 10 seconds. These were used for calculating the AHI. On the other hand, the oxygen saturation signal was employed to calculate the oxygen desaturation index (ODI) — the average number of oxygen desaturations per hour. If an oxygen desaturation was associated with body movement, it was not taken into account, as body movement during sleep can cause desaturations. The average of the AHI and ODI was used for screening. A sensitivity of 100%, and a specificity of 86% were obtained. Although the proposed approach integrates breathing sounds with a key physiological parameter, it is obtrusive and requires specialised hardware, which potentially limits its wide use. Additionally, it was tested on a considerably small population. So, the set of rules might not properly generalise.

Similar to the previous study, Castillo-Escario et al. (2019) used breathing sounds and oxygen saturation for detecting apnoeas and hypopneas. They recorded breathing sounds with a smartphone from 13 subjects during HSAT, like Alakuijala and Salmi (2016). This is a strength of both studies, as data was collected in typical sleep conditions, where these alternative methods are expected to be used. So, it avoids the mismatch between the data used for developing the proposed methods, and real-world data. The oxygen saturation signal was obtained from the HSAT device, which was placed on the participant's chest along with the smartphone. This signal was used for extracting audio recording regions associated with a desaturation. Then, individual apnoea-hypopnea events were detected in such regions as silent segments based on sample entropy. Sample entropy is a measure of the complexity of a signal. For example, snoring regions have higher sample entropy values than silent regions, since the former display more complex patterns than the latter. Using the detected apnoea-hypopnea events, the AHI was estimated as follows:

$$\mathrm{AHI}_{\mathrm{estimated}} = \frac{\text{number of apnoea-hypopnea events}}{\text{duration of the audio recording}} \tag{2.18}$$

A sensitivity and a specificity of 100% were reported when using the estimated AHI to screen for OSA. However, it is important to take into account that the proposed approach was tested on a considerably small population, requires specialised hardware, and is ob-

trusive. Therefore, this rule-based approach might not generalise properly, and has the same limitations as PSG.

Narayan et al. (2018) also used breathing sounds recorded with a smartphone to screen for OSA. But, in this study, it was placed on a bedside table, and no physiological parameters were employed. Audio recordings were made from 91 participants during PSG. Three sleep laboratories were used. PSG data was scored by a registered polysomnographic technologist and used as reference. Breathing sounds and apnoea events were detected as follows. For every non-overlapping 120-second audio recording segment:

1. Compute the fast Fourier transform (FFT).
2. Calculate the median of the spectral power for every millisecond of the FFT to obtain a spectral magnitude-time series.
3. Calculate the root mean square (RMS) for every non-overlapping 300-sample window of the spectral magnitude-time series to get an RMS envelope.
4. Detect the peaks in the RMS envelope to identify an array of maxima.
5. Compute the area of each array of maxima using the trapezoidal rule.
6. Define a threshold based on the scored data to classify each area as either breathing or non-breathing sounds.
7. Detect apnoea events as cessations of breathing sounds for more than 10 seconds.

For a visualisation of this algorithm, the reader is referred to Figure 1 of (Narayan et al., 2018). After having detected apnoea events, an 'acoustic respiratory index', analogous to the AHI, was calculated from the number of detected events and the duration of the audio recording. Based on a threshold optimised on the scored data, participants were screened for clinically relevant OSA, like Almazaydeh et al. (2013), with a sensitivity of 94% and a specificity of 63%. Although this study screened for OSA only using breathing sounds recorded unobtrusively with readily available hardware, this was done in controlled conditions, and based on a handcrafted set of rules. Then, as other studies, its performance in realistic sleep conditions, where it would have to deal with different room acoustics and levels of background noise, remains to be investigated.

Similar to Al-Mardini et al. (2014), Saha et al. (2020) used tracheal respiratory sounds, oxygen saturation, and body movement for estimating the AHI. Body movement and respiratory sounds were recorded from 69 participants using a custom-made wearable device during PSG. This device consisted of a 3-dimensional accelerometer and a microphone, and was placed on the participant's throat. As the previous study, PSG data was scored by sleep technicians and employed as reference. The oxygen saturation signal

was obtained from the PSG device. Individual apnoea-hypopnea events were detected as follows. Breathing and snoring events were first detected in the audio signal. Desaturations — drops in the oxygen saturation signal — were employed to look for potential apnoea-hypopnea events in the audio and accelerometer signals, as breathing sounds and body movement are reduced during such events. Then, using 10-second windows with 80% overlap, features from the audio, oxygen saturation, and accelerometer signals were computed for the analysis segment: from 20 seconds before the start of the desaturation to its lowest point. These included drop in oxygen saturation, reduction in respiratory-related movements in the z- and x-axis of the accelerometer, percentage of breathing in the analysis segment, and mean energy of the breathing events in the analysis segment.

The features were normalised, weighted, and summed to get a score between 0 and 1. Such weights were selected based on the importance of each feature. Using a threshold, a segment was classified as an apnoea-hypopnea event or discarded. In the same way as Castillo-Escario et al. (2019), the AHI was estimated with Equation 2.18. The performance was evaluated with a Bland-Altman plot (Bland and Altman, 1986): 94% of the estimated AHIs were within the agreement limits. A sensitivity of 91% with a specificity of 89% were reported when using the estimated AHIs to screen for clinically relevant OSA. The main strength of the study by Saha et al. is the detection of individual apnoea-hypopnea events, since the AHI can be directly estimated. This might facilitate the adoption of alternative methods for OSA screening by clinicians, as the clinical diagnosis of OSA is mainly based on the AHI. However, it is important to bear in mind that the oxygen saturation signal was not obtained with the custom-made device but with the PSG device, and data was collected obtrusively in controlled conditions. Therefore, some of the limitations of PSG are also present in the proposed approach.

Machine Learning Methods

Having looked at rule-based approaches to screen for OSA, we will now focus on studies that have implemented machine learning methods. Two studies developed a logistic regression classifier. Yadollahi and Moussavi (2009) used tracheal respiratory sounds and oxygen saturation for estimating the AHI. Later in this thesis we will also investigate the integration of acoustic features with physiological information to robustly screen for OSA. Respiratory sounds were recorded with a high-quality microphone placed on the suprasternal notch, and oxygen saturation was measured with a pulse oximeter. Data was collected from 40 subjects during PSG. This was scored by a registered polysomnographic technologist, and used as reference. Audio recording segments associated with a $\geq 4\%$ desaturation were extracted from the drop to the following rise in oxygen saturation. Based on energy, breaths and snores were first detected in the extracted segments. Then these segments were classified as either apnoea-hypopnea or non-apnoea-hypopnea

segments. Total sound energy of the breathing events, proportion of breathing events, proportion of snore events, and amount of desaturation in a segment were provided as input to the classifier. If the output of the logistic regression classifier was below 0.5, the segment was classified as an apnoea-hypopnea segment. Otherwise, it was classified as a non-apnoea-hypopnea segment. The AHI was estimated using the number of apnoea-hypopnea segments, and the duration of the audio recording, as shown in Equation 2.18.

As we saw in Chapter 1, the AHI is calculated as the average number of apnoea-hypopnea events per hour of sleep rather than hour of recording. Using the duration of the audio recordings is nonetheless a reasonable approximation, as estimating the time spent sleeping requires specialised hardware, an electroencephalograph, which Yadollahi and Moussavi (2009) were and we are trying to avoid. Similar to Saha et al. (2020), the performance was evaluated with a Bland-Altman plot: 93% of the estimated AHIs were within the agreement limits. A sensitivity of 89% with a specificity of 92% were reported when using the estimated AHIs to screen for OSA. That is, predicting whether a subject is healthy: AHI < 5 events/hour, or has OSA: AHI ≥ 5 events/hour. However, it is important to take into account that the audio recordings were performed obtrusively in controlled conditions, and additional hardware was required to measure oxygen saturation. This might limit the performance of the proposed screening system in typical sleep conditions, and hinder its wide use.

Kim et al. (2018) also used a logistic regression classifier. But they proposed a different approach to screen for OSA. Instead of detecting individual apnoea-hypopnea events to estimate the AHI, as was done in the previous study, Kim et al. (2018) predicted the OSA severity — normal, mild, moderate or severe — directly from features computed from whole-night audio recordings. They collected audio recordings from 120 participants — three times the number of subjects in the previous study — during PSG with a microphone placed on the ceiling at 1.7 m from the bed. PSG data was scored by a sleep physiologist, and used as reference. To predict the OSA severity, an 'acoustic biomarker' was developed. It is a set of 98 acoustic features that allows the differentiation between OSA severity groups. Such features include MFCCs, spectral centroid, zero-crossing rate, root mean square, linear predictive coding, first 3 formants, sub-band energy distribution, and weighted sound intensities, amongst others. These were extracted from 5-second non-overlapping windows, and their mean and standard deviation were computed for each feature as its representative values for the whole night. Only the audio recording segments corresponding to the second and third sleep stages, derived from the scored PSG data, were considered.

In addition to the acoustic biomarker, a 'quantised transition matrix' was proposed with the aim of exploiting the temporal pattern of breathing sounds. It quantises the absolute magnitude of the audio signal into four levels: (1) silence segments with a duration

of less than 20 seconds, (2) low-level segments (e.g., those with healthy breathing), (3) high-level segments (e.g., those with snoring), and (4) silence segments with a duration of more than 20 seconds (e.g., those with apnoea events). The quantised transition matrix represents the distribution of transitions between these four levels in a whole night. It was provided along with the acoustic biomarker as input to the logistic regression classifier. 10-fold cross-validation was performed. In each fold, 90% of the data was used for training and the remaining 10%, for testing. An average recall of 88% was reported. It is worth noting that number of participants for each OSA severity group was the same. A sensitivity of 97% with a specificity of 83% were obtained when screening for OSA. However, as the study by Yadollahi and Moussavi (2009), this study used sleep audio recordings made in controlled conditions. Additionally, the proposed approach requires specialised hardware and manual scoring of sleep stages, which is not practical. It poses the same limitations as PSG: obtrusiveness, and expensive and time-consuming scoring. Exploiting the temporal pattern of breathing sounds is nonetheless a powerful concept given the silent and long-duration nature of apnoea events. This idea will be further investigated in this thesis.

The studies considered so far have analysed breathing sounds during sleep to screen for OSA. Although reasonable performance has been achieved by integrating acoustic features with physiological parameters, an alternative approach might be to exploit the relationship between breathing sounds and speech to screen for OSA using the latter. Anatomic upper airway abnormalities have been associated with OSA (Partinen et al., 1988), and the anatomical properties of the vocal tract — like soft tissue characteristics and vocal tract structure — affect the acoustic parameters of speech. Based on this, Elisha et al. (2012) developed a GMM classifier to screen for OSA from speech. They hypothesised that there are differences in the acoustic parameters of speech between healthy subjects and those with OSA. To develop the screening system, audio recordings were collected from 87 participants, who read a 1-minute text. The text — in Hebrew — was designed to emphasise vowels: /a/, /e/, /i/, /o/ and /u/, and nasal phonemes: /m/ and /n/.

Elisha et al. (2012) extracted 103 acoustic features from 30-ms frames with 50% overlap. These were obtained from the time and frequency domains. 7 GMM classifiers were trained: one for each vowel, one for the nasal phonemes, and one for all other phonemes and silence. The parameters for each GMM classifier were learned from data for healthy subjects and those with OSA. Participants were classified as either having OSA or being healthy using a weighted sum of the outputs of the classifiers, and a threshold optimised on the training data. A sensitivity of 83% with a specificity of 81% were reported when using automatic phoneme segmentation, whereas a sensitivity and a specificity of 92% were achieved when using manual phoneme segmentation. One possible limitation (and advantage) of this study is that it indirectly screens for OSA. So, the subject does not need

to be asleep, but breathing sounds are not directly evaluated, which might not be easily accepted by clinicians.

Up to this point, we have looked at studies that have implemented traditional machine learning techniques to screen for OSA from breathing sounds and physiological parameters or indirectly from speech. We will now consider two studies that stand out for having used acoustic features on their own along with deep learning techniques to screen for OSA. Nakano et al. (2019) trained CNNs to predict from tracheal respiratory sounds the main diagnostic parameters for OSA: respiratory events (e.g., apnoeas and hypopneas) and sleep stages. Similar to the study by Yadollahi and Moussavi (2009), tracheal respiratory sounds were recorded during PSG with a microphone attached to the participant's throat over the trachea. Data from 1,852 participants was collected over 8 years, which, to the best of our knowledge, makes this study the largest ever in the field of acoustic analysis of sleep-disordered breathing.

PSG data was scored by registered polysomnographic technologists, and used as reference. Ten respiratory events were considered: central apnoea with termination, obstructive apnoea with termination, central hypopnea with termination, obstructive hypopnea with termination, apnoea without termination, hypopnea without termination, irregular breathing or body movement, normal breathing, snoring, and 'other' events. 'Termination' refers to whether the respiratory event ended within the 60-second analysis window. This differs from the studies previously considered, which collapsed hypopneas, and central and obstructive apnoeas into the apnoea class, and all other events, into the non-apnoea class, since their limited amount of data hindered a more detailed classification approach. Hypopneas, and central and obstructive apnoeas are nonetheless considered the same for calculating the AHI. Likewise, four sleep stages were considered: wake, non-REM 1, non-REM 2 or 3, and REM. Spectrograms were computed for 60-second segments or analysis windows with 50% overlap. These were labelled automatically using the scored PSG data, and provided as input to the classifiers.

Similar to Demir et al. (2018), Nakano et al. (2019) approached a sound classification problem as an image classification task. Two different CNNs were trained to classify respiratory events, and sleep stages from spectrogram images. The input to both networks was a 64×300×3 tensor corresponding to 64 frequency bins (between 22 and 700 Hz), 300 frames (i.e., a 60-second segment with a frame rate of 0.2 seconds), and 3 colour channels (i.e., red, green, and blue). The networks consisted of 3 convolutional layers with ReLU activation, each one followed by a max-pooling layer, and 2 fully connected layers with tanh and softmax activations, respectively. The number of artificial neurons in the output layer was 10 for the CNN classifying respiratory events, and 4 for the CNN classifying sleep stages. That is, one artificial neuron for each class considered. Sleep status (i.e., sleep vs. wake classification) was predicted with a sensitivity of 92% and a specificity of

72%. The output of the networks was used for directly calculating the AHI following the standard definition (Equation 1.1).

In the same way as Yadollahi and Moussavi (2009), the AHI estimation performance was evaluated with a Bland-Altman plot: 94% of the estimated AHIs were within the agreement limits. A sensitivity of 98% with a specificity of 76% were reported when using the estimated AHIs to screen for OSA. The study by Nakano et al. (2019) has three key strengths. Firstly, using a large amount of data to develop the proposed system, which allowed the estimation of an important physiological parameter: sleep stage, and a detailed classification approach, unlike previous studies. Secondly, directly calculating the AHI, the main parameter for OSA diagnosis. This might advance the adoption of alternative diagnostic methods amongst clinicians. Thirdly, relying only on acoustic evidence, which would potentially facilitate the wide use of the proposed approach. However, it is worth noting that the sleep audio recordings used were collected obtrusively in controlled conditions at the same facility. So, the developed method shares one of the limitations of PSG: obtrusiveness, and its performance in a bedroom at home, where these alternative methods are intended to be used, is still to be examined.

So far, all the studies presented in this chapter have used sounds that the researchers could hear for classifying sleep-disordered breathing events, and screening for OSA. We will now look at a study that used sounds beyond the range of human hearing. Tiron et al. (2020) developed an approach that uses passive acoustic analysis, and sonar reflections for estimating the AHI, and screening for OSA. To develop such an approach, they collected breathing sounds and ultrasonic reflections with an Android or iOS smartphone from 248 participants during PSG. The smartphone was placed on a bedside table with the microphone and speaker facing the participant, and the audio recordings were sampled at 48 kHz to be able to capture the ultrasonic reflections. PSG data was scored by a registered polysomnographic technologist, and used as reference.

For passively detecting breathing sounds, the audio recordings were bandpass-filtered in the range of 250–8,000 Hz, which removed the active sonar signal. Then, similar to Duckitt et al. (2016), MFCCs were computed for non-overlapping 25-ms frames, and provided as input to a CNN classifier. This consisted of convolutional layers with ReLU activation and dense layers, and was trained to detect body movement (e.g., duvet noise), breathing, snoring, breathing pauses, and recovery breaths. On the other hand, the active sonar signal was a swept sinusoidal waveform in the range of 18–22 kHz played on the smartphone speaker. The reflections of that signal were recorded with the smartphone microphone, and used for deriving respiratory effort and sleep stages based on body movement.

Time-frequency features from both the passive audio and active sonar reflections were provided as input to different classifiers. For classifying sleep stages, gender and age were

additionally used as features to consider age-related variations in sleep architecture. The AHI was estimated using a linear regression model. Time spent sleeping, derived from the predicted sleep stages, was provided as an additional feature to the model. Finally, screening for OSA was carried out with a logistic regression classifier, which classified the participants as either having an AHI < 15 events/hour or having an AHI ≥ 15 events/hour (i.e., clinically relevant OSA). Unfortunately, a more detailed description of the features and classifiers implemented is not possible, as proprietary algorithms were used by Tiron et al.

As the studies by Yadollahi and Moussavi (2009), and Nakano et al. (2019), the performance on the AHI estimation task was evaluated with a Bland-Altman plot: 90% of the estimated AHIs were within the agreement limits. A sensitivity of 88% with a specificity of 80% were reported for the clinical threshold of AHI ≥ 15 events/hour. Exploiting active sonar reflections to derive key physiological parameters to screen for OSA using readily available hardware is the main strength of the proposed approach, as it overcomes one of the limitations of the acoustic analysis of sleep-disordered breathing: the lack of physiological data. Also, from all the studies considered in the present chapter, this is the only one that was deployed on mobile devices, which might improve accessibility to sleep-disordered breathing diagnosis, as no additional hardware is required, and preserves user privacy, since the data is processed locally on the smartphone. However, the study by Tiron et al. (2020) has the same limitation as the studies considered before: the audio recordings were collected in a sleep laboratory. So, once more, how their proposed approach performs in a different acoustic environment has yet to be evaluated.

Lastly, Table 2.2 summarises the studies presented in this section. As before, performance comparison between studies is not entirely appropriate, since a custom dataset was employed in each study.

Table 2.2 Using acoustics to screen for OSA

Study	**Yadollahi et al. (2009)**	**Elisha et al. (2012)**	**Almazaydeh et al. (2013)**	**Al-Mardini et al. (2014)**	**Alakuijala et al. (2016)**	**Narayan et al. (2018)**	**Kim et al. (2018)**	**Nakano et al. (2019)**	**Castillo et al. (2019)**	**Saha et al. (2020)**	**Tiron et al. (2020)**
Task	AHI	OSA screening	Apnoea events	OSA screening	OSA screening	OSA screening	OSA severity	Apnoea and sleep status (AHI)	AHI	AHI	OSA screening and AHI
Audio recordings	Tracheal sounds at clinic	Speech	Breathing sounds	Tracheal sounds with a smartphone	Breathing sounds	Ambient sound at clinic	Ambient sound at clinic	Tracheal sounds at clinic	Ambient sound at home	Tracheal sounds at clinic	Ambient sound at clinic
Data	40 subjects	87 subjects	50 subjects	15 subjects	211 subjects	91 subjects	120 subjects	1,852 subjects	13 subjects	69 subjects	248 subjects
Reference	PSG	PSG	Subjects were awake	PSG	HSAT	PSG	PSG	PSG	HSAT	PSG	PSG
Features	Energy and oxygen saturation	Time and frequency features	VAD	Energy, oxygen saturation and body movement	Snoring	Energy	Acoustic biomarker and sleep stage	Spectrogram images	Sample entropy and oxygen saturation	Energy, oxygen saturation and body movement	MFCCs and active sonar reflections
Classifier	Sigmoid functions sum	GMM	Rule based	Rule based	Rule based	Rule based	Logistic regression	CNN	Rule based	Rule based	CNN and logistic regression
Results	Sensitivity: 89% Specificity: 92%	Sensitivity: 83% Specificity: 81%	Accuracy: 97%	Sensitivity: 100% Specificity: 86%	Sensitivity: 93% Specificity: 35%	Sensitivity: 94% Specificity: 63%	Sensitivity: 97% Specificity: 83%	Sensitivity: 98% Specificity: 76%	Sensitivity: 100% Specificity: 100%	Sensitivity: 91% Specificity: 89%	Sensitivity: 88% Specificity: 80%

2.4 Summary

This chapter introduced the sound analysis problem in the context of sleep-disordered breathing, and set out the evaluation framework that will be used throughout this thesis to assess the performance of the systems that will be presented. Generative and discriminative machine learning methods that can be employed for such an analysis were introduced as well. Special attention was paid to deep learning methods. The current chapter also reviewed previous attempts to analyse sleep-disordered breathing events. These have focused on detecting snoring, and screening for OSA from sleep audio recordings. Considering that the acoustic analysis of sleep-disordered breathing is a data-scarce field, conventional machine learning techniques have been usually implemented because of their robustness to limited amounts of (labelled) data. This brought about research question RQ4, which will be addressed in Chapters 4 and 5. Deep learning methods, which require considerably more training data than traditional machine learning techniques, have been nonetheless implemented for the same tasks, since they can model complex relationships between the input features and the expected output.

Many studies have proposed rule-based methods rather than machine learning techniques to screen for OSA. The latter might be nonetheless more robust than the former, as handcrafted rules are unlikely to effectively capture the great variability of breathing sounds, and deal with different room acoustics and levels of background noise. Some approaches have integrated acoustic features with physiological information such as oxygen saturation, sleep stages, and body movement. A few studies have detected individual apnoea-hypopnea events to then estimate the AHI, since OSA diagnosis is clinically based on this index. These have used the duration of the audio recording instead of the time spent sleeping for its estimation, as sleep stage information is required to compute the time spent sleeping. Other approaches, instead of detecting individual apnoea-hypopnea events, have predicted the OSA severity directly from features computed from whole-night audio recordings. Lastly, one of the considered studies exploited the relationship between breathing sounds and speech to screen for OSA using a read text.

Two studies stood out for having made use of acoustic features on their own, and deep learning techniques for OSA screening. These derived physiological parameters from acoustic evidence. Nakano et al. (2019) employed CNNs to predict respiratory events and sleep stages from tracheal respiratory sounds. Tiron et al. (2020) used passive acoustic analysis to detect breathing events with a CNN, and active sonar reflections to derive respiratory effort and sleep stages. This study is the only one that considered co-sleeping: the breathing sounds of two participants sleeping on the same bed can be potentially distinguished based on active sonar reflections. However, it was only simulated with breathing phantoms. This led to research question RQ5, and will be considered in Chapter 6.

The reviewed studies recognised that there is no widely accepted definition of snoring, which motivated research question RQ2, and will be addressed in Chapter 3. Although reasonably good performance was reported by the studies considered, they exhibit some limitations. Most studies collected sleep audio recordings in the same controlled conditions — a sleep laboratory. This might limit the performance of the developed systems in realistic sleep conditions — a bedroom at home — where different room acoustic characteristics and background noise conditions, for example, can affect their performance. Some studies collected sleep audio recordings when the participants were under sedation. In this case, the data might not epitomise natural sleep physiology, and the performance of the proposed approaches on natural sleep remains to be investigated. Another common limitation is the use of specialised hardware, for instance, ceiling-mounted microphones. It would potentially hinder the wide use of the developed methods, as they would be conditioned by equipment availability. Additionally, some studies carried out sleep audio recordings obtrusively, for instance, with a microphone attached to the participant's throat or a smartphone attached to their chest. Therefore, the proposed approaches share some limitations with PSG such as obtrusiveness and requiring specialised hardware. Ideally, systems to screen for sleep-disordered breathing should be robust enough to be applied in typical sleep conditions, and make use of readily available hardware. This prompted research questions RQ1 and RQ3, which will be tackled in the upcoming chapter, and Chapters 4 and 5, respectively.

Chapter 3

Sleep-disordered Breathing Sound Corpora

"Laugh and the world laughs with you,
snore and you sleep alone."
– A. Burgess

Healthcare research is a data-scarce field (Meijerink et al., 2020). Most of the studies on the acoustic analysis of sleep-disordered breathing collect their own data, and it is seldom made available to others for research. One of the reasons is that audio recordings are not usually made during HSAT or PSG, the gold standard for sleep-disordered breathing diagnosis. A second reason is the clinical nature of the data, and the associated medical privacy concerns. Another reason is data scarcity itself: the collected data is a valuable resource and may have commercial value.

The present chapter will describe the process of data collection, and the development of two sleep-disordered breathing corpora: the *Snoring Sound Corpus* and the *OSA Sound Corpus*. The desirable characteristics for an acoustic corpus of sleep-disordered breathing are summarised in Table 3.1. Data collection was done in a bedroom at home with a smartphone, since improving access to diagnosis is one of the main challenges that sleep medicine is still facing (Dafna et al., 2013; Phillips, 2007). Creating the Snoring Sound Corpus required manual annotation of sleep audio recordings. For this purpose, an annotation scheme was defined, and inter-annotator agreement was evaluated. Data for the OSA Sound Corpus was collected during HSAT. HSAT data was manually scored by a Registered Polysomnographic Technologist, and used as reference, which required the synchronisation of sleep audio recordings with the scored HSAT data.

Before examining strategies for analysing breathing sounds, it is important to understand their characteristics to inform decisions for their analysis. The present chapter will also characterise breathing and the main forms of sleep-disordered breathing —

Table 3.1 Desirable characteristics for an acoustic corpus of sleep-disordered breathing

Characteristic	**Motivation**
Collected in a variety of realistic sleep environments	To facilitate scaling up the technology developed in the general population
Collected with readily available hardware	To improve access to sleep-disordered breathing diagnosis
Annotated with a consistent and detailed annotation scheme	Sleep-disordered breathing is a range of conditions that result from the collapse of the upper airway

snoring and sleep apnoea — from an acoustic perspective using the corpora that were recorded. Since there is no widely accepted definition of snoring, we will propose one from its acoustic characterisation. Some of the work presented in this chapter has previously appeared in (Romero et al., 2019) and (Romero et al., 2020a).

3.1 Snoring Sound Corpus

Sleep audio recordings were collected from 31 male and 13 female participants recruited from the general population without particular criteria. Each participant was recorded for five nights with an iOS device (e.g., Apple iPhone or iPod touch) using a bespoke app. Recordings were made in a bedroom at home, and the device was placed at head level within arm's reach (e.g., on a bedside table). A single audio channel was used with a sample rate of 16 kHz, and a resolution of 16 bits per sample. Passion for Life Healthcare (UK) Limited, one of the sponsors of this study, recruited the participants and managed the recordings. Data collection and storage protocols were subjected to the ethical review procedures of the University of Sheffield.

3.1.1 Annotation of Sleep Audio Recordings

A subset of the collected sleep audio recordings was chosen to be manually annotated using Praat (Figure 3.1). This consisted of data from six male participants. Age, height and weight were collected as part of the demographic information. The participants — aged between 18 and 64 years — had an average body mass index (BMI, calculated as weight / height^2) of 26.63 ± 3.67 kg/m^2, which evinces that most of them did not have a healthy weight, as 25 ≤ BMI < 30 kg/m^2 denotes overweight (Nuttall, 2015), an independent risk factor for habitual snoring (Ma et al., 2017; Ursavas et al., 2008). 25 2-minute segments with at least 20% of snoring (i.e., snore events amounting to 24 seconds) were selected for each participant. These were identified using a simple GMM that was trained

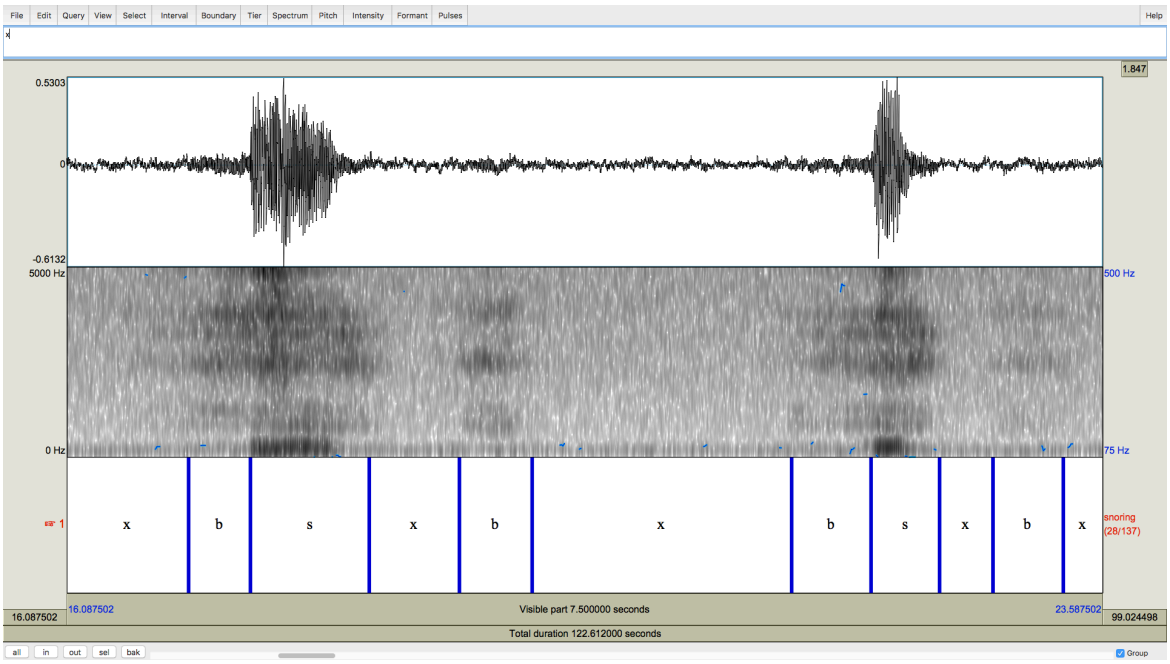


Fig. 3.1 Example of manual annotation in Praat. The upper panel displays the waveform; the middle panel, the spectrogram; and the lower panel, the manual annotation. Snore (s), breath (b), and silence (x) events are marked in the 'snoring' tier.

after manually selecting and annotating a few recordings. Additionally, 27 2-minute segments containing mainly other sound events (e.g., speech, a car passing by, etc.) were also included to balance the amount of data available for each of the classes considered. Therefore, the Snoring Sound Corpus consists of 354 minutes of manually annotated sleep audio recordings.

As discussed in Chapter 2, most studies commonly define two classes of sound events for annotation (and classification): (1) snore or breath, and (2) non-snore or non-breath (Dafna et al., 2013; Emoto et al., 2018; Nonaka et al., 2016). However, considering additional audio events that are normally present in sleep audio recordings might be useful. For instance, given the great variability of breathing sounds, a more detailed annotation can potentially result in more robust classifiers. For this reason, the annotation scheme presented in Table 3.2 was defined. This considers 6 sound events: 'snore' (s), 'breath' (b), 'silence' (x), 'wheezing' (w), 'noisy in-breath' (n), and 'other' (o). Even though there is a commonly understood perception of snoring, there is no widely accepted definition of it. Therefore, given this commonly understood perception, annotators subjectively marked snore regions in the audio recordings of the Snoring Sound Corpus. It is worth mentioning that the proposed annotation scheme allows events to be merged into a single class if needed. For example, it is possible to merge 'breath', 'silence', 'wheezing', 'noisy in-breath' and 'other' into 'other' (i.e., non-snore) to obtain an annotation scheme similar to that

Table 3.2 Annotation scheme

Event	Label	Description
Snore	s	Pitched breathing sound
Breath	b	Unpitched breathing sound, high frequency noise
Silence	x	Hiss, low background noise, no other structure
Wheezing	w	Whistling sound during breathing
Noisy in-breath	n	Similar to a snore but without a strong pitch
Other	o	Speech, street noise, alarm clock, aircraft noise, birds chirping, etc.

used by other studies. In Figure 3.2 the corpus statistics are shown. After collapsing 'wheezing' and 'noisy in-breath' into the 'snore' class and removing 70% of 'silence', 35.5% of the Snoring Sound Corpus consists of data from the 'snore' class, and the remaining 64.5% consists of about the same amount of data from each of the other classes.

With the aim of reducing the time required to perform an annotation — around 30 minutes for each 2-minute segment — an HMM was developed to automatically segment and annotate a sleep audio recording. This was done using HTK (Young et al., 2006). The HMM was trained with the first 15 annotations performed manually, and only considered the 'snore', 'breath', 'silence', and 'other' classes. The 'wheezing' and 'noisy in-breath' classes were merged into the 'snore' class, since a very small amount of data was available for them. The HMM consisted of seven states for the 'snore' class, five states for the 'breath' class, and three states for the 'silence' and 'other' classes. The number of states was heuristically selected according to the complexity of each sound event, and their output distribution was modelled by a mixture of Gaussians. With this model the time required to perform an annotation was approximately reduced by 50%, since the task then consisted of manually adjusting the segmentation, and revising the labels.

3.1.2 Inter-annotator Agreement Evaluation

Most studies evaluate inter-annotator agreement to ensure that reliably and consistently annotated data is used. Usually, three observers annotate a subset of audio recordings, and their agreement is evaluated using well-defined metrics like Cohen's kappa (Dafna et al., 2013; Emoto et al., 2018; Nonaka et al., 2016). Cohen's kappa is a widely accepted agreement metric that considers the possibility of agreement by chance. It can take values between -1 and 1. Values below 0 indicate that agreement is worse than guessing, and 1 indicates perfect agreement. Cohen's kappa (K) is defined as follows:

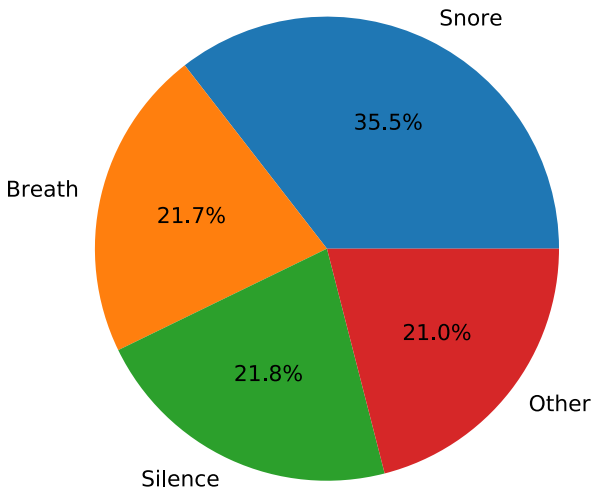


Fig. 3.2 Class distribution of the Snoring Sound Corpus after removing 70% of 'silence', and collapsing 'wheezing' and 'noisy in-breath' into the 'snore' class

$$K = \frac{P_o - P_e}{1 - P_e} \tag{3.1}$$

where P_o is the proportion of segments in which the annotators agreed, and P_e is the proportion of segments for which agreement is expected by chance. Cohen's kappa can be interpreted as "the proportion of agreement after chance agreement is removed from consideration" (Cohen, 1960, p. 40).

Segmentation and annotation of two 2-minute sleep audio recordings was performed by three observers: HER, GJB and NM, with the aim of evaluating inter-annotator agreement before carrying out the annotation of all the collected data. Cohen's kappa was calculated at frame level for pairs of annotators, and the obtained results are presented in Table 3.3. All kappas were between 0.43 and 0.85, which suggested a reasonable level of agreement, and reflected the great variability of breathing sounds during sleep. Since the initial inter-annotator agreement was considered good enough, each sleep audio recording in the corpus was annotated once by one observer. In related studies the reported Cohen's kappa is strikingly high. For example, Emoto et al. (2018) reported $K \geq 0.98$, and Dafna et al. (2013), $K \geq 0.97$. A better agreement was achieved in these studies in comparison with ours probably because a basic annotation scheme (i.e., snore and non-snore) was used for labelling data that had been previously segmented, whereas in the present study the annotators both segmented and annotated the sleep audio recordings. We will now consider the OSA Sound Corpus.

Table 3.3 Inter-annotator Cohen's kappa

Annotator pair	Audio recording 1	Audio recording 2
HER – GJB	0.85	0.70
HER – NM	0.44	0.54
GJB – NM	0.43	0.68

3.2 OSA Sound Corpus

To develop the system to screen for OSA that will be introduced in Chapter 5, audio recordings were collected from 100 participants during home sleep apnoea testing (HSAT). Participants slept on their own in a bedroom at home. HSAT was performed with a SOMNOmedics SOMNOtouch RESP,[1] which records respiratory effort, nasal airflow, oxygen saturation, heart rate, snoring, and body position with sensors attached to the body. The measurement and use of these physiological parameters are presented in Table 3.4. Audio recordings were made with an Android or iOS smartphone placed at head level within arm's reach using a custom app. Data for one or two nights was collected. As with the Snoring Sound Corpus, Passion for Life Healthcare (UK) Limited recruited the participants and managed the recordings, and data collection and storage protocols were subjected to the ethical review procedures of the University of Sheffield.

The first 20 and the last 2 minutes of the audio recordings for each night were not included, since the participants were usually awake during these periods, and the breathing sounds could not be properly recorded due to the presence of speech and noise (e.g., an alarm). HSAT data was scored by a Registered Polysomnographic Technologist, who manually annotated obstructive apnoeas, central apnoeas, and hypopneas based on airflow and respiratory effort information. Events with a > 90% reduction in airflow for more than 10 seconds associated with respiratory effort were marked as obstructive apnoeas, whereas those not associated with respiratory effort were annotated as central apnoeas. Events with a > 30% reduction in airflow for more than 10 seconds were labeled as hypopneas.

Data was not included in the OSA Sound Corpus if it was not long enough, since at least 3 hours of data are required to obtain clinically significant results. Therefore, data was not scored if it had a duration of < 3 hours. Furthermore, data was discarded if the corresponding audio recordings were missing or sensor data was corrupted. For instance, sensors sometimes fall off as people move during sleep, and signals are then not properly recorded. The OSA Sound Corpus consists of data for 60 nights from 45 participants, which amount to over 416 hours of data.

[1] www.somnomedics.de/en/solutions/sleep_diagnostics/polygraphy-devices/somnotouch-resp/

Table 3.4 Physiological parameters measured during HSAT

Physiological parameter	Measurement	Use
Respiratory effort	Thoracic and abdominal belts	Differentiation between obstructive and central apnoeas
Nasal airflow	Nasal cannula	Detection of apnoeas and hypopneas
Oxygen saturation	Pulse oximeter	Detection of apnoeas and hypopneas
Heart rate	Pulse oximeter	Computation of heart rate variability
Snoring (low-sampled sound)	Derived from nasal airflow	Determining snoring occurrence
Body position	Accelerometer	Assessing positional occurrence of snoring or OSA

Table 3.5 along with Figures 3.3 and 3.4 present the demographic information (SD: standard deviation). As noted in Table 3.5, 62.2% of the sleep audio recordings were made on iOS devices, whereas 37.8% of them, on Android devices. In this way, the corpus consists of recordings made with different microphones, which might contribute to the development of a robust screening system, as it would not be tailored to a particular microphone response. The corpus is made up of more data from male than female participants. It can also be noted that the average BMI of the participants in the corpus is 30.17 kg/m^2, which evidences that most of them are overweight or obese, since $25 \leq BMI < 30$ kg/m^2 indicates overweight and $BMI \geq 30$ kg/m^2, obesity (Nuttall, 2015). Overweight and obesity are risk factors for OSA (Young et al., 2004). In fact, the two nights with the highest AHIs are from the participant with the highest BMI, as can be observed in Figure 3.3. However, it is worth noting that the BMI and AHI do not correlate linearly. It can also be seen that the corpus consists of a broad range of AHI values. Considering Figure 3.4, about three quarters of the nights in the corpus correspond to subjects with OSA: those with an AHI ≥ 5 events/hour. The amount of nights corresponding to severe OSA (i.e., AHI ≥ 30 events/hour) is nonetheless the lowest of the four severity groups followed by that corresponding to moderate OSA (i.e., $15 \leq$ AHI < 30 events/hour).

Considering that sleep stages alter respiration, intra-night variability is expected. As discussed in Chapter 1, apnoeas and hypopneas are more likely to occur during REM sleep due to the absence of muscle tone (Chokroverty and Avidan, 2016; Colten and Altevogt, 2006a), for example. Inter-night variability can be expected as well, since several factors such as alcohol consumption before going to bed, nasal congestion, and sleep position can determine if someone snores at all. Lastly, variability across participants is also expected because there are distinct craniofacial phenotypes (Yaggi and Strohl, 2010; Young et al.,

Table 3.5 Demographics of the OSA Sound Corpus

	Count/Range	**Mean ± SD**
Male (participants)	26	
Female (participants)	19	
Age (years)	25–71	41.58 ± 13.91
BMI (kg/m^2)	19.40–45.01	30.17 ± 6.44
Recording duration (hours/night)	3.33–9.93	7.27 ± 1.33
AHI (events/hour)	1.6–113.0	18.85 ± 25.11
iOS (devices)	28	
Android (devices)	17	

2004), snoring can be generated by various structures in the upper airway — as we will see later in this chapter — and OSA has different degrees of severity. However, given that the required information (e.g., sleep stages, craniofacial phenotype, etc.) was not available, these forms of variability in the collected data were not investigated.

3.2.1 Synchronisation Between Audio Recordings and HSAT

Although the audio recordings and the HSAT data were timestamped, their clocks were not synchronised. Since the timestamps could not be used by themselves to synchronise the data, an algorithm for automatically synchronising the HSAT data with the corresponding audio recordings was developed. It used the first 40 minutes of the audio recordings, and the snore signal from HSAT as follows:

1. The audio signal, sampled at 16 kHz, was downsampled to the HSAT sampling frequency of 256 Hz.
2. An approximate initial time difference between the snore and audio signals was calculated from the timestamps.
3. The approximate initial time difference was used to extract the segment from the snore signal that approximately corresponded to the audio segment.
4. The snore and audio segments were half-wave rectified and normalised to the 0–1 range.

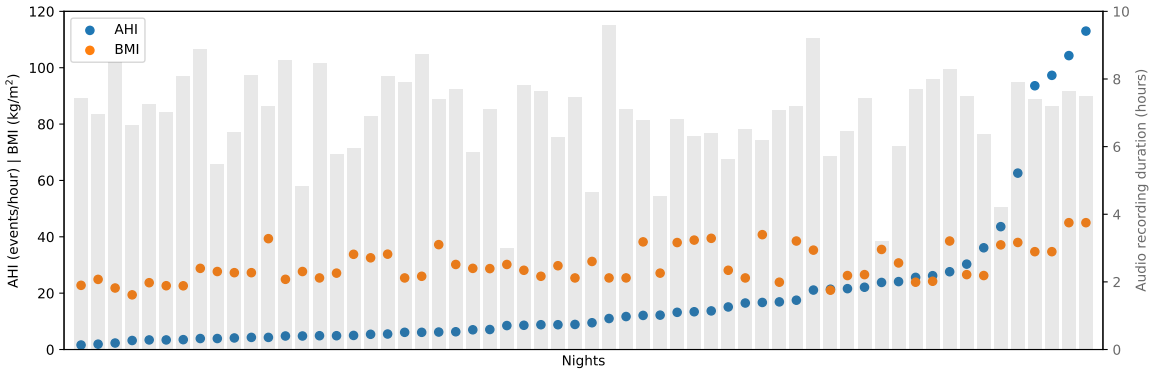


Fig. 3.3 Audio recording duration (bars, right axis), AHI (blue dots, left axis) and BMI (orange dots, left axis) for each night in the OSA Sound Corpus. Nights sorted by AHI.

5. The crosscorrelation function (C) of the snore (s) and audio (a) segments was computed as:

$$C(\tau) = \sum_{n=0}^{N-1} s(n) \cdot a(n - \tau)$$

where N was the number of samples in the segment, and $\tau \in [0, N)$ was the time delay. For a 40-minute segment sampled at 256 Hz, $N = 614,400$.

6. The initial time difference was adjusted using the time delay indicated by the peak in the crosscorrelation function.

7. The adjusted time difference was added to the timestamps of the HSAT data.

The audio recordings and the snore signal desynchronised towards the end of the night in a limited number of cases. This was likely caused by the HSAT clock not having the exact frequency reported by the manufacturer: 256 Hz. Decreasing the sampling frequency to a slightly lower value — for example, 255.997 Hz — to downsample the audio signal was found to improve the synchronisation. Lastly, snoring, apnoea-hypopnea events, inhalations, exhalations, and desaturations were annotated in the audio recordings using the scored HSAT data. An example of this annotation is provided in Figure 3.5. It shows a 30-second audio recording segment containing snores and an apnoea event. These are annotated along with the inhalations, exhalations, and desaturations. Consistent with what will be discussed in the following section, snoring occurs during inhalation in this particular example.

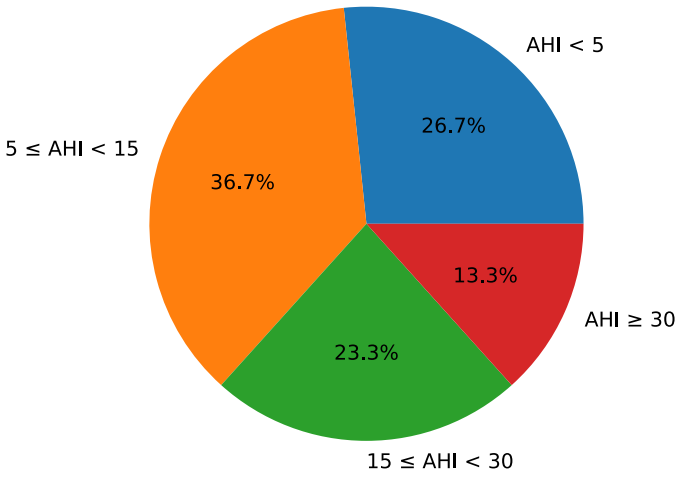


Fig. 3.4 AHI (events/hour) distribution of the OSA Sound Corpus. AHI < 5: normal, 5 ≤ AHI < 15: mild OSA, 15 ≤ AHI < 30: moderate OSA, and AHI ≥ 30: severe OSA. The percentage of nights in each interval is shown on the pie chart.

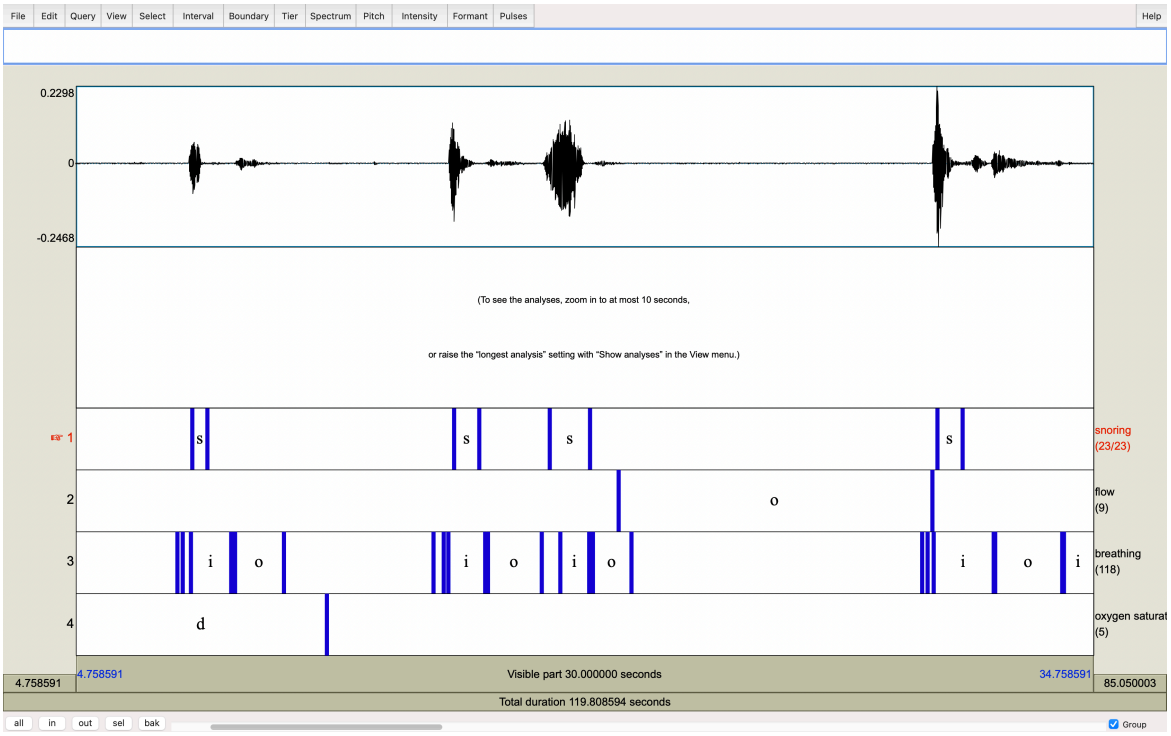


Fig. 3.5 Annotation of a sleep audio recording using the scored HSAT data. Snore events (s) are labeled in the 'snoring' tier. An obstructive apnoea (o) is marked in the 'flow' tier. Inhalations (i) and exhalations (o) are annotated in the 'breathing' tier. An oxygen desaturation (d) is marked in the 'oxygen saturation' tier.

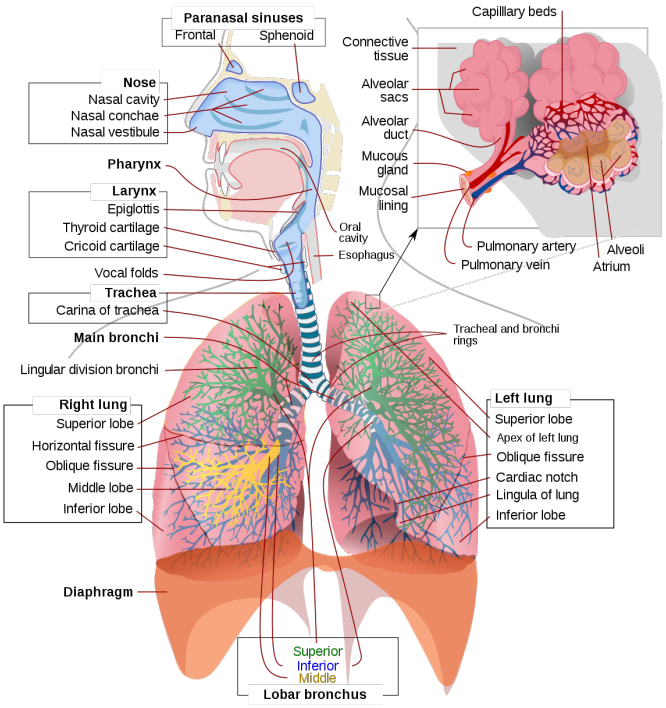


Fig. 3.6 Respiratory system (Ruiz Villareal, 2007)

3.3 Acoustic Characterisation of Sleep-disordered Breathing

Before characterising sleep-disordered breathing from an acoustic perspective, we will first consider the physiology of respiration. Human body cells require a constant supply of oxygen for the chemical reactions that generate energy from food, and the carbon dioxide generated as a by-product in these reactions must be continuously removed from the body to avoid unphysiological variations in pH.[2] Respiration moves oxygen from the atmosphere to the tissues, and carbon dioxide from the tissues to the atmosphere. Specifically, the respiratory system exchanges oxygen and carbon dioxide between the atmosphere and blood, while the blood carries oxygen and carbon dioxide between the respiratory system and the tissues (Sherwood, 2016).

Air flows into and out of the lungs through the pharynx, larynx, trachea and bronchi (Figure 3.6), thanks to periodic changes in alveolar[3] pressure by alteration of the volume of the lungs. According to Boyle's law:

$$P_0 V_0 = P_1 V_1 \tag{3.2}$$

[2] A measurement of the acidity or alkalinity of a solution (e.g., blood).
[3] Alveoli are tiny air sacs inside the lungs in which the oxygen and carbon dioxide exchange takes place.

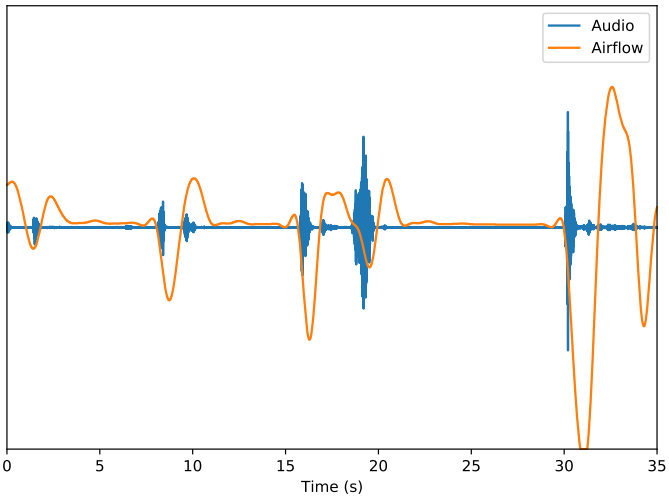


Fig. 3.7 Audio waveform and airflow for a 35-second segment from the OSA Sound Corpus

the pressure (P) exerted by a gas decreases as its volume (V) increases and vice versa. Since air flows from high pressure to low pressure, the alveolar pressure has to be lower than the atmospheric pressure during inspiration or breathing in, and has to be higher than atmospheric pressure during expiration or breathing out. The respiratory muscles change the volume of the lungs by altering the volume of the thoracic cavity. During inspiration or inhalation the external intercostal muscles and diaphragm contract to enlarge the thoracic cavity, alveolar pressure falls, and air flows into the lungs until alveolar pressure matches atmospheric pressure. Accessory inspiratory muscles in the neck, the sternocleidomastoid and scalenus, contract during deep inspiration to enlarge the thoracic cavity even further by elevating the sternum and the first two ribs. After inspiration, expiration or exhalation takes place. The diaphragm and external intercostals relax, the lungs recoil, alveolar pressure increases, and air flows out of the lungs until alveolar pressure equals atmospheric pressure (Sherwood, 2016).

Variations in air pressure that reach our ears is what we perceive as sound. The characteristics of sounds produced in the vocal tract — speech, singing, and breathing sounds — are determined, among other factors, by the way air flows through it. The present study leverages this fact. For instance, breathing sounds are louder when there is an obstruction in the vocal tract as we will see later in this section. We will now characterise inhalation and exhalation, snoring, and OSA from an acoustic perspective.

3.3.1 Inhalation and Exhalation

The distinction between inhalation and exhalation is not usually considered in most studies related to the acoustic analysis of sleep-disordered breathing. However, their

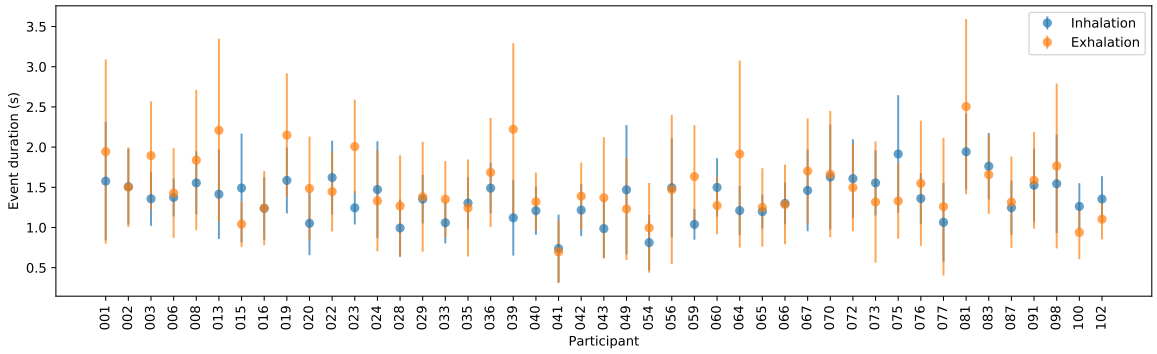


Fig. 3.8 Mean duration of inhalations and exhalations for each participant in the OSA Sound Corpus. Standard deviations are shown as error bars.

characteristics may provide salient information to screen for sleep-disordered breathing. For example, distinguishing inhalation and exhalation might potentially facilitate the distinction between snoring and other forms of sleep-disordered breathing such as catathrenia (nocturnal groaning). As will be discussed later on, snoring is a loud breathing event that mainly occurs during inhalation, whereas catathrenia, which is also a loud breathing event, mostly happens during exhalation and has a longer duration than snoring (Iriarte et al., 2015; Yu et al., 2020). Inhalations and exhalations from the OSA Sound Corpus were analysed. These were characterised in terms of temporal envelope, duration and spectral shape. Their correlation with airflow and snoring was also investigated.

Temporal Envelope and Airflow

Figure 3.7 presents the waveform for a typical 35-second segment from a sleep audio recording, and the corresponding airflow. Inhalations are seen in the flow signal as values below the RMS and exhalations, as values above the RMS. No correlation was found between sound intensity and airflow, since partial upper airway obstructions or snoring generally cause loud events and decreased flow with respect to healthy breathing, which is not always audible. For instance, the snore event between 18–20 seconds was associated with a small airflow, whereas the snore event between 30-32 seconds was associated with the largest airflow in the segment, and the exhalation event between 32–34 seconds was barely audible.

Duration

Figure 3.8 presents the mean duration of inhalations and exhalations along with the standard deviation for each participant in the OSA Sound Corpus. In general, a participant's exhalations had a longer mean duration than inhalations, and exhalations varied more in their duration. For a limited number of participants, inhalations were longer than

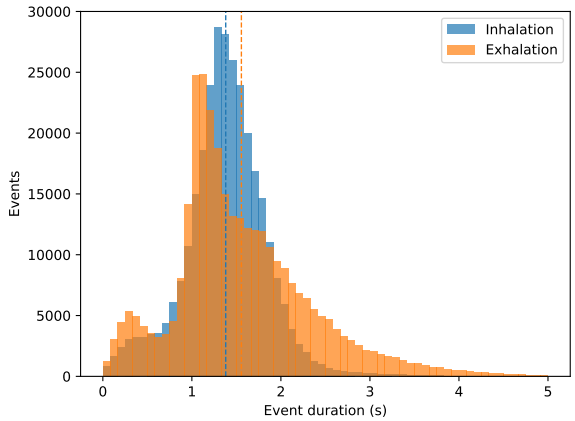


Fig. 3.9 Histogram of the overall duration of inhalations and exhalations of the participants with non-severe OSA in the OSA Sound Corpus. Means are shown as dashed lines.

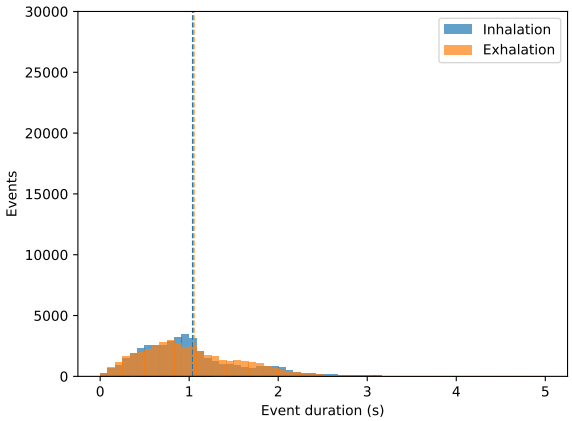


Fig. 3.10 Histogram of the overall duration of inhalations and exhalations of the participants with severe OSA in the OSA Sound Corpus. Means are shown as dashed lines.

exhalations or both events had about the same mean duration. For instance, the participant 041 has severe OSA and shallow breathing was observed: the mean event duration was below 1 second, whereas the participant 081 has mild OSA and their breathing events took longer than those of the participant 041.

Figures 3.9 and 3.10 show the histograms for the overall duration of inhalations and exhalations of the participants with non-severe and severe OSA in the corpus, respectively. Consistent with the participant-based statistics, inhalations of the participants with non-severe OSA had a mean duration of 1.38 $\pm$ 0.47 seconds, whereas those of the participants with severe OSA had a mean duration of 1.04 $\pm$ 0.62 seconds. Exhalations of the subjects with non-severe OSA had a mean duration of 1.56 $\pm$ 0.78 seconds, whereas those of the subjects with severe OSA had a mean duration of 1.05 $\pm$ 0.63 seconds. As expected, shorter durations were reported for the participants with severe OSA, since shallow breathing is characteristic of sleep-disordered breathing.

Spectral Shape

In the OSA Sound Corpus, 79% of snoring happened during inhalation, whereas 21% occurred during exhalation. Similar observations have been made by related studies. For instance, Levartovsky et al. (2016) found that 97% of snoring occurred during inhalation after analysing data from participants referred to PSG. Figure 3.11 shows the waveform for a 40-second segment from a sleep audio recording and the corresponding spectrograms. Typical inhalations are seen around 27–28 and 31–32 seconds. Typical exhalations can be observed around 28.5–29 and 33–33.5 seconds. Sound energy was usually throughout the

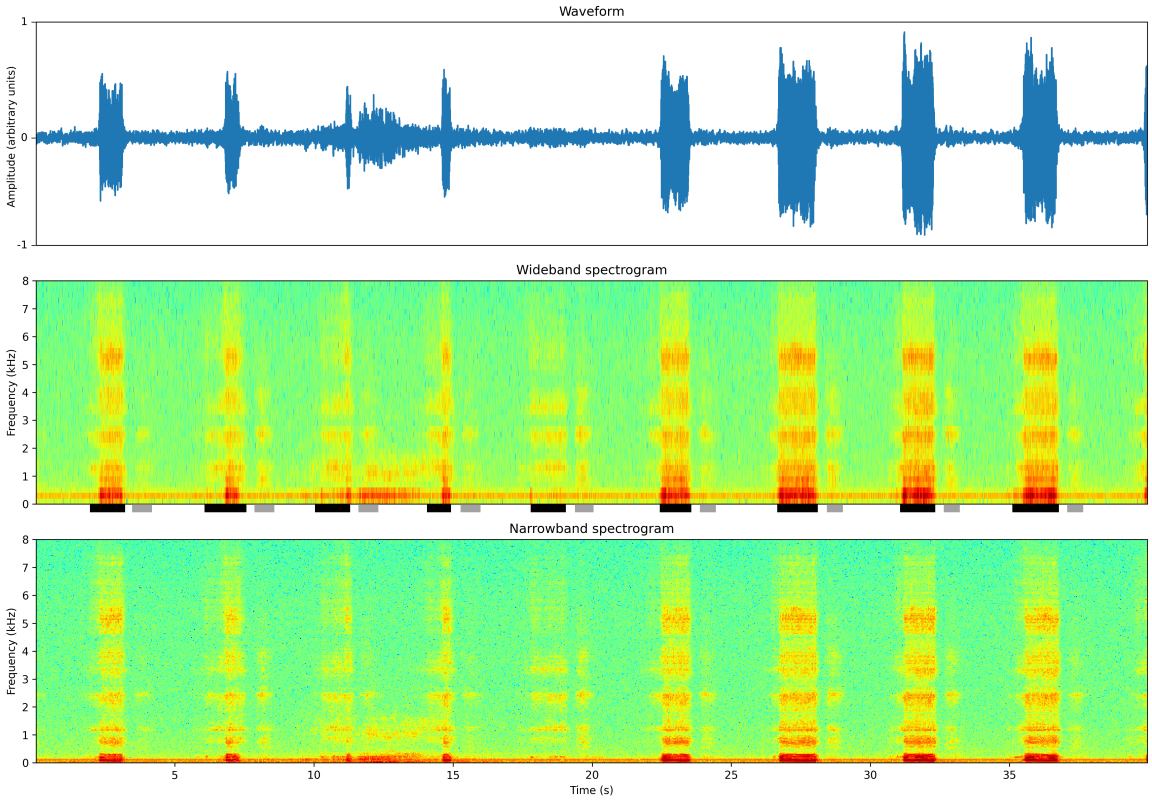


Fig. 3.11 Waveform, wideband and narrowband spectrograms for a 40-second segment from the Snoring Sound Corpus. Inhalations are marked in black, and exhalations are indicated in grey under the wideband spectrogram.

whole spectral range (i.e., 0–8 kHz) for inhalations, and in the 0–4 kHz range for exhalations, as most of the snoring events took place during inhalation and, as will be described next, snoring has a broad spectrum (i.e., 0–8 kHz range).

3.3.2 Snoring

As mentioned before, although there is a commonly understood perception of snoring, there is no generally agreed acoustic definition of it. Dafna et al. (2013) defined snoring as "a breathing sound that [occurs] during an inspiration with an intensity >20 dB" (p. 5) or "a noisy inhalation sound made during sleep that [is] >20 dB" (p. 12). They also noted that other studies have defined snoring as an acoustic event with an oscillatory component or simply "any sound perceived as such by the observer holding the microphone" (p. 12). Huang et al. (1995) defined snoring as a "respiratory noise in general" (p. 97). The duration, pitch, temporal envelope, and spectral shape of the 4,130 snore events in the Snoring Sound Corpus were analysed to define snoring from an acoustic perspective.

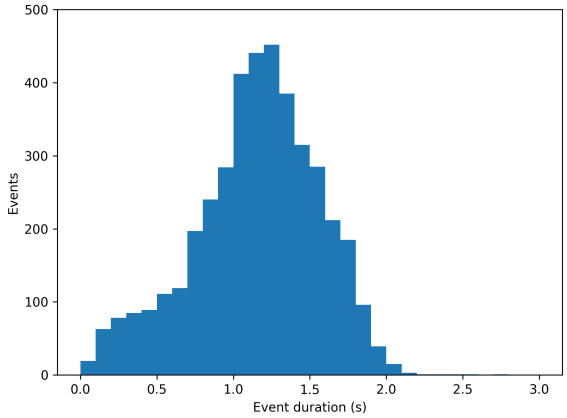


Fig. 3.12 Histogram of the duration of snore events in the Snoring Sound Corpus

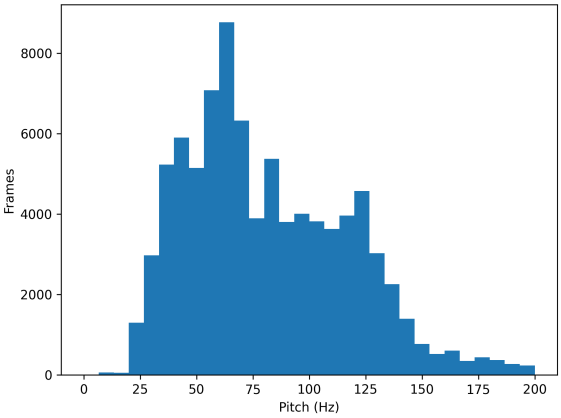


Fig. 3.13 Histogram of the pitch of snore events in the Snoring Sound Corpus

Duration

The snore events in the corpus were extracted taking into account the manual annotation, and their duration was calculated from the start and end timestamps. Figure 3.12 presents the histogram for the event duration. A mean duration of 1.15 ± 0.41 seconds was observed. It can be seen that the duration of snoring follows a normal distribution to some extent: a kurtosis of -0.06 was noted. On the other hand, a mean inter-snore interval of 3.13 ± 1.62 seconds was observed. Similar findings have been published. For example, Levartovsky et al. (2016) reported a mean duration of 1.40 ± 0.07 seconds for men in REM sleep. It is worth mentioning that, unlike Levartovsky et al., we considered snore events in all sleep stages, as this information was not available in the Snoring Sound Corpus.

Pitch

Upper airway dilator muscles exhibit atonia or relative loss of muscle tone during sleep, which causes narrowing and increased resistance in the upper airway. This generates turbulent airflow that results in the vibration of structures like the soft palate, epiglottis, pharyngeal walls or tongue (Pevernagie et al., 2010). Such a vibration is what is perceived as snoring and generates pitch, similar to the vibration of the glottis or vocal cords for speech production. Figure 3.14 summarises this. For this reason, in Section 3.1.1 a snore was succinctly described as a '*pitched* breathing sound' for annotation purposes. In a study on the biomechanics of snoring, Huang et al. (1995) described that "snoring consists of a series of impulses caused by the rapid obstruction and reopening of the upper airway", and found that "this cycle of closure and opening occurs in the region of fifty times per

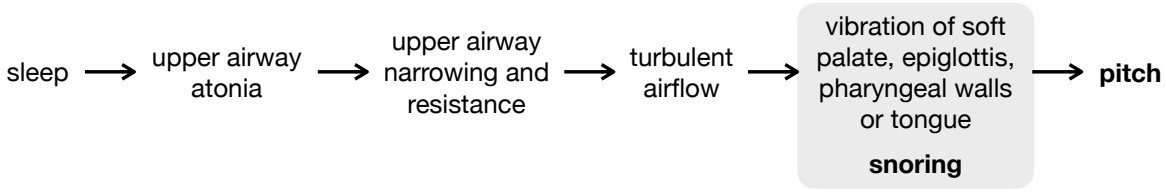


Fig. 3.14 Snoring and pitch

second during a snore" (p. 96). In another study, Liu et al. (2007) noted that "the soft palate vibrates at a frequency of 20–80 Hz during snoring" (p. 864).

The snore events in the corpus were segmented into 100 ms frames with a frame rate of 10 ms, which allowed a minimum pitch estimation of 10 Hz. Pitch was computed for each frame with the YIN algorithm (De Cheveigne and Kawahara, 2002) searching between 10 and 200 Hz. This range was based on what related studies have reported (Huang et al., 1995; Liu et al., 2007). A mean pitch of 85.03 ± 43.44 Hz was observed. Figure 3.13 shows the histogram for the pitch of the snore events in the corpus. Pitch depends on the particular vocal tract characteristics of each participant, whether air is passing through the nose, mouth or both, and where the obstruction is in the vocal tract. For instance, snoring can be caused by the soft palate flapping forwards and backwards, by the collapse of the pharyngeal walls either from front to back or side to side, etc. (Huang et al., 1995). However, the information required to do a further characterisation was not available in the data used, as the determination of the obstruction site, for example, requires drug-induced sleep endoscopy (Kezirian et al., 2011).

Temporal Envelope

The top panel of Figure 3.11 displays the waveform for a typical 40-second segment from the Snoring Sound Corpus. Limited information can be derived from the temporal envelope. It can be seen that snore events occur periodically to some extent. In this example, there is a snore event every 4 seconds approximately. Consistent with the study by Levartovsky et al. (2016), it can also be noted that snoring sticks out from the background noise floor.

Spectral Shape

The middle panel of Figure 3.11 presents the wideband spectrogram for a 40-second segment of a sleep audio recording. A 5-ms window was used for analysis to obtain a good temporal resolution. Snore events display high energy on a broad frequency range, and vertical striations that correspond to pitch pulses can be observed. The bottom panel of Figure 3.11 depicts a narrowband spectrogram of the the same 40-second segment. In this

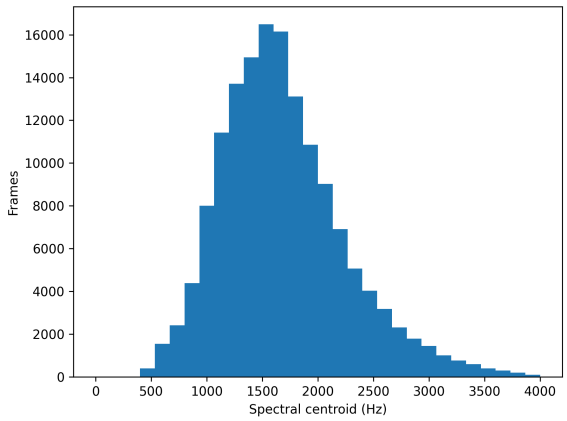


Fig. 3.15 Histogram of the spectral centroid of snore events in the Snoring Sound Corpus

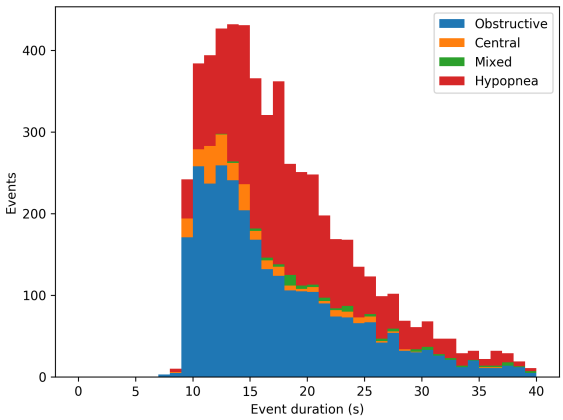


Fig. 3.16 Histogram of the duration of different forms of apnoea events in the OSA Sound Corpus

case, a 40-ms window was used to achieve a better spectral resolution. Some of the formant structure can be distinguished, and the first four formants are around 1,083, 2,286, 3,703 and 5,197 Hz for a typical snore event.

Other studies have reported related findings. Huang (1995) examined a limited number of simulated snore events, and noted that a snore consists of "a series of pulses repeating every 26.5 ms" (p. 3647). Fiz et al. (1996) analysed snore events from 17 subjects, and identified two snoring patterns. The first pattern had a fundamental frequency and several harmonics through a broad frequency range. The second pattern did not have clearly identified harmonics and its sound energy was scattered on a narrow frequency range. Kim et al. (2018) did not describe the spectral shape of snoring, but found that the first three formants were useful to distinguish snore events from non-snore events. Finally, Almazaydeh et al. (2013) noted that a narrower upper airway commonly results in a higher first formant frequency.

Spectral Centroid

With the aim of getting a statistical insight of what is observed in the spectrograms, the spectral centroid was additionally calculated for all the snore events in the corpus. The spectral centroid is a measure of the shape of the spectrum or of its 'centre of gravity' (Giannakopoulos and Pikrakis, 2014). A spectral centroid of 1,687.69 $\pm$ 563.74 Hz was reported. The histogram showing its distribution is presented in Figure 3.15, which suggests that most of the energy of the snore events is located in the lower frequency range. However, Levartovsky et al. (2016) reported a spectral centroid of 3,094 $\pm$ 80 Hz for men in REM sleep. The dissimilarity between their reported values and ours might be due to the differences in data collection and annotation (e.g., definition of snoring), and having con-

sidered snoring in all sleep stages in our study. Heavy snorers tend to snore more during REM sleep, which is characterised by atonia, so the vocal track is at its narrowest state, whereas light snorers evenly snore throughout all sleep stages (Hoffstein et al., 1991).

3.3.3 Apnoea Events

Thus far we have characterised inhalation, exhalation, and snoring from an acoustic perspective. Most of these events are typically audible. By contrast, apnoea events tend to be silent, as they are caused by a partial or complete collapse of the upper airway that results in reduced or absent airflow. However, snoring usually antecedes an apnoea, and breathing resumes after it with a loud gasp. For this reason, the acoustic events that bound an apnoea indicate its occurrence. Here, apnoea events were only characterised in terms of duration.

5,772 manually scored apnoea-hypopnea events from the OSA Sound Corpus were analysed, which included obstructive apnoeas, central apnoeas, mixed apnoeas (i.e., those that begin as central and end as obstructive events), and hypopneas. These events were defined in Chapter 1. Obstructive events had a duration of 17.94 ± 8.30 seconds; central events, 14.96 ± 5.09 seconds; mixed events, 32.22 ± 12.41 seconds; and hypopneas, 19.37 ± 7.98 seconds. This will inform the decision on the segment length for analysis in the OSA screening task that will be presented in Chapter 5. Figure 3.16 displays the stacked histogram for the duration of the different events in the OSA Sound Corpus. Most of the events are obstructive apnoeas and hypopneas, as central sleep apnoeas are common in subjects on high dose opiates, individuals with some neurological disorders, and those with heart failure (Muza, 2015). According to the standard definition, apnoea events have a duration of a least 10 seconds (American Academy of Sleep Medicine, 2020), but a very small number of events with a duration of less than 10 seconds can be noted in the histogram. This was probably caused by manually scoring events *close* to the standard definition, for example, a >90% reduction in airflow for 9 seconds.

3.4 Summary

Healthcare research is a data-scarce field. For this reason, we have collected our own data in the present study, and created the Snoring Sound Corpus and OSA Sound Corpus. Data collection was done with readily available hardware in typical sleep conditions: in a bedroom at home with a smartphone. The Snoring Sound Corpus consists of 354 minutes of manually annotated sleep audio recordings from 6 male participants. An annotation scheme was defined to manually annotate the collected data, and inter-annotator agreement was evaluated using Cohen's kappa. The proposed annotation scheme allowed a greater level of detail in comparison with related studies, and Cohen's kappa was $0.61 \pm$

0.17. The Snoring Sound Corpus will be used in Chapters 4 and 6 to develop and evaluate systems for classification of sleep-disordered breathing events, and snorer diarisation, respectively. The OSA Sound Corpus consists of audio recordings and physiological data (e.g., airflow, respiratory effort, heart rate, oxygen saturation, sleep status, body position, and snoring) collected from 45 participants for 1 or 2 nights during HSAT. Audio recordings amount to more than 416 hours of data. Physiological data was manually scored by a Registered Polysomnographic Technologist, and used as reference. The OSA Sound Corpus will be employed in Chapter 5 to develop and evaluate a system to screen for OSA.

In this chapter the physiology of respiration was introduced as well. Respiration moves oxygen from the atmosphere to the tissues, and carbon dioxide from the tissues to the atmosphere. Air flows into and out of the lungs thanks to periodic changes in alveolar pressure by alteration of the volume of the lungs. Based on it, sleep-disordered breathing was characterised from an acoustic perspective, and this was related to physiological parameters. No correlation was found between sound intensity and airflow, since partial upper airway obstructions or snoring generally cause loud events and decreased flow with respect to healthy breathing, which is not always audible. In general, a participant's exhalations had a longer average duration than inhalations, and exhalations varied more in their duration. 79% of snoring happened during inhalation, whereas 21% occurred during exhalation. Sound energy was usually throughout the whole spectral range (i.e., 0–8 kHz) for inhalations, and in the 0–4 kHz range for exhalations, as snoring has a broad spectrum and mainly occurs during inhalation. Similar findings have been reported by related studies (Huang, 1995; Huang et al., 1995; Levartovsky et al., 2016; Liu et al., 2007).

From the acoustic characterisation of snoring that was reported in this chapter, it can be defined as follows:

> *Snoring is a pitched breathing sound during sleep generated by the vibration of at least one of the following upper airway structures: soft palate, epiglottis, pharyngeal walls or tongue. It is a periodic loud event with a duration of about 1 second, occurs approximately every 3 seconds, its pitch is near 85 Hz, and its spectral centroid is around 1,700 Hz.*

It is worth noting that this definition is based on data from male participants only. While it is true that snoring is more prevalent in men than in women (Chuang et al., 2017), differences between male and female snoring remain to be studied.

In the histograms shown, the great variability of breathing sounds during sleep could be noted, since these are generated in the vocal tract as speech, and the anatomical characteristics of the vocal tract influence their production. Lastly, apnoea events tend to be silent, as they are caused by a partial or complete collapse of the upper airway that results in reduced or absent airflow. However, these are commonly preceded by snoring, and followed by the resumption of breathing with a loud gasp. Therefore, the acoustic events

that surround an apnoea provide cues that suggest its occurrence. Chapter 5 will exploit this to screen for OSA from sleep audio recordings. Using the Snoring Sound Corpus introduced in this chapter, and based on the acoustic characterisation that was carried out, we will investigate the classification of sleep-disordered breathing events in the following chapter.

Chapter 4

Classification of Sleep-disordered Breathing Events

"And thus they give the time, that nature meant for peaceful sleep and meditative snores, to ceaseless din and mindless merriment and waste of shoes and floors."
– L. Carroll

The success of speech technology has inspired research into developing acoustic-based healthcare solutions (Cummins et al., 2018; Latif et al., 2020). Using the Snoring Sound Corpus introduced in the previous chapter and based on the acoustic characterisation of sleep-disordered breathing carried out, the classification of sleep-disordered breathing events — like snoring — in sleep audio recordings is considered in the present chapter. This is done by taking inspiration from speech technology tools. In Chapter 2, it was noted that there is no widely accepted definition of snoring, since it displays great variability in an individual and across individuals. Such variability arises from the fact that breathing sounds are generated in the vocal tract in the same manner as speech, and several factors can have an effect on the way someone snores and can determine if someone snores at all. These include, among others, the sleeping position, having a congested nose, alcohol consumption before going to bed, having a cold, and taking some medications like muscle relaxants.

As we saw in Chapter 2, a general workflow is commonly followed when analysing sleep breathing sounds: data collection, manual annotation, feature extraction, classification, and evaluation. The first two steps were already considered in the preceding chapter, whereas the remaining ones are the focus of the current chapter. Considering such a workflow, there are different approaches to optimising a classifier. One can optimise the features or the *learning protocol* (i.e., information gathering mechanism), as Valiant (1984) in his seminal paper called them. For example, features can be handcrafted based on the characteristics of the data or automatically learned from it. One can also

optimise the classifier architecture or the *deduction procedure*, as Valiant designated it. For instance, generative or discriminative approaches can be implemented using traditional or state-of-the-art machine learning methods. This is usually conditioned by the amount of data available to train the classifier. Additionally, the classifier can be further informed by providing it with more information about the task, e.g., incorporating a 'language' model.

For this reason, this chapter examines different features and classifier architectures for the classification of sleep-disordered breathing events. The features examined include MFCCs, which have been traditionally used as features for ASR and rare sound event detection (Duckitt et al., 2016; Gutierrez et al., 2016), bottleneck features from auditory modelling, and bottleneck features from the autocorrelation function. Bottleneck features are an unsupervised feature representation learned from data, so these allow large amounts of unlabelled data, which would be costly to annotate, to be leveraged. Several classifier architectures including GMM, HMM and DNN are investigated as well. Some of the work presented in this chapter has previously appeared in (Romero et al., 2019).

The present chapter is organised into five sections. Section 4.1 starts by setting out the evaluation framework to assess the performance of the different features and classifier architectures studied. Classifying sleep-disordered breathing events using MFCCs as input to a GMM and HMM is the focus of Section 4.2. Classification with an HMM and DNN employing bottleneck features from an auditory model is examined in Section 4.3, whereas Section 4.4 considers bottleneck features from the autocorrelation function. Lastly, Section 4.5 summarises the chapter.

4.1 Classifier Performance Evaluation

Before considering different features and classifier architectures for the classification of sleep-disordered breathing events, we will first set out an evaluation framework to assess their performance. This will be done both at frame and event level. As we discussed at the beginning of Chapter 2, an audio waveform is usually split into short frames (e.g., 25- or 30-ms frames) for feature extraction, since an audio signal is quasi-stationary at this level, and signal processing techniques assume that the properties of a signal are stationary or slowly vary with time. We also noted that acoustic events — for instance, a snore or a breath — can last for a few seconds and consist of hundreds of frames. Going forward in this thesis, we will divide the audio signal in frames or events for feature extraction and classification. This also determines the way the classifier performance is evaluated.

Frame-level evaluation compares the output of the classifier with the reference (i.e., manual annotation) frame by frame. Whereas, at event level, similar consecutive frames output by the classifier are merged into an event, and compared with the reference event

by event. Event-level evaluation does not assess the segmentation quality — correctly marking the start and end of a particular event. Frame-level evaluation addresses this. Proper segmentation is necessary to compute the time spent snoring, for example, which can be useful to evaluate treatment efficacy or the progression of the condition.

At event level, the snore-event error rate (SER) (Equation 4.1), the proportion of correct events, deletions (missing events), substitutions (incorrect events), and insertions (additional events), will be reported for the snore class and compared with a baseline. At frame level, recall or sensitivity (Equation 2.1), precision (Equation 2.3), F-measure (Equation 2.4), and accuracy (Equation 2.5) will be presented for the snore class and compared with a baseline. These metrics will be computed from the number of true positives (snore classified as snore), false positives (non-snore classified as snore), true negatives (non-snore classified as non-snore), and false negatives (snore classified as non-snore). Lastly, confusion matrices will be reported for all classes as well.

$$\text{SER} = \frac{\text{substitutions} + \text{deletions} + \text{insertions}}{\text{reference}} \tag{4.1}$$

4.2 Classification with MFCCs

As previously discussed, MFCCs have been successfully used as features for rare sound event detection (Duckitt et al., 2016; Gutierrez et al., 2016), since they are highly effective in modelling the frequency content of audio signals (Xu et al., 2004), and offer a very compact feature representation that is less correlated than spectral features, which facilitates their modelling (Bridle and Brown, 1974; Mermelstein, 1976). For the first set of experiments, we will use 12 MFCCs along with deltas and accelerations[1] to classify sleep-disordered breathing events in sleep audio recordings. The audio signal was pre-emphasised, a Hamming window (Equation A.2) was applied to 25-ms frames with a frame rate of 10 ms, and a bank of 26 filters was used. Lastly, MFCCs, deltas and accelerations were normalised to zero mean. The Snoring Sound Corpus introduced in Chapter 3 was used to perform the experiments reported in the present chapter. The train set consisted of data from four subjects, and the test set, of data from two different subjects. Therefore, the systems and results presented here are 'subject-independent'. We will look at the classifier architecture next.

4.2.1 Classifier Architecture

With the aim of setting up a baseline, the first set of experiments examined two classifier architectures: GMM, and HMM. The HMM consisted of three states for the 'silence' (x) and

[1]The extraction of MFCCs, and the computation of deltas and accelerations are considered in detail in Appendix A.

Table 4.1 Event-level evaluation

Configuration	Correct	Insertions	Substitutions	Deletions	SER
GMM + 12 MFCCs	92.51%	55.49%	6.79%	0.70%	62.98%
HMM + 12 MFCCs	88.25%	42.60%	9.90%	1.84%	54.35%
HMM + 25 MFCCs	87.87%	56.63%	10.22%	1.90%	68.76%
HMM + 12 MFCCs + PM	89.40%	32.13%	6.41%	4.19%	42.73%
HMM + 12 MFCCs + LM	85.78%	5.90%	7.37%	6.86%	20.13%
HMM + RM bottleneck	90.60%	8.38%	6.67%	2.73%	17.78%
DNN + RM bottleneck	88.83%	8.19%	5.97%	5.21%	19.37%
HMM + ACF bottleneck	88.32%	29.27%	8.19%	3.49%	40.95%
DNN + ACF bottleneck	92.25%	27.37%	4.32%	3.43%	35.11%

'other' (o) classes, five states for the 'breath' (b) class, and seven states for the 'snore' (s) class. The annotation scheme used was defined in Chapter 3. The output distribution of each of these states was modelled by a mixture of seven Gaussians, and the number of states was heuristically selected according to the complexity of each acoustic event. For example, silence is simply low background noise, whereas snoring displays a great variability, since it can be caused by the vibration of different structures in the upper airway, and can occur during inhalation or exhalation, as we saw in the previous chapter.

Tables 4.1 and 4.2 report the performance of the different classifier configurations considered in this chapter evaluated at event and frame level, respectively. As expected, the HMM (HMM + 12 MFCCs) performed better than the GMM (GMM + 12 MFCCs), since an HMM considers the temporal characteristics of the events unlike a GMM. In Chapter 2 we noted that an HMM model infers a sequence of hidden labels from a sequence of observations in contrast to a GMM, which infers a label from a single observation. Specifically, an HMM models an event — for instance, a snore or a breath — as a sequence of different states (Jurafsky and Martin, 2014a), whereas a GMM simply models an event as similar consecutive observations. HMMs have nonetheless limitations. One of these is the Markov assumption: the probability of the current state depends only on the previous state (Rosenblatt, 1974). Therefore, HMMs do not consider the wider history of previous states, which might provide additional temporal context for inference.

At event level (Table 4.1), the GMM achieved a SER of 62.98% due to a very high proportion of insertions, 55.49%, whereas the HMM obtained an SER of 54.35% thanks to a smaller proportion of insertions, 42.60%, in comparison with the GMM. Nevertheless, both classifier architectures resulted in a very high proportion of insertions, as they output snore events with unrealistic durations. A great amount of events with a duration << 1 second were output, and a snore event has an average duration of 1 second, as we

Table 4.2 Frame-level evaluation

Configuration	Accuracy	Precision	Recall	F-measure
GMM + 12 MFCCs	93.93%	95.50%	83.30%	88.98%
HMM + 12 MFCCs	95.35%	93.67%	90.29%	91.95%
HMM + 25 MFCCs	94.78%	94.79%	87.05%	90.76%
HMM + 12 MFCCs + PM	96.59%	96.52%	91.72%	94.06%
HMM + 12 MFCCs + LM	96.18%	92.43%	94.80%	93.60%
HMM + RM bottleneck	97.49%	95.88%	95.60%	95.74%
DNN + RM bottleneck	94.72%	91.50%	90.46%	90.98%
HMM + ACF bottleneck	94.14%	88.83%	91.62%	90.20%
DNN + ACF bottleneck	88.82%	75.56%	91.71%	82.86%

saw in the previous chapter. Later in this chapter, we will investigate the inclusion of a 'language' model to enforce realistic event durations.

At frame level (Table 4.2), the GMM attained an F-measure of 88.98%, as high precision, 95.50%, but low recall, 83.30%, were reported. The HMM achieved an F-measure of 91.95%, since a balance between precision, 93.67%, and recall, 90.29%, was obtained. It showed an improvement in recall and accuracy with no important impact on precision, which is always a desired characteristic of a classifier. A higher recall evidences that less false negatives (i.e., snore frames incorrectly classified as non-snore) were output by the classifier. This indicates that the HMM classifier is better suited for distinguishing snore from non-snore frames than the GMM, since the former considers the temporal characteristics of the events, which allows it to better deal with the great variability of snoring, as we discussed before.

Additionally, in the frame-level confusion matrices for each classifier presented in Figures 4.1 and 4.2, it can be seen that the HMM performed better at snore, breath and other classification. But the GMM was better at silence classification than the HMM, as less silence frames were classified as breath frames by the GMM. This was probably caused by the similarities between silence (i.e., low background noise) and quiet breathing, and the HMM having a better breath recall, which resulted in more silence frames being incorrectly classified as breath frames. Lastly, for the experiments that will be discussed next, the described HMM will be used as baseline for comparison.

4.2.2 Incorporating Periodicity Information

As we saw in Chapter 3, snoring is a *pitched* breathing sound, as it results from the vibration of soft tissue in the vocal tract due to a partial obstruction of it. ASR systems based on MFCCs normally take into account only 12 MFCCs, since they capture information about

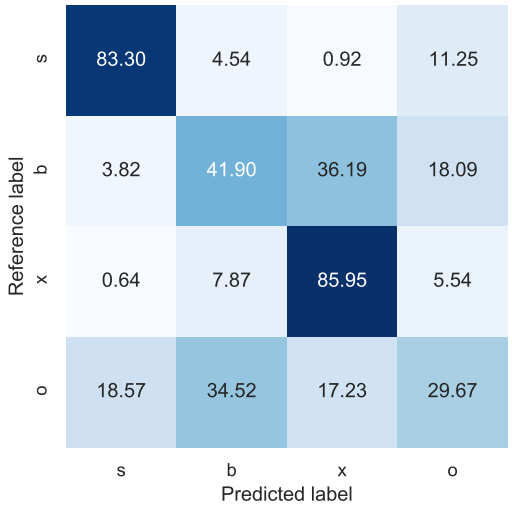


Fig. 4.1 Frame-level confusion matrix for the classification of sleep-disordered breathing events with a GMM using 12 MFCCs. s: snore, b: breath, x: silence, o: other.

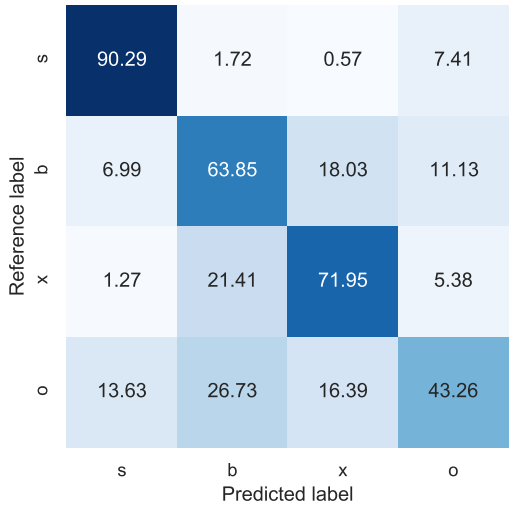


Fig. 4.2 Frame-level confusion matrix for the classification of sleep-disordered breathing events with an HMM using 12 MFCCs (baseline). s: snore, b: breath, x: silence, o: other.

the filter or the state of the vocal tract. However, as we previously discussed, these do not contain information about the source or the characteristics of the vocal tract, such as pitch, since it is not relevant for most ASR tasks. Incorporating information about the source — specifically, periodicity information — should aid in the classification of sleep-disordered breathing events, as this would potentially allow a better differentiation between pitched and unpitched breathing sounds.

Increasing the Number of MFCCs

One possible approach to incorporating periodicity information may be considering higher-order MFCCs, as a more detailed representation of the spectrum might provide spectral cues for pitch. Before evaluating this hypothesis in the classification of sleep-disordered breathing events, we will first illustrate it with a speech signal in Figure 4.3. It presents the MFCCs for the /a/ phoneme uttered by the same male speaker. The top panels display 12 MFCCs for this phoneme produced with a pitch of 102 Hz and 420 Hz, respectively. There are differences between them but the overall pattern is about the same in both cases, so the effect of the pitch difference is not strikingly clear. The bottom panels show 25 MFCCs for the same configurations. This time, there is a clear difference between both plots from the 12th to the 18th coefficients, which suggests that, at least for this particular example, including higher-order MFCCs does provide spectral cues for pitch.

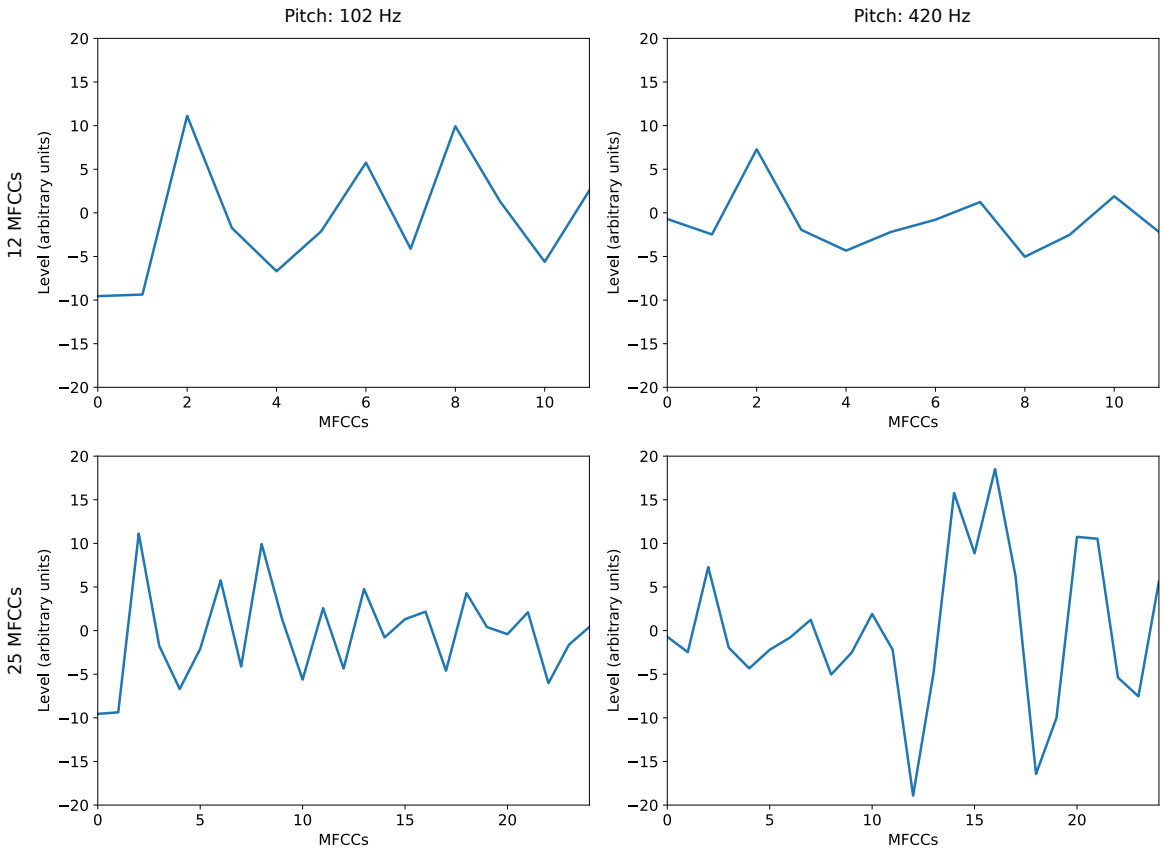


Fig. 4.3 12 (top panels) and 25 MFCCs (bottom panels) for the /a/ phoneme uttered by the same male speaker with a pitch of 102 Hz (left panels) and 420 Hz (right panels)

To evaluate this hypothesis in the classification of sleep-disordered breathing events, an HMM classifier using 25 MFCCs along with deltas and accelerations was trained and compared with the proposed baseline, which uses 12 MFCCs. When using 25 MFCCs (HMM + 25 MFCCs) a SER of 68.76% was obtained at event level (Table 4.1). It was higher than the baseline SER, 54.35%, as the proportion of correct snore events was about the same but the proportion of insertions, substitutions and deletions increased. Similar to the baseline, this configuration resulted in a great amount of events with an unrealistic duration (i.e., $<<$ 1 second).

At frame level (Table 4.2), using 25 MFCCs resulted in a slightly lower F-measure, 90.76%, in comparison with the baseline, 91.95%, since the precision was increased but the recall was decreased. This evidences that less false positives (i.e., non-snore frames incorrectly classified as snore) but more false negatives (i.e., snore frames incorrectly classified as non-snore) were output by the classifier. It can be seen in the frame-level confusion matrix in Figure 4.4 that the performance on breath classification was improved, as less breathing frames were incorrectly classified as snore frames thanks to incorpo-

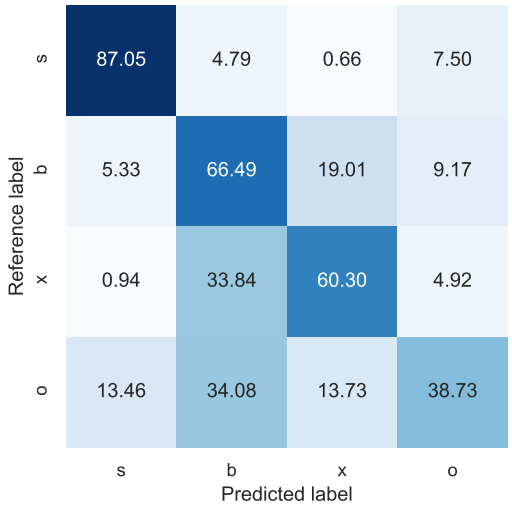


Fig. 4.4 Frame-level confusion matrix for the classification of sleep-disordered breathing events with an HMM using 25 MFCCs. s: snore, b: breath, x: silence, o: other.

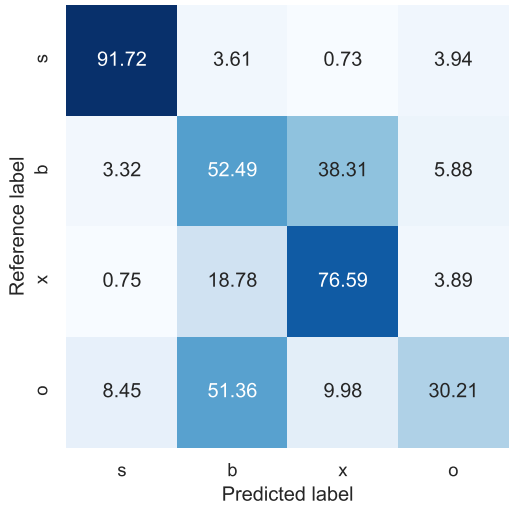


Fig. 4.5 Frame-level confusion matrix for the classification of sleep-disordered breathing events with an HMM using 12 MFCCs and the periodicity measure. s: snore, b: breath, x: silence, o: other.

rating periodicity information. However, the performance on silence classification was decreased an absolute 12%, since periodicity information does not aid in the differentiation between silence and breath frames because both are unpitched sound events. As a result, more silence frames were incorrectly classified as breath frames.

Higher-order MFCCs provide some spectral cues for pitch. In fact, previous studies have estimated pitch from MFCCs for speech reconstruction (Rao and Ghosh, 2017; Shao and Milner, 2005). However, using higher-order MFCCs to incorporate periodicity information did not provide an improvement in the classification of sleep-disordered breathing events, as they are more susceptible to the influence of noise in comparison with lower-order MFCCs (Qaisar, 2019). For this reason, the performance was, in general, lower than the baseline. Since using higher-order MFCCs did not provide an improvement, we will consider other approaches to capture periodicity information next.

Periodicity Measure

In the previous section, we considered higher-order MFCCs in an attempt to capture pitch information to further inform the classifier but it did not provide an improvement, since this was an indirect way of obtaining such an information from a representation that purposely discards it. An alternative might be including a measure derived from a representation that is directly related to the periodicity of an audio signal, for example, from the autocorrelation function. Once more, this should improve the classifier performance,

as it would potentially aid in the differentiation between unpitched and pitched breathing sounds.

The short-term autocorrelation function has been used for decades as the basis for many pitch determination algorithms (Hess, 1983). In general, these algorithms split the audio signal into short frames with a duration of 20–50 ms, since at least two pitch periods are required within a frame to detect periodicity. Following this, the autocorrelation function is computed for each frame. Periodic signals give a prominent peak in the autocorrelation function, which is then detected by a peak detector algorithm, and marked as the pitch estimate for a particular frame. Examples of pitch determination algorithms based on the autocorrelation function are the Simplified Inverse Filtering Technique (SIFT) (Cosi et al., 1984), the McLeod Pitch Method (MPM) (McLeod and Geoff, 2005), and the YIN algorithm (De Cheveigne and Kawahara, 2002), which was used in the previous chapter for determining the pitch of snore events in the Snoring Sound Corpus. The autocorrelation function can be defined as:

$$r(d,q) = \sum_{n=q}^{q+K-d} x(n)x(n+d) \tag{4.2}$$

where $x(n)$ is the audio signal, K is the frame length, q is the starting point of the frame, $q + K$ is its ending point, and d is the delay time or lag. The autocorrelation function of a periodic signal displays a strong peak when the delay d equals the period of the signal (Hess, 2008). Based on this, a 'periodicity measure' was proposed to further inform the classifier. It is computed as follows:

1. RMS normalisation and clipping are applied to the audio signal. This preprocessing step results in a more discriminative periodicity measure.
2. The normalised and half-wave rectified audio signal is split into 25-ms frames with 10-ms overlap. These are the same parameters for extracting MFCCs.
3. The autocorrelation function is computed for each frame as shown in Equation 4.2.
4. The largest peak after zero lag is detected in the autocorrelation function.
5. The ratio between the largest peak after zero lag and the autocorrelation value at zero lag is calculated. This is the periodicity measure.
6. The periodicity measure is appended to the MFCCs.

This is illustrated in Figure 4.6. The top panel displays an RMS-normalised and half-wave rectified snore frame and the bottom panel, its autocorrelation function. The green

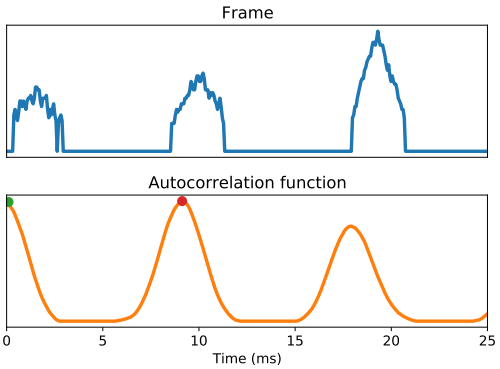


Fig. 4.6 A 25-ms snore frame from a sleep audio recording and its autocorrelation function

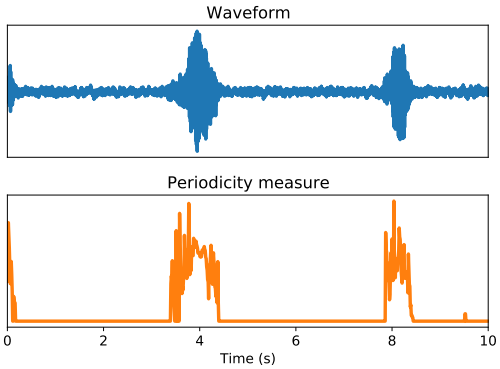


Fig. 4.7 A 10-second segment from a sleep audio recording and its periodicity measure

dot indicates the autocorrelation value at zero lag and the red dot, the autocorrelation value at the largest peak after zero lag. The autocorrelation value at zero lag provides information on the energy of the frame, and the autocorrelation value at the largest peak after zero lag occurs at the period of the signal. The peak detection algorithm also works in a frame-by-frame manner. For each frame, 50-sample windows with 10-sample overlap are analysed within the snoring pitch range[2] looking for the maximum autocorrelation value inside the window. If such a maximum value is not at an edge, it is marked as a peak. After all peaks have been found, the largest peak is selected, which is then used for the calculation of the periodicity measure. In Figure 4.7 a 10-second segment from a sleep audio recording with snore events and its periodicity measure are shown. It can be seen that the periodic sections of the signal or those with snoring have a periodicity measure greater than zero, whereas non-periodic segments or those corresponding to low background noise have a periodicity measure very close to zero or zero.

To evaluate the performance of the proposed periodicity measure in the classification of sleep-disordered breathing events, an HMM classifier using this feature along with 12 MFCCs, deltas and accelerations was trained and compared with the baseline. At event level (Table 4.1), including the periodicity measure (HMM + 12 MFCCs + PM) resulted in a considerably lower SER, 42.73%, in comparison with the baseline SER, 54.35%, as the proportion of insertions, 32.13%, was reduced. This evidences that including the periodicity measure allowed the classifier to output snore events with a more realistic duration, since such a feature successfully discriminates between snore and non-snore frames, as can be seen in Figure 4.7.

At frame level (Table 4.2), a better F-measure, 94.06%, with respect to the baseline, 91.95%, was achieved when including the periodicity measure, since it resulted in a better

[2] As we saw in Chapter 3, this is approximately from 20 to 200 Hz.

precision, 96.52%, and recall, 91.72%, which means that less false positives (i.e., non-snore frames incorrectly classified as snore) and false negatives (i.e., snore frames incorrectly classified as non-snore) were output by the classifier. It can be observed in the frame-level confusion matrix in Figure 4.5 that incorporating periodicity information improved the performance on snore and silence classification but not on breath and other classification, as more breath frames were incorrectly classified as silence and more other frames, as breath in comparison with the baseline. This was probably caused by the fact that snore events are evidently periodic and silence is clearly not periodic, so they were easily distinguished based on periodicity. However, breath and other events can be periodic or not, which rendered the periodicity measure not very useful for discriminating these events.

4.2.3 Including a Language Model

Up to this point we have considered the different events in a sleep audio recording in isolation. However, considering these as part of a sequence of events might potentially result in a more robust classifier, since the temporal pattern of respiration would be considered. One possible strategy is including a 'language' model. To describe what a language model is, the fundamental equation of statistical audio event classification will be used:

$$\hat{L} = \underset{L}{\operatorname{argmax}}\, P(X|L)P(L) \tag{4.3}$$

where X is a sequence of acoustic features, L is a sequence of labels, $P(X|L)$ is the acoustic model, and $P(L)$ is the language model. In this chapter, we have used generative machine learning approaches such as GMM and HMM to model $P(X|L)$. N-gram language models predict the next audio event from the previous $N-1$ events. An N-gram is a sequence of N events. Then an N-gram language model is a statistical model that computes the last event of an N-gram from the previous events (Jurafsky and Martin, 2014b). The simplest approach is a bigram (i.e., a 2-event sequence) language model. That is, the next audio event is predicted from the previous event. In the context of the classification of sleep-disordered breathing events, such an approach can be applied to model the probability of a certain breathing event being followed by another particular event. For example, it would model the probability of a snore being followed by a breath. A bigram model is a reasonable approach, as only four different events are considered in the task at hand.

A bigram language model was learned from the training data: statistics on the occurrence of every possible 2-event sequence (e.g., snore-breath, snore-silence, snore-other, breath-snore, etc.) were calculated from the manual annotation. With the aim of evaluating the contribution of a language model in the classification of sleep-disordered breathing events, an HMM classifier using 12 MFCCs with deltas, accelerations and the learned

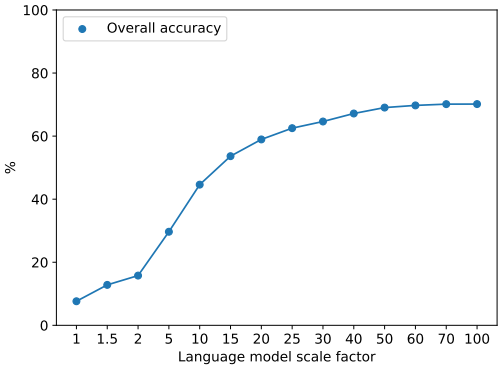


Fig. 4.8 Accuracy for all classes achieved at different language model scale factors

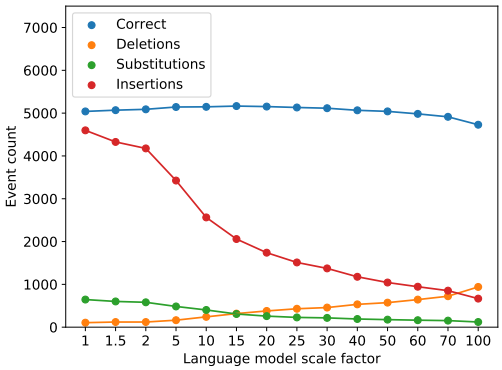


Fig. 4.9 Number of correct events, deletions, substitutions and insertions for all classes at different language model scale factors

bigram language model was trained and compared to the baseline. A scale factor is typically applied to the language model. It works as a weight parameter that instructs the classifier how important the language model should be for classification. Specifically, it scales $P(L)$ in Equation 4.3 according to its relevance. The scale factor was optimised on a subset of the training data, and selected as the value that allowed a balance between overall accuracy and the number of correct events, deletions, substitutions and insertions. All classes were taken into account to compute these metrics. This is depicted in Figures 4.8 and 4.9. Based on this, a scale factor of 30 was used.

At event level (Table 4.1), including the language model (HMM + 12 MFCCs + LM) resulted in a notable improvement in SER, 20.13%, in comparison with the baseline, 54.35%. This is because the language model encouraged realistic event durations, which was reflected in a considerably lower proportion of insertions, 5.90%. For the same reason, the proportion of deletions was increased to a small degree, as some short events were merged into a single event. Although the impact of including a language model at frame level (Table 4.2) was not as obvious as the impact at event level, a better F-measure, 93.60%, was obtained with respect to the baseline, 91.95%, thanks to a better recall, 94.80%, but a slightly lower precision, 92.43%. This means that less false negatives (i.e., snore frames incorrectly classified as non-snore) but more false positives (i.e., non-snore frames incorrectly classified as snore) were output by the classifier, since, by encouraging realistic snore event durations, some non-snore frames were incorporated into a snore event. Lastly, it can be noted in the frame-level confusion matrix in Figure 4.10 that including the language model also resulted in a moderately better performance at breath, silence and other classification thanks to considering the temporal pattern of breathing sounds during sleep. The performance on the 'other' class was still the lowest amongst all classes,

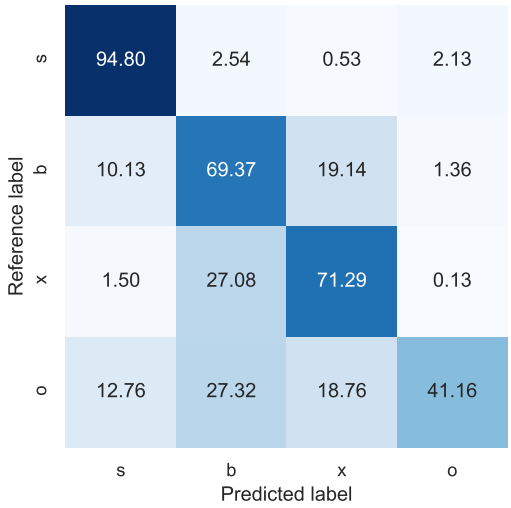


Fig. 4.10 Frame-level confusion matrix for the classification of sleep-disordered breathing events with an HMM using 12 MFCCs and including a bigram language model. s: snore, b: breath, x: silence, o: other.

as its great variability (e.g., it includes coughing, speech, etc.) did not allow it to properly exploit the language model.

4.3 Classification with Bottleneck Features from an Auditory Model

Although MFCCs have been successfully used as features for rare sound event detection (Duckitt et al., 2016; Gutierrez et al., 2016; Janjua et al., 2019), as we just saw in the previous section, these were initially developed for ASR. One would expect that using features specifically designed for a particular kind of data should result in more robust systems. Considering that labelled sleep audio recordings are not widely available, as discussed in Chapter 3, a possible approach to designing features is leveraging large amounts of unlabelled data to learn new feature representations.

In the present study, we have applied unsupervised learning to learn bottleneck features from an auditory-motivated time-frequency representation that better captures pitch information in comparison with MFCCs, since it offers an expanded low-frequency characterisation of the spectrum of the audio signal (Wang and Brown, 2006). Bottleneck features were learned from a large amount of unlabelled data: sleep audio recordings collected to create the Snoring Sound Corpus that were not manually annotated. This novel feature representation was used to train and test different classifier architectures, in the same way as it was done in the previous section for MFCCs. The feature learning process, and the obtained results will be presented next.

4.3.1 Auditory Nerve Firing Rate Map

Auditory-motivated representations have been successfully used in sound understanding applications (Cooke et al., 2001; Coy and Barker, 2007; Rahali et al., 2014). One such representation is the *auditory nerve firing rate map* or *cochleagram*, which represents the average firing rate of auditory nerve fibres, and is typically used in computational auditory scene analysis (CASA)[3] systems. Specifically, it is a time-frequency representation obtained with a computational model of the human peripheral auditory system consistent with the cochlear frequency selectivity. This selectivity is modelled with a gammatone filterbank: a bank of filters with overlapping pass bands derived from psychophysical and physiological studies. Each gammatone is a bandpass filter that models a human critical band, whose impulse response is a tone windowed by a gamma function:

$$g_{f_c}(t) = t^{N-1} \exp\left(-2\pi t b(f_c)\right) \cos\left(2\pi f_c t + \phi\right) u(t) \tag{4.4}$$

where f_c is the filter centre frequency in Hz, N is the filter order, $b(f_c)$ is the bandwidth for centre frequency f_c, ϕ is the phase, and $u(t)$ is the unit step function. A filter order $N = 4$ gives a good approximation to the impulse response function of auditory nerve fibres. Gammatone filters are equally spaced on the equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990; Patterson et al., 1992):

$$\mathrm{E}(f) = 21.4 \log_{10}(0.00437 f + 1) \tag{4.5}$$

where $E(f)$ is the number of ERBs and f is the frequency in Hz. The ERB scale is a warped frequency scale comparable to the critical-band scale of the human auditory system, whose filter centre frequencies are equally spaced in accordance with their bandwidth. The bandwidth, $b(f)$, for fourth-order filters, $N = 4$, is defined by:

$$b(f) = 25.2 + 0.110 f \tag{4.6}$$

The response of each filter or frequency channel can be seen as the instantaneous firing rate within an auditory nerve fibre. This is obtained by smoothing the time series associated with each frequency channel using a Hann window (Equation A.3) (e.g., on 20-ms frames), and downsampling the output (e.g., at 5-ms intervals) (Brown and Cooke, 1994; Wang and Brown, 2006).

For the experiments that will be presented next, we will use rate maps computed for 25-ms frames with 10-ms overlap using a gammatone filterbank with 64 channels equally

[3]Auditory scene analysis aims to "distinguish individual sound sources from a complex mixture of sounds" (Wang and Brown, 2006, p. 1), and CASA "is the computational field of study that aims to achieve human performance in auditory scene analysis by using one or two microphone recordings of the acoustic scene" (Wang and Brown, 2006, p. 11).

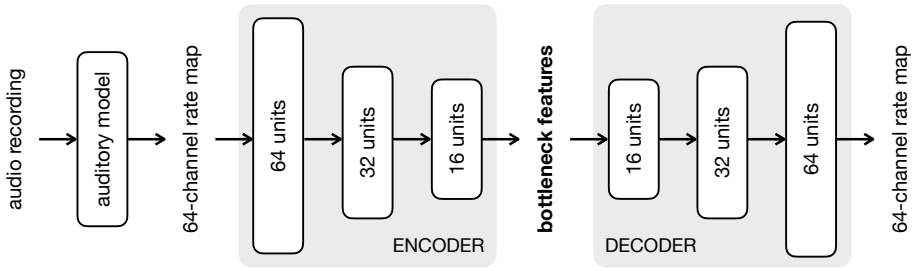


Fig. 4.11 Rate map autoencoder

spaced on the ERB scale between 80 and 7,500 Hz. The frame size and rate were the same as those used for computing MFCCs, and the number of channels was selected based on the frequency range considered and the sampling rate of the sleep audio recordings. Rate maps were root-compressed, and normalised to the 0–1 range. An example is presented in the top panel of Figure 4.12.

4.3.2 Bottleneck Features

Bottleneck features from auditory nerve firing rate maps were learned with a 6-layer autoencoder DNN. The first three layers encode a rate map, and the remaining ones decode it. Specifically, the input to the DNN is a 64-channel rate map, which is encoded to a compressed 16-channel representation using three dense layers of 64, 32 and 16 sigmoid units, respectively. After having encoded the rate map, three additional layers of 16, 32 and 64 sigmoid units, respectively, decode or reconstruct the encoded rate map back to its original 64-channel form. Bottleneck features are the compressed 16-channel representation from the encoder part of the network. The second half of the DNN is only used during training and discarded afterwards. A diagram of the rate map autoencoder is presented in Figure 4.11, and an example of the process, in Figure 4.12. The DNN was developed using TensorFlow (Abadi et al., 2016), and trained in 60 epochs with a learning rate of 0.001, a batch size of 256, and mean squared error as the loss function. The number of epochs and batch size were defined based on the performance on the validation dataset, the learning rate was the standard value that is commonly used (Wu et al., 2009), and the loss function was selected considering the objective of the autoencoder: to reconstruct the original rate map from the compressed representation as closely as possible (Creswell et al., 2017).[4]

An HMM classifier with the same settings as the baseline was trained using bottleneck features, instead of MFCCs, with deltas and accelerations (HMM + RM bottleneck) to examine their performance in the classification of sleep-disordered breathing events. This can be seen as a tandem system (Rath et al., 2014), since acoustic features were extracted

[4]Preliminary experiments conducted to determine the key hyperparameters of the DNNs developed in this chapter are reported in Appendix B.

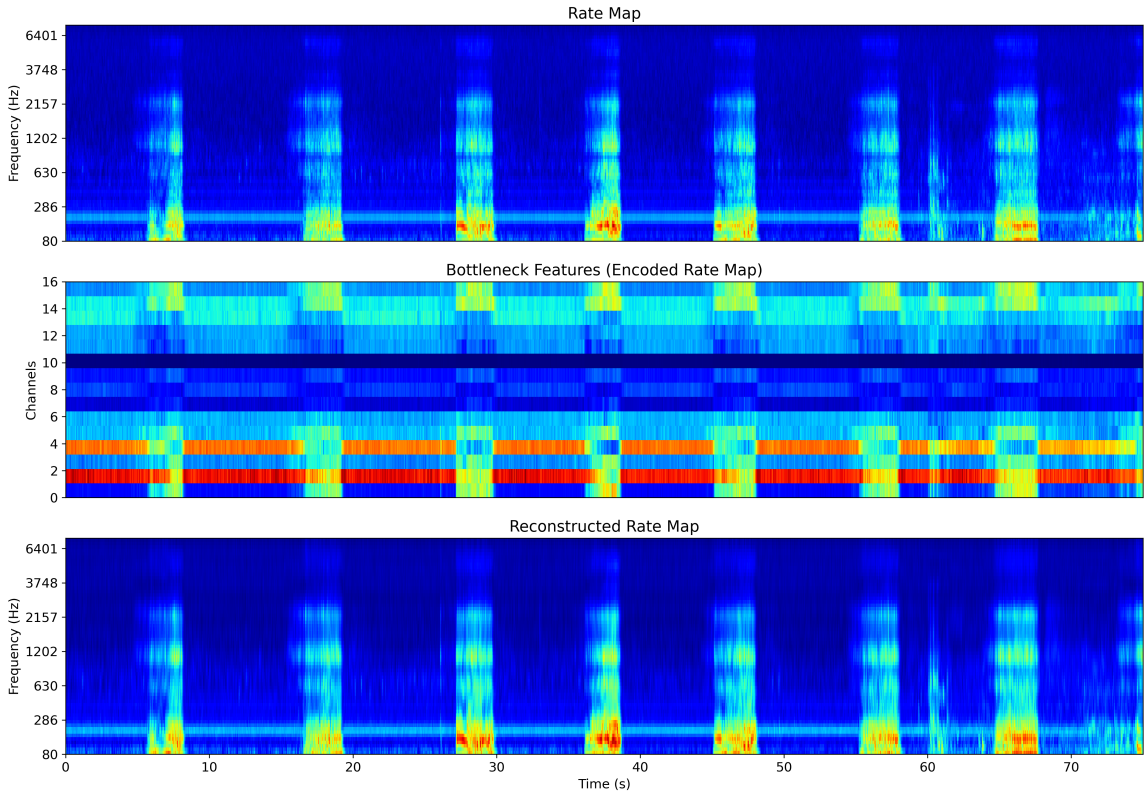


Fig. 4.12 Rate map, bottleneck features or encoded rate map, and reconstructed rate map for a 75-second segment from a sleep audio recording

by a DNN but modelled with a conventional HMM-GMM. At event level (Table 4.1), a SER of 17.78% was achieved, since the proportions of substitutions, 6.67%, and deletions, 2.73%, were lower, and the proportion of correct events, 90.60%, was higher than the best performing classifier configuration using MFCCs (i.e., the one including a language model) with a small impact on the proportion of insertions. In fact, using bottleneck features from rate maps as input to an HMM classifier resulted in the best SER of all configurations considered in the present chapter.

At frame level (Table 4.2), consistent with the event level evaluation, the best F-measure, 95.74%, was obtained when using bottleneck features from rate maps, as the recall, 95.60%, was increased an absolute 12.30% with respect to the baseline while the precision, 95.88%, was statistically the same as the baseline. This evidences that less false negatives (i.e., snore frames incorrectly classified as non-snore) were output by the classifier with respect to the baseline while keeping the same low number of false positives (i.e., non-snore frames incorrectly classified as snore) as the baseline. Also, the highest accuracy (i.e., proportion of true positives and true negatives) was achieved with this configuration as well. It can be noted in the frame-level confusion matrix in Figure 4.13 that

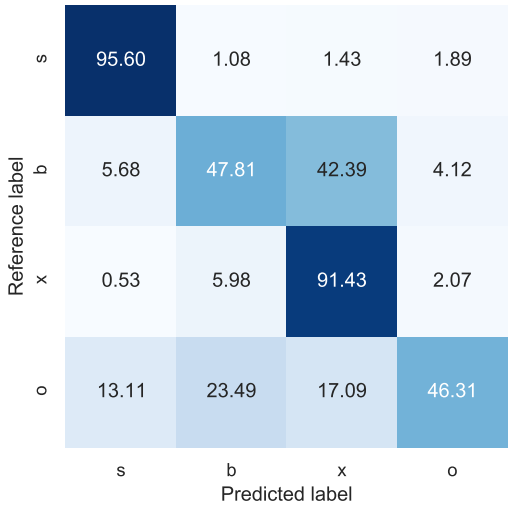


Fig. 4.13 Frame-level confusion matrix for the classification of sleep-disordered breathing events with an HMM using bottleneck features from rate maps. s: snore, b: breath, x: silence, o: other.

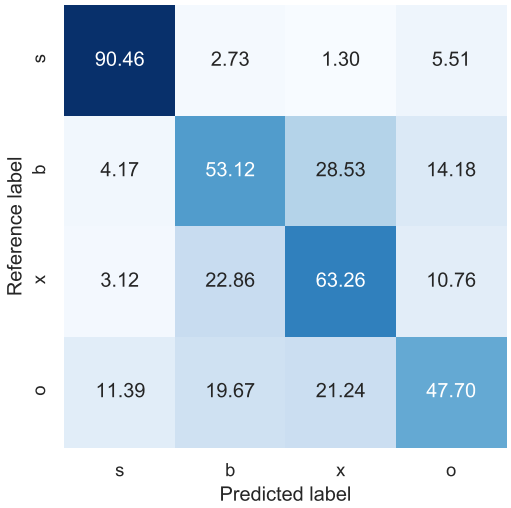


Fig. 4.14 Frame-level confusion matrix for the classification of sleep-disordered breathing events with a DNN using bottleneck features from rate maps. s: snore, b: breath, x: silence, o: other.

the performance at silence classification was improved an absolute 19.48%, because less silence frames were incorrectly classified as breath, but the performance at breath classification was deteriorated, since more breath frames were incorrectly classified as silence with respect to the baseline. We are nonetheless interested mainly in snoring, as it is directly related to sleep-disordered breathing.

We have just seen that using features specifically designed for sleep-disordered breathing data, instead of standard features like MFCCs, resulted in a more robust system. This highlights the 'added value' of learning from data as an alternative to handcrafting feature extraction procedures given the inherent data variability and complexity. For example, our classifiers have to deal with differences in background noise, room acoustics, distance to the microphone, microphone characteristics, breathing pathway (i.e., through the nose, mouth or both), etc. Additionally, rate maps and, by extension, bottleneck features derived from them have advantages over standard acoustic features such as MFCCs. Rate maps offer better spectral resolution at lower frequencies, which is useful to capture pitch information — a key characteristic of snore events — thanks to the ERB-scale spacing of the filter centre frequencies.

Up to this point we have implemented HMM classifiers, since these have been traditionally used in sequence modelling tasks — like ASR — for their simplicity and effectiveness (Gales and Young, 2007), and our focus has been on considering different features with the aim of developing a robust system to classify sleep-disordered breathing

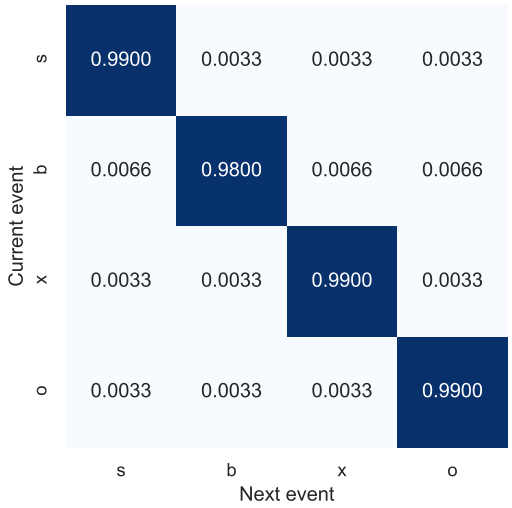


Fig. 4.15 Transition matrix for Viterbi decoding on the DNN output. s: snore, b: breath, x: silence, o: other.

events. However, there are more complex classifier architectures — such as DNN — that might also lead to more robust systems. We have already used a DNN to learn a novel feature representation from the output of an auditory model, and can use another DNN to classify sleep-disordered breathing events. This will be discussed next.

4.3.3 Deep Neural Network

A DNN was trained to classify sleep-disordered breathing events using bottleneck features. It consisted of three hidden layers of 96 sigmoid units, and one layer of 4 softmax units: one for each class (i.e., snore, breath, silence, and other). The output of the network can be interpreted as the probability of an audio event belonging to a particular class. The DNN was developed with TensorFlow (Abadi et al., 2016), and trained in 60 epochs using a learning rate of 0.001, a batch size of 256, and categorical crossentropy as the loss function. As before, the number of epochs and batch size were selected based on the performance on the validation dataset, the learning rate was the standard value (Wu et al., 2009), and the loss function was defined taking into account the objective of the network: to correctly classify the frames in the audio signal into four classes or categories. Reference labels were provided as one-hot vectors.[5]

Viterbi decoding was applied to the output of the DNN with the aim of incorporating global information, and therefore considering the events as part of a sequence rather than in isolation. The Viterbi algorithm is a computationally efficient technique for finding the most probable sequence of events by applying dynamic programming or recursion

[5]One-hot vectors are used to represent a finite set of values. One element is set to 1, and the remaining elements, to 0, for example, snore: [1,0,0,0], breath: [0,1,0,0], silence: [0,0,1,0], and other: [0,0,0,1].

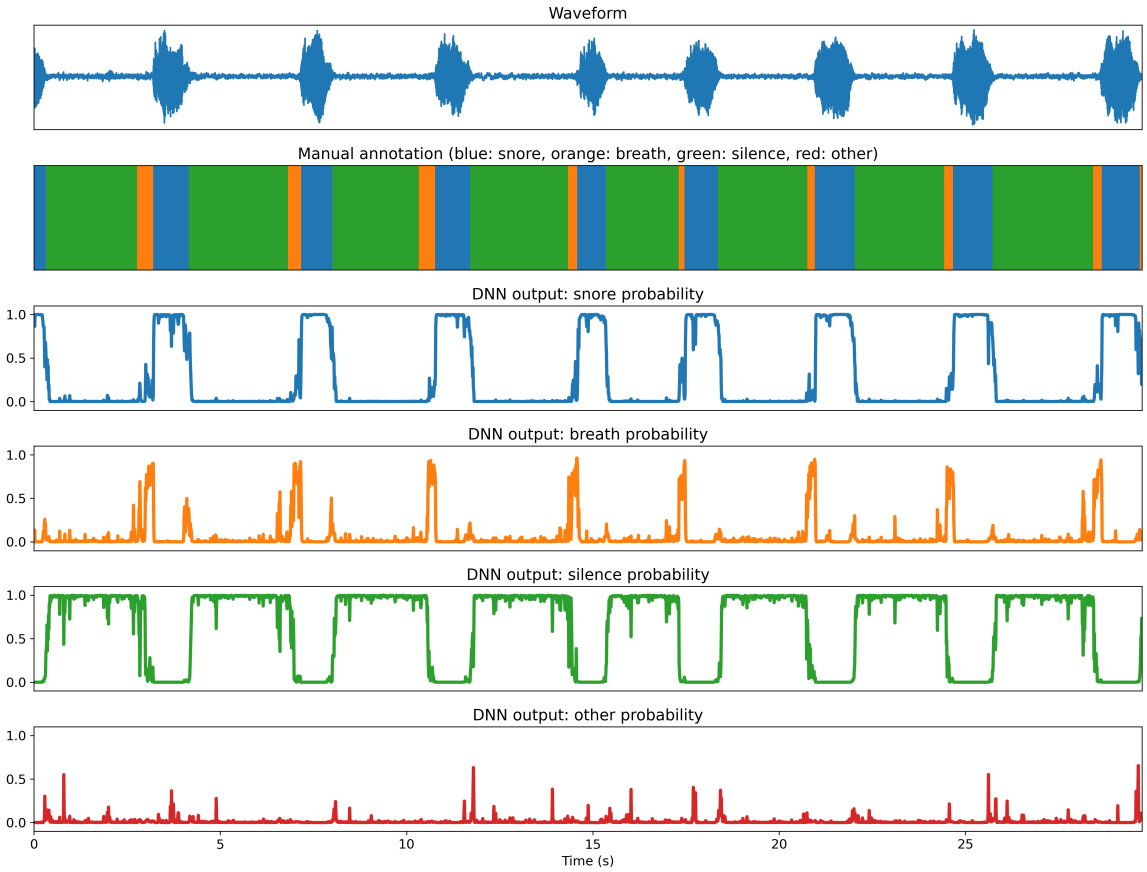


Fig. 4.16 Waveform, manual annotation, and DNN output for a 30-second segment from a sleep audio recording

instead of computing the probability of every possible sequence of events. It takes as input a transition matrix (i.e., an HMM), and the output of the DNN (i.e., the scores or class posterior probabilities for each vector of acoustic features) to create a trellis. Based on this, it computes the optimal path or most probable sequence of events (i.e., states) by exploiting the Markov assumption (Viterbi, 2006). As we previously discussed in the present chapter, it assumes that the probability of the current state is dependent only on the previous state (Rosenblatt, 1974). Transition probabilities were learned from the training data. The resulting transition matrix is shown in Figure 4.15. In this way, the proposed approach can be seen as a hybrid system (Rath et al., 2014), since HMM state probabilities were directly represented by posterior probabilities from the DNN, which mapped acoustic features into class probabilities. That is, the DNN played the role of the acoustic model.

Figure 4.16 shows the waveform for a 30-second segment from a sleep audio recording along with its manual annotation, and the output of the proposed DNN. It can be seen

that the DNN correctly classified all the events in this particular example. At event level (Table 4.1), the performance of the hybrid system (DNN + RM bottleneck) was very similar to that of the tandem system (HMM + RM bottleneck). A SER of 19.37% was obtained, as the proportion of insertions, 8.19%, and substitutions, 5.97%, were about the same but the proportion of deletions, 5.21%, slightly increased and the proportion of correct events, 88.83%, moderately decreased with respect to the tandem system. At frame level (Table 4.2), a lower F-measure, 90.98%, was achieved by the hybrid system in comparison with the tandem system, since the precision, 91.50%, and recall, 90.46%, decreased, which reflects that more false positives (i.e., non-snore frames incorrectly classified as snore) and false negatives (i.e., snore frames incorrectly classified as non-snore) were output by the DNN.

Consistent with the above results, it can be observed in the frame-level confusion matrix in Figure 4.14 that the performance of the hybrid approach was somewhat similar to that of the tandem approach. However, the performance at snore and silence classification was decreased an absolute 5.14% and 28.17%, respectively. This kind of behaviour — lower performance of DNNs in comparison with traditional techniques like HMMs — has been reported by related studies. These have pointed out that "a common observation with DNNs is that they need more training data than, for example, GMM-HMM systems with MFCCs features" (Schröder et al., 2016, p. 1). Data scarcity has been a challenge in this study. In Chapter 7, we will follow up on a number of techniques that could be potentially used in future work to address this, for example, data simulation, transfer learning, data augmentation, etc.

4.4 Classification with Bottleneck Features from the Autocorrelation Function

In Chapter 3 we noted that snoring is a pitched breathing sound. In other words, it is a periodic acoustic event. For this reason, a periodicity measure derived from the autocorrelation function was proposed earlier in this chapter to capture pitch information. However, this measure was obtained with a set of handcrafted rules rather than learned from data. Features could be learned from the raw waveform, for instance, but enough data for an end-to-end system was not available in this study. These could also be learned from a spectral representation, as we did with rate maps, or with other representation that emphasises pitch, like the autocorrelation function. Similar to the bottleneck features from rate maps previously described, we can learn bottleneck features from the autocorrelation function with a DNN. As before, such a learned representation should capture pitch information, which might result in a more robust classifier, since it would aid in the distinction between pitched and unpitched breathing events.

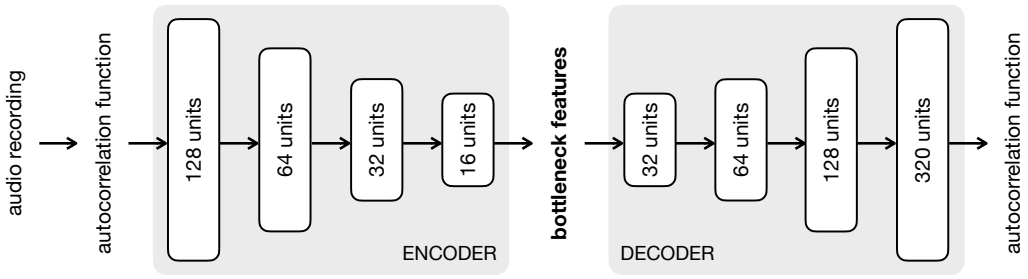


Fig. 4.17 Autocorrelation function autoencoder

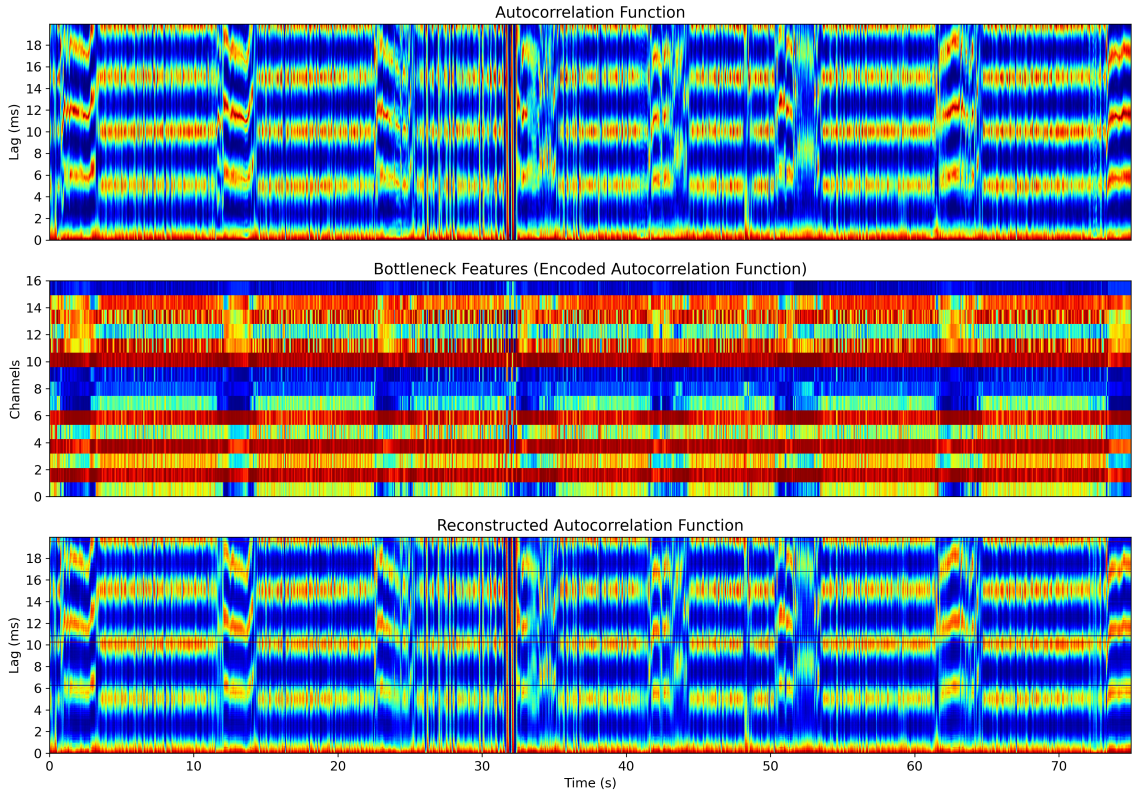


Fig. 4.18 Autocorrelation function, bottleneck features or encoded autocorrelation function, and reconstructed autocorrelation function for a 75-second segment from a sleep audio recording

The autocorrelation function (Equation 4.2) was computed on 25-ms frames with 10-ms overlap — as was done for MFCCs and rate maps — within the period range (i.e., > 50 Hz) to reduce the number of autocorrelation function coefficients by only keeping the relevant ones. A rectangular window was applied to each frame, since in this case we only worked in the time domain. In the same manner as we did in the previous section, bottleneck features from the autocorrelation function were learned using a large amount of unlabelled data: audio recordings collected to create the Snoring Sound Corpus that were not selected for manual annotation. This novel feature representation was used to train and test different classifier architectures. The feature learning process, and the obtained results will be presented next.

4.4.1 Bottleneck Features

Unsupervised learning was applied once more to learn bottleneck features from the autocorrelation function with the aim of obtaining a compressed feature representation that captured periodicity information. Bottleneck features from the autocorrelation function were learned using an autoencoder DNN, in the same way as those from an auditory model.

The input to the DNN is a frame-based 320-coefficient autocorrelation function.[6] The DNN consists of four layers of 128, 64, 32 and 16 sigmoid units, respectively, which encode or compress the autocorrelation function to a 16-channel representation. Another four layers of 32, 64, 128, 320 sigmoid units, respectively, decode or reconstruct the autocorrelation function back to its original 320-coefficient form from its 16-channel representation. That is, the expected output is the input itself. The bottleneck features are the compressed 16-channel representation. This is illustrated in Figure 4.17. The second half of the network is only used during training, and discarded afterwards. An example of this process is depicted in Figure 4.18. The autoencoder DNN was developed in TensorFlow (Abadi et al., 2016), and trained in 60 epochs with a learning rate of 0.001, a batch size of 256 frames, and mean squared error as the loss function. These parameters were chosen in the same way as those of the rate map autoencoder.

To examine the performance of the proposed bottleneck features in the classification of sleep-disordered breathing events, an HMM classifier with the same settings as the baseline was trained using these features with deltas and accelerations (HMM + ACF bottleneck) instead of MFCCs. Once more, this can be seen as a tandem system. At event level (Table 4.1), a lower SER, 40.95%, was achieved in comparison with the baseline thanks to a reduction in the proportion of insertions, 29.27%, as bottleneck features derived from

[6] A 25-ms frame from an audio signal sampled at 16 kHz and its autocorrelation function consist of 400 samples or coefficients. However, we only keep the autocorrelation function coefficients within the period range: > 50 Hz. A frequency of 50 Hz is equivalent to a period of 20 ms, which corresponds to 320 coefficients.

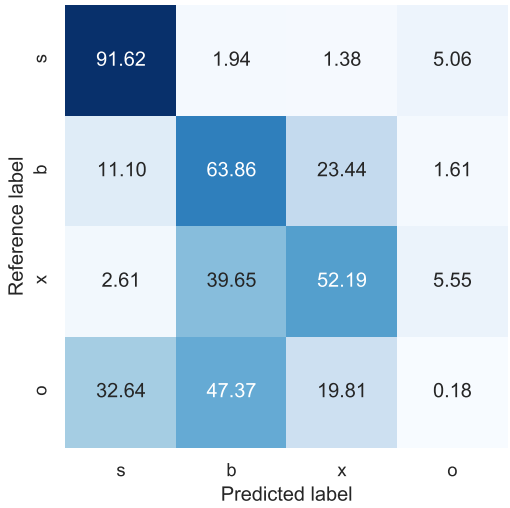


Fig. 4.19 Frame-level confusion matrix for the classification of sleep-disordered breathing events with an HMM using bottleneck features from the autocorrelation function. s: snore, b: breath, x: silence, o: other.

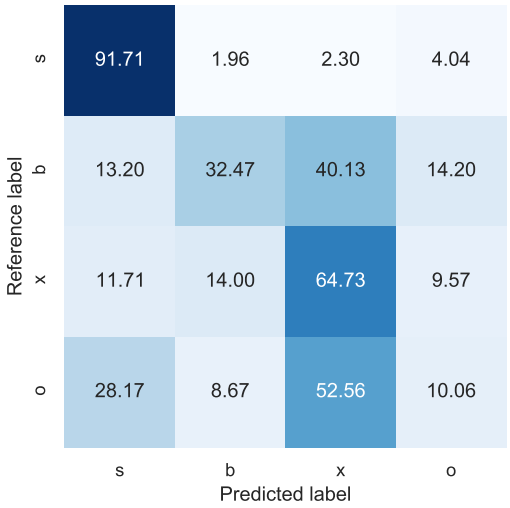


Fig. 4.20 Frame-level confusion matrix for the classification of sleep-disordered breathing events with a DNN using bottleneck features from the autocorrelation function. s: snore, b: breath, x: silence, o: other.

the autocorrelation function allowed a better distinction between pitched and unpitched acoustic events, for example, between snoring and breathing. However, similar to the baseline, some events with unrealistic durations (i.e., << 1 second) were observed. This was caused by the fact that features derived from the autocorrelation function effectively capture periodic characteristics but do not provide comprehensive spectral information, since this function is purely in the time domain rather than a time-frequency representation, like MFCCs or rate maps.

At frame level (Table 4.2), an F-measure, 90.20%, slightly lower than that of the baseline was obtained, since the precision, 88.83%, was reduced. More false positives (i.e., non-snore frames incorrectly classified as snore) were output by the classifier when using bottleneck features from the autocorrelation function, because 'other' events — like speech or music — can be pitched as well. For this reason, 32.64% of other frames were confused with snoring, as can be seen in the frame-level confusion matrix in Figure 4.19. It can also be observed that the performance at snore, 91.62%, and breath, 63.86%, classification was nearly the same as the baseline, but the performance at silence, 52.19%, and other, 0.18%, classification was considerably lower than the baseline, since a great amount of silence and other frames were incorrectly classified as breath. Bottleneck features from the autocorrelation function worked well for discriminating between pitch and unpitched frames (i.e., snore vs. non-snore classification) but not for distinguishing between different unpitched classes (i.e., breath, silence, and other). This confirms that

the frequency content of the audio signal, which the autocorrelation function does not encompass, is crucial for classifying acoustic events.

4.4.2 Deep Neural Network

In the same way as we did in the previous section with bottleneck features from rate maps, a DNN was trained to classify sleep-disordered breathing events using bottleneck features from the autocorrelation function. The DNN consisted of three dense layers of 192 sigmoid units, and one layer of four softmax units: one for every class. The network was developed in TensorFlow (Abadi et al., 2016), and trained in 60 epochs with a learning rate of 0.001, a batch size of 256, and categorical crossentropy as the loss function. These parameters were selected in the same manner as those of previous networks. Reference labels were provided as one-hot vectors, and Viterbi decoding was applied to the output of the DNN. This can be seen as a hybrid system.

At event level (Table 4.1), a slightly better SER, 35.11%, was attained by the hybrid system (DNN + ACF bottleneck) with respect to the tandem system, as the proportions of substitutions, 4.32%, and insertions, 27.37%, were reduced. This was different from the behaviour reported for the tandem and hybrid systems using bottleneck features from an auditory model. However, this was not the case at frame level (Table 4.2). Although the recalls of the tandem and hybrid systems were very similar, the lowest F-measure, 82.86%, of all configurations considered in this chapter was obtained with the hybrid system using bottleneck features from the autocorrelation function, since the precision, 75.56%, considerably decreased in comparison with the tandem system. Consistent with the performance of the tandem system, many false positives were output by the hybrid system, which suggests, once more, that bottleneck features from the autocorrelation function are not suitable for distinguishing different pitched events (i.e., snore vs. other), as no additional cues can be exploited by the classifier (e.g., the frequency content of the audio signal).

Considering the frame-level confusion matrix in Figure 4.20, it can be noted that a lower performance on breath classification, 32.47%, was obtained with the hybrid system in comparison with the tandem system, as 40.13% of breath frames were incorrectly classified as silence. Breath and silence are unpitched events, so their confusion by the DNN is reasonable, since only periodicity information was supplied to it. Lastly, one of the reasons why DNNs commonly require more training data than HMMs to properly generalise (Schröder et al., 2016) is the difference in the number of parameters that need to be learned in each approach. Here, the HMM classifier had over 9,000 parameters, whereas the DNN had more than 135,000 parameters.[7] That is, the DNN had to learn around 15 times more parameters than the HMM from the same amount of data.

[7]This was nonetheless a very small DNN. In Chapter 5 we will use a DNN with over 6 million parameters.

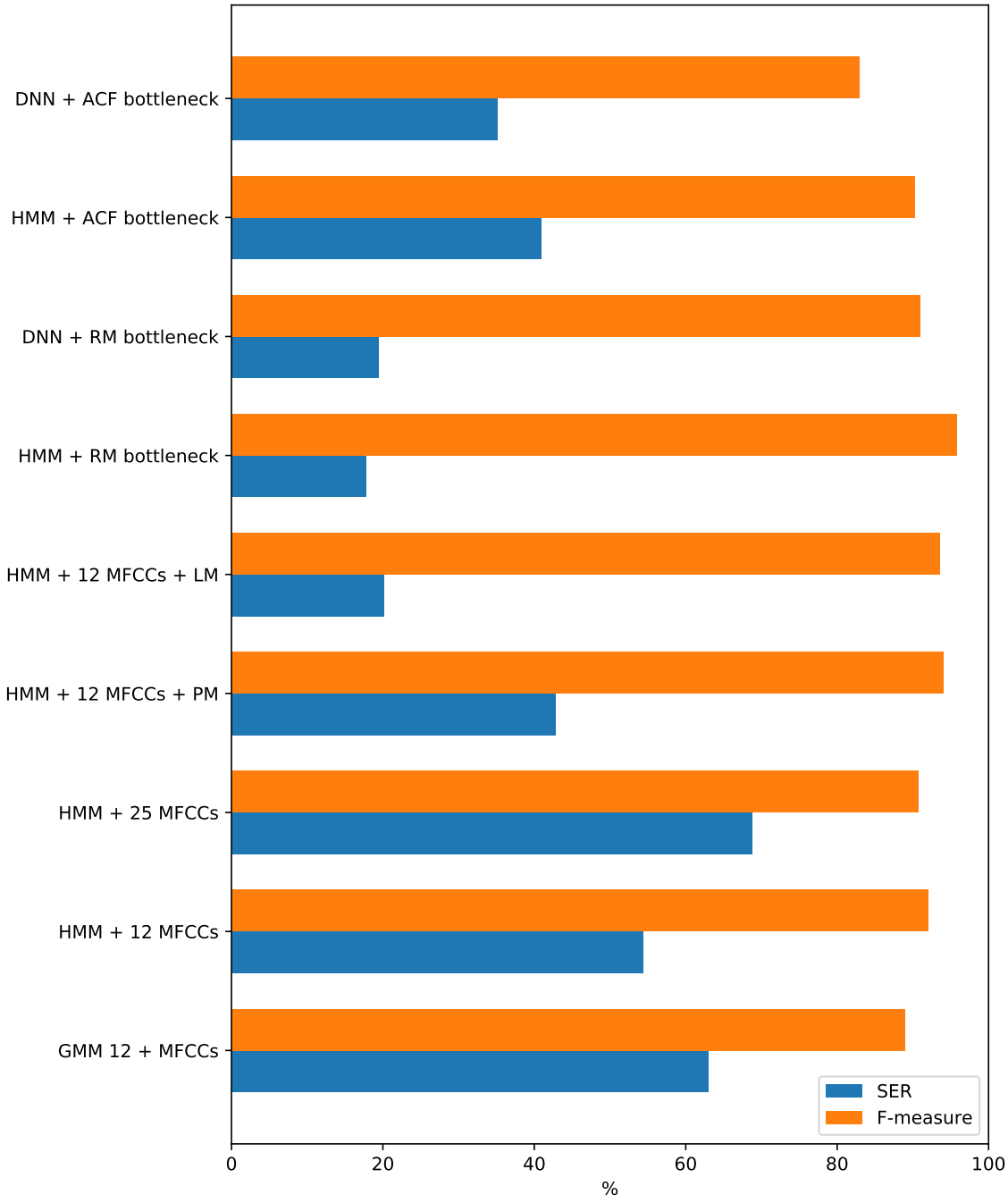


Fig. 4.21 SER (lower is better) and F-measure (higher is better) for the classification of sleep-disordered breathing events with different configurations

4.5 Summary

Research into developing acoustic-based healthcare solutions has been motivated by the success of speech technology, which offers a well defined and extensive framework for acoustic analysis in general. In this chapter the classification of sleep-disordered breathing events — like snoring — in sleep audio recordings was considered by extrapolating some ASR concepts. Different features and classifier architectures were examined with the aim of developing a robust classifier using the Snoring Sound Corpus introduced in Chapter 3. Since classifiers are optimised with respect to defined metrics, an evaluation framework was set out to assess their performance at event and frame level. Figure 4.21 summarises the performance of the different configurations considered at event level with the SER, and at frame level with the F-measure.

An HMM classifier using 12 MFCCs (HMM + 12 MFCCs) along with deltas and accelerations was set up as the baseline (SER: 54.35%, F-measure: 91.95%). The best performing system (SER: 17.78%, F-measure: 95.74%) was obtained with an HMM using bottleneck features from auditory nerve firing rate maps (HMM + RM bottleneck), an auditory-motivated time-frequency representation. Bottleneck features were learned from unlabelled data with a DNN rather than derived from handcrafted feature extraction procedures. This system showed a statistically significant improvement (t-test: p-value < 0.02) over the baseline. Looking at individual snorer's results, no important difference in performance due to snorer variation was observed: the F-measures for snorer 1 and snorer 2 were 93.93% and 96.40%, respectively. The second best performing system (SER: 20.13%, F-measure: 93.60%) was achieved with an HMM using 12 MFCCs along with a bigram language model (HMM + 12 MFCCs + LM). This was learned from data and included to consider the events in a sleep audio recording as part of a sequence of observations rather than in isolation. For example, it modelled the probability of a snore being followed by a breath. The language model encouraged realistic event durations.

Bottleneck features from the autocorrelation function were also proposed with the aim of learning a feature representation that captured pitch information, since snoring is a pitched breathing sound. Reasonable performance (SER: 40.95%, F-measure: 90.20%) was achieved with an HMM using these features (HMM + ACF bottleneck). Although bottleneck features from the autocorrelation function effectively captured periodicity information, they did not encompass the frequency content of the audio signal, which is crucial for classifying acoustic events, since the autocorrelation function is a time-domain representation. This was reflected in a high SER. In contrast, MFCCs are derived from a time-frequency representation but, by design, do not capture periodicity information. Whereas bottleneck features from rate maps are derived from a time-frequency representation that does capture pitch information. For this reason, the latter yielded the best performance.

Classification of sleep-disordered breathing events with a DNN (i.e., a hybrid system) was also examined. Although its performance was close to that of an HMM (i.e., a tandem system), the latter performed better. Discriminative approaches commonly require more training data than generative ones, since the former need to learn a greater number of parameters. This has been noted by related studies (Schröder et al., 2016). The performance of the DNN could have been improved with more training data. However, data scarcity has been a challenge in this study. In Chapter 7, we will follow up on different methods that might be used in future work to address this, for instance, data augmentation, transfer learning, etc.

Unlike most of previous studies (Dafna et al., 2013; Emoto et al., 2018; Nonaka et al., 2016; Xie et al., 2021), which have worked on sleep audio recordings collected in controlled conditions with specialised hardware (e.g., in a sleep clinic with a ceiling-mounted microphone), we worked on recordings made in typical sleep conditions with readily available hardware (i.e., in a bedroom at home with a smartphone). This made the task more challenging, as our classifiers had to deal with different (noisy) acoustic environments and microphone characteristics. For the same reason, this study contributes to improving accessibility to sleep-disordered breathing diagnosis. As noted above, having used a limited amount of data is nonetheless one of the limitations of the present study, since the great variability of breathing sounds during sleep might not have been exhaustively considered. At the time of producing this thesis, there were no immediate plans to make the Snoring Sound Corpus publicly available, as it was collected in collaboration with Passion for Life Healthcare, one of the PhD sponsors, and has particular commercial value to them. However, the work reported in this chapter can be validated on data collected and annotated as closely as possible to our data.

With each chapter of this thesis we are incrementally going through more complex scenarios. In the present chapter we concentrated on detecting snoring, a highly prevalent form of sleep-disordered breathing characterised by short loud acoustic events. In the upcoming chapter we will focus on screening for OSA, the most severe form of sleep-disordered breathing distinguished by (nearly) silent events that last for more than 10 seconds, which makes their detection a more challenging task. As we saw in Chapter 2, previous studies have used snoring to screen for OSA (Alakuijala and Salmi, 2016; Fiz et al., 1996). This approach has limitations. While it is true that 70–95% of subjects with OSA snore, not all snorers suffer from OSA. Furthermore, due to the high prevalence of snoring in the general population, it is a poor predictor of OSA (Maimon and Hanly, 2010). For these reasons, we will not employ snoring to screen for OSA. However, similar to the inclusion of a language model to consider the temporal pattern of respiration proposed in the current chapter, we will also exploit such a temporal pattern for screening.

Chapter 5

Screening for Obstructive Sleep Apnoea

"Many things — such as loving, going to sleep, or behaving unaffectedly — are done worst when we try hardest to do them."
– C. S. Lewis

As discussed in Chapter 1, obstructive sleep apnoea (OSA) is the most severe form of sleep-disordered breathing. Unlike snoring, apnoeas are (nearly) silent events, as they are caused by the collapse of the upper airway during sleep, which results in the cessation of breathing and oxygen desaturation. An apnoea is typically preceded by snoring events indicating a partial collapse of the upper airway. Then, when it collapses completely, the apnoea occurs. After a short while — typically longer than 10 seconds — the need for oxygen causes an arousal, and breathing resumes with a loud gasp. Detecting apnoeas from acoustic evidence alone can be a challenging task, since silence during respiration might be caused by different factors, for example, recovering muscle tone after an arousal and, therefore, breathing quietly.

Testing a subject for the presence of a highly prevalent, high-risk and treatment-available condition with an unobtrusive and inexpensive method is known as screening (World Health Organization, 2020). In this chapter, we exploit the temporal pattern of respiration to screen for OSA from sleep audio recordings. Instead of detecting individual apnoeas or hypopneas, we predict the presence or absence of these events in large analysis windows, for example, 1-minute segments. Using this larger context allows us to effectively capture the temporal pattern in which the apnoea takes place, as described above. The substrate for this process is a time-frequency representation (e.g., auditory nerve firing rate map) of such large analysis windows. Examples are presented in Figure 5.1. In the top panel, the snore events preceding the apnoea can be observed as well as the

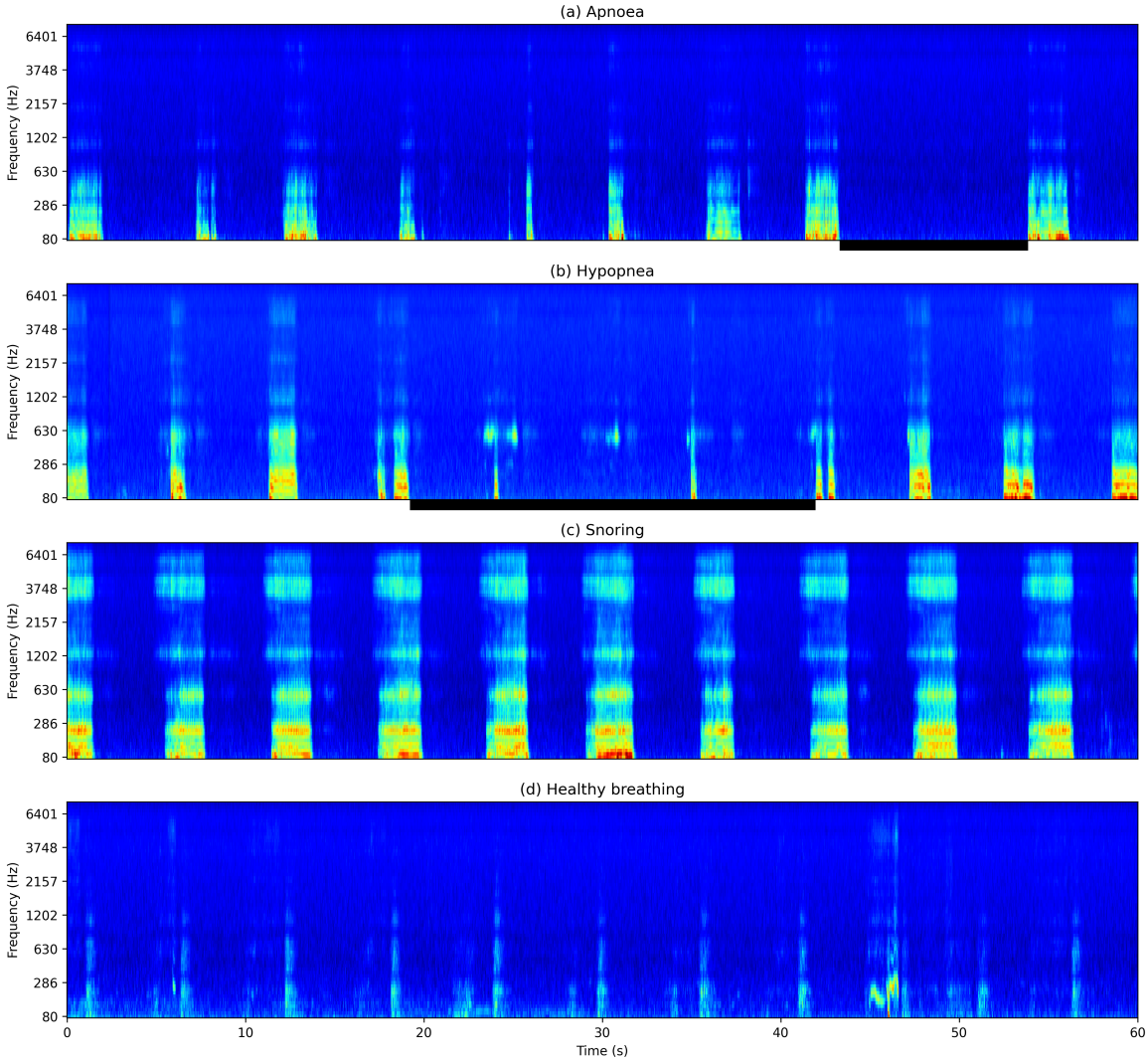


Fig. 5.1 Rate maps for 1-minute segments from the OSA Sound Corpus. Dark blue areas represent low energy, whereas yellow and red regions, high energy. (a) Apnoea: complete silence corresponding to absence of breathing is observed between 42 and 54 seconds. (b) Hypopnea: low-energy events corresponding to shallow breathing are seen between 20 and 40 seconds. (c) Snoring: periodic high-energy events are observed throughout the segment. (d) Healthy breathing: periodic low-energy events are seen throughout the segment.

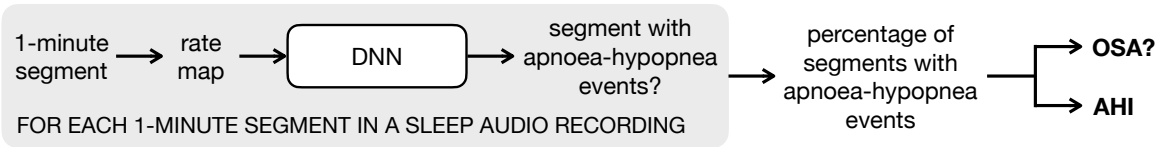


Fig. 5.2 Overview of the system to screen for OSA and predict the AHI

recovery breath after it, which jointly provide context for deeming the silent 12-second segment between them an apnoea.

In the introductory chapter we discussed that PSG — the gold standard for OSA diagnosis — is inconvenient, as it consists in staying overnight in a sleep laboratory while several physiological parameters are measured with sensors attached to the body (Mansukhani et al., 2020). This motivated the acoustic analysis of sleep breathing sounds as a completely unobtrusive alternative to PSG. In the present chapter, after investigating how far can we get using acoustic features on their own to screen for OSA, we will also look into the integration of acoustic evidence with one direct measurement of physiological data — oxygen saturation — and examine if the inconvenience of wearing one sensor results in a better performance in comparison with using acoustic features only. Additionally, since OSA diagnosis is clinically based on the AHI and with the aim of getting a metric closer to the clinical standard, we investigate the estimation of the AHI — instead of just predicting the presence of OSA — from sleep audio recordings, and from them along with physiological data.

An overview of the system to screen for OSA and predict the AHI is shown in Figure 5.2. It takes as input a 1-minute audio recording segment, computes a time-frequency representation, and provides it as input to a DNN with convolutional layers, which classifies each segment as having apnoea-hypopnea events in it or not. OSA screening is based on the percentage of segments classified as having apnoea-hypopnea events in a whole night and a defined threshold. If the value is above the threshold, the participant has OSA. Otherwise, they do not. Lastly, the same percentage value is input to a regression model that maps it into an AHI value. To develop and evaluate the proposed system, the OSA Sound Corpus was created. Some of the work presented in this chapter has previously appeared in (Romero et al., 2020c).

The present chapter is organised into five sections. Screening for OSA using acoustic features is the focus of Section 5.1, whereas the integration of acoustic features with physiological information to screen for OSA is examined in Section 5.2. Section 5.3 considers the task of estimating the AHI, and Section 5.4, the deployment of the developed screening system on mobile devices. Finally, Section 5.5 summarises the chapter.

5.1 Screening for OSA Using Acoustic Features

As mentioned at the beginning of this chapter, given the silent nature of apnoea events, the temporal pattern of respiration was exploited to screen for OSA from sleep audio recordings. These were made in typical sleep conditions with readily available hardware. An overview of the screening system is presented in Figure 5.2. The presence or absence of apnoea-hypopnea events in large audio recording segments is predicted with a DNN

from an auditory-motivated time-frequency representation, and screening is based on the percentage of predicted segments with apnoea-hypopnea events in a night. In addition to using acoustic features to screen for OSA, we investigated their integration with physiological information. Each of these approaches will be considered in detail next.

5.1.1 Acoustic Features

Auditory Nerve Firing Rate Maps

In the previous chapter, auditory nerve firing rate maps were successfully used for the classification of sleep-disordered breathing events, since they provide an expanded low-frequency representation of the spectrum of the audio signal (Brown and Cooke, 1994; Wang and Brown, 2006). These properties are also useful for the task at hand in the present chapter, as we are still dealing with breathing sounds, including snoring. For this reason, rate maps were employed again to screen for OSA. These were computed for large non-overlapping segments to effectively capture the temporal pattern of respiration, since apnoea-hypopnea events usually last for 20–30 seconds, as we saw in Chapter 3. Rate maps were computed on non-overlapping 25-ms frames to reduce the number of features in comparison with those used in the previous chapter, which were computed using a frame rate of 10 ms. Since the whole rate map was provided as input to the DNN, rate maps were normalised over all time-frequency bins to the 0–1 range (Han et al., 2012):

$$z_{ij} = \frac{x_{ij} - \min(x)}{\max(x) - \min(x)} \tag{5.1}$$

where x is the rate map, x_{ij} is its value at time i and frequency bin j, and z_{ij} is the corresponding normalised value. Each segment was labeled using the scored HSAT data as either having apnoea-hypopnea events in it or not. These were not required to start and end within the segment. Examples are presented in Figure 5.1. The first panel presents the rate map for a 1-minute audio recording segment containing an obstructive apnoea. The snore events preceding it, and the recovery breath following it can be observed. The second panel shows a segment with a hypopnea. Unlike the apnoea in the first panel, in which there is no breathing at all, shallow breathing seen as low-energy events takes place during the hypopnea. Snoring events are displayed in the third panel. These are observed as periodic high-energy events. Finally, healthy breathing seen as periodic low-energy events is presented in the last panel.

Bottleneck Features

Continuing with the idea — introduced in Chapter 4 — of learning feature representations to potentially achieve more robust systems, bottleneck features from auditory

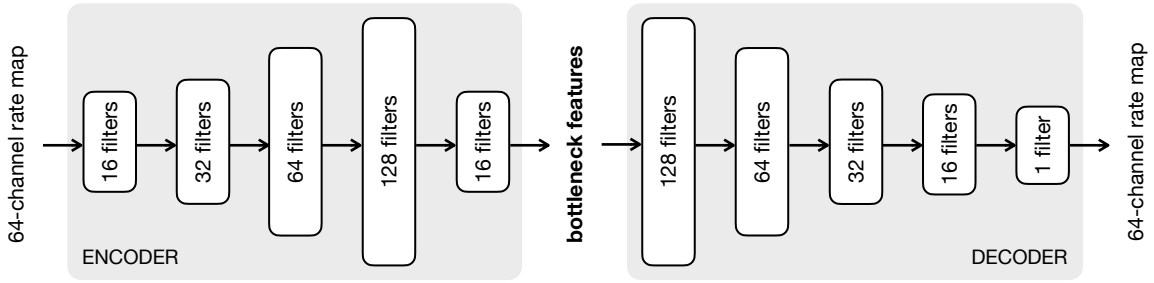


Fig. 5.3 CNN for the extraction of bottleneck features

nerve firing rate maps were learned in an unsupervised way with an autoencoder CNN. These were similar to the bottleneck features proposed in the previous chapter for the classification of sleep-disordered breathing events. However, instead of dense layers, convolutional layers were used here. The neural network was trained with the audio recordings that were not included in the OSA Sound Corpus: 25 nights from 19 participants amounting to over 158 hours of data.

In order to effectively capture the local temporal pattern, 2-second segments (i.e., 80 25-ms frames) with 1-second overlap were used as input to an autoencoder CNN based on an architecture that has been successfully employed in the extraction of features from medical images (Starke et al., 2020). The network consists of 5 convolutional layers (Albawi et al., 2017) of 16, 32, 64, 128 and 16 filters with a kernel size of 3×3 that encode a 64-channel rate map to a compressed 16-channel representation. These are followed by another 5 convolutional layers of 128, 64, 32, 16 and 1 filters with a kernel size of 3×3 that decode or reconstruct the encoded rate map back to its initial 64-channel form. The last convolutional layer only has one filter, since rate maps are 2-dimensional. Successfully reconstructing rate maps requires the bottleneck features to capture the important 'image' characteristics. For this reason, the relevant information is assumed to be encoded within these features. Every encoding layer is followed by a max-pooling layer, whereas every decoding layer is followed by an upsampling layer. Although the number of filters is increased with each encoding layer to obtain more abstract features, max-pooling layers reduce the dimensionality of each convolutional layer output to finally get a compressed feature representation, and avoid overfitting. Rectified linear unit (ReLU) activation was used throughout the network with the exception of the last layer of the encoder and the last layer of the decoder, which used sigmoid activation to obtain bottleneck features and reconstructed rate maps in the original 0–1 range. This is illustrated in Figure 5.3. The second half of the network was discarded after training. Bottleneck features are the compressed 16-channel representation. The autoencoder was trained in 128 epochs with a batch size of 256, a learning rate of 0.001, and mean squared error (MSE) as the loss func-

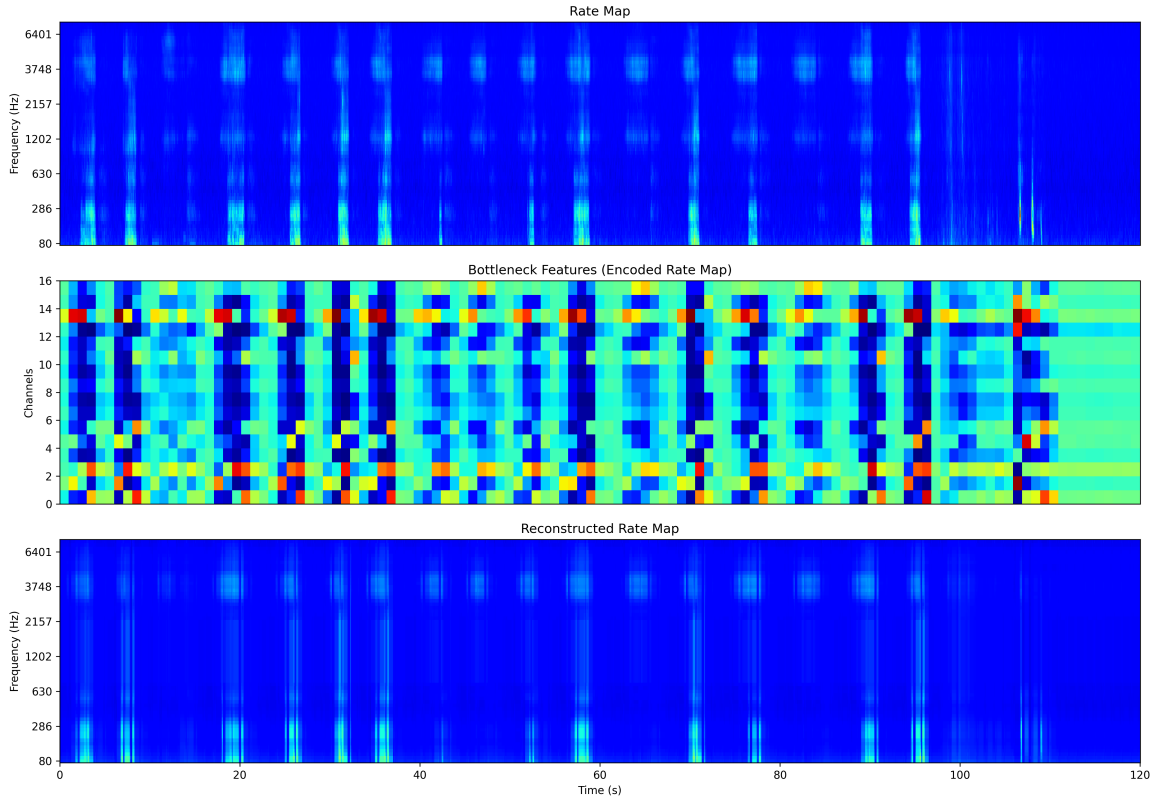


Fig. 5.4 Rate map (top panel), bottleneck features or encoded rate map (middle panel), and reconstructed rate map (bottom panel) for a 2-minute sleep audio recording

tion.[1] It was developed using TensorFlow (Abadi et al., 2016). An example of the process is shown in Figure 5.4. This evidences that the bottleneck features are capturing salient aspects of the rate map, as it is effectively reconstructed from them.

5.1.2 Predicting the Presence of Apnoea-hypopnea Events

Rate maps (and the other features considered) were classified with a DNN. The network consists of three convolutional layers of 32, 16 and 64 filters with a kernel size of 3×3 followed by a dense layer of 1,024 tanh units, and a dense layer of 2 softmax units, one for each of the classes considered: 'segment with no apnoea-hypopnea events' and 'segment with apnoea-hypopnea events'. Each convolutional layer is followed by a max-pooling layer. Further feature extraction is performed by the convolutional layers, and classification is done by the dense layers. This is illustrated in Figure 5.5. The DNN was trained in about 128 epochs with a learning rate of 0.001, a batch size of 64, and categorical crossentropy as the loss function. The loss was monitored on a validation set to prevent overfit-

[1]Preliminary experiments conducted to determine the key hyperparameters of the DNNs developed in this chapter are reported in Appendix B.

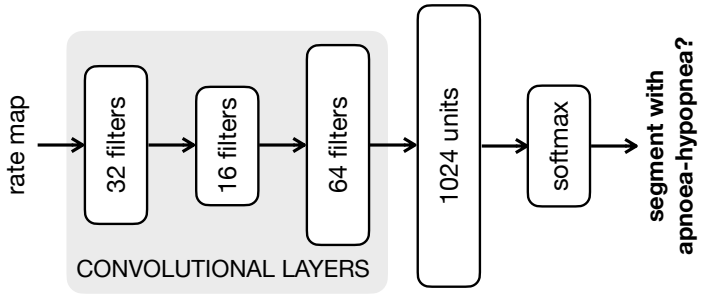


Fig. 5.5 DNN for the classification of sleep audio recording segments

ting. It was developed in TensorFlow (Abadi et al., 2016), and its hyperparameters were chosen heuristically. Several kernel sizes, number of convolutional layers, and number of filters were contemplated while attempting to keep the network small for proper generalisation with the limited amount of data available.

The percentage of segments in a night classified as having apnoea-hypopnea events was used for screening. That is, the screening system output whether a night was negative — below an AHI value — or positive — above an AHI value — based on that percentage. For example, when screening for severe OSA (i.e., AHI $\geq$ 30 events/hour), if the system predicted that a night was negative, the participant very likely had an AHI $<$ 30 events/hour or non-severe OSA. If the night was predicted as positive, the participant very likely had an AHI $\geq$ 30 events/hour or severe OSA. The threshold (i.e., the percentage of apnoea-hypopnea segments) that optimised the screening capability was determined on the validation set with the distance to corner metric (Habibzadeh et al., 2016) calculated on the ROC curve, as this metric allows a balance between sensitivity and specificity. Lastly, sensitivity, specificity, ROC curve, and AUC were evaluated for the screening system, as these directly reflect its diagnostic capability.

5.1.3 Experiments

Given the limited amount of data available, leave-one-participant-out cross-validation was carried out. The screening system was tested on one participant, and validated and trained on data from all the other participants. This was done for each subject in the OSA Sound Corpus: 45 distinct folds were considered. The validation set consisted of data from 5 healthy subjects: AHI $<$ 5 events/hour, and 5 subjects with OSA: AHI $\geq$ 5 events/hour. These were randomly selected. The training set consisted of data from the remaining 34 participants. It was balanced in order to have the same number of segments with apnoea-hypopnea events and segments without these, since most of the data in the OSA Sound Corpus consists of segments with no apnoea-hypopnea events. Various AHI

Table 5.1 Screening for OSA with Rate Maps Using Different Segment Lengths

Segment length	AHI cut-off	5	10	15	20	25	30
	Nights below	16	31	22	43	49	52
	Nights above	44	29	38	17	11	8
15 seconds	**Sensitivity**	0.66	0.62	0.64	0.59	0.73	0.88
	Specificity	0.50	0.48	0.50	0.56	0.73	0.73
	AUC	0.65	0.63	0.65	0.66	0.80	0.90
30 seconds	**Sensitivity**	0.64	0.66	0.64	0.71	0.64	0.75
	Specificity	0.69	0.55	0.50	0.51	0.63	0.69
	AUC	0.65	0.65	0.66	0.68	0.75	0.82
45 seconds	**Sensitivity**	0.68	0.69	0.64	0.65	0.73	0.88
	Specificity	0.62	0.65	0.58	0.56	0.73	0.79
	AUC	0.72	0.73	0.74	0.75	0.81	0.89
60 seconds	**Sensitivity**	0.73	0.66	0.64	0.71	0.64	0.88
	Specificity	0.81	0.74	0.74	0.72	0.80	0.85
	AUC	0.75	0.74	0.72	0.75	0.87	0.94

cut-off points[2] were considered to screen for OSA: 5, 10, 15, 20, 25, and 30 events/hour. Also, different segment lengths were studied to properly capture the temporal pattern of respiration: 15, 30, 45 and 60 seconds. By definition, an apnoea-hypopnea event has a duration of at least 10 seconds (American Academy of Sleep Medicine, 2020). So, a 15-second segment should be able to include an event and some of the context in which it takes place. However, as discussed in Chapter 3, most apnoea-hypopnea events have a duration of 20–30 seconds. For this reason, we examined other segment lengths.

Since most of the related studies have used data recorded in controlled conditions with specialised hardware (e.g., in a sleep laboratory with a tracheal microphone) unlike our data, we used features that directly reflect the physiological effects of OSA and are not affected by background noise — oxygen saturation deltas — and standard acoustic features — log-scaled short-time Fourier transform (logSTFT) — as baseline. It is worth noting that measuring oxygen saturation is obtrusive and requires specialised hardware. Such features were computed for large non-overlapping segments, in the same way as rate maps. Similar to the delta features used in the previous chapter for capturing dynamic acoustic information, oxygen saturation deltas were derived from the oxygen saturation signal from HSAT using a 199.2-ms window (i.e., a 51-sample window with a sampling rate of 256 Hz). Delta features were used, as the variations of this parameter (e.g., desaturations) rather than its instantaneous values reflect the physiological effects of OSA. Similar to rate maps, logSTFT features were computed on non-overlapping 25-ms windows, and normalised over all time-frequency bins to the 0–1 range.

[2]OSA severity based on the AHI was discussed in Chapter 1.

5.1.4 Results and Discussion

Screening for OSA with Rate Maps Using Different Segment Lengths

Table 5.1 presents the results when using 15-second, 30-second, 45-second and 60-second rate maps to screen for OSA at different AHI cut-off points. The number of nights below and above each AHI cut-off point is shown in the table. Looking at sensitivity, specificity and AUC, using 60-second rate maps consistently yielded the best performance at all AHI cut-off points with the exception of the sensitivity at 10 events/hour, the AUC at 15 events/hour, and the sensitivity at 25 events/hour. This evidences that providing enough context is crucial to distinguish segments with apnoea-hypopnea events from those without them, since such events are (nearly) silent and typically have a duration of 20–30 seconds. These events can be nonetheless as short as 10 seconds by definition (American Academy of Sleep Medicine, 2020). 15-, 30- and 45-second segments did not capture the temporal pattern of respiration effectively in all cases, as an apnoea-hypopnea event could make up the whole segment and, therefore, there would not be important variations in the features that could be exploited. For example, the whole segment could consist of silence. Sensitivity, specificity and AUC tended to be higher for the AHI cut-off points of 25 and 30 events/hour, since the nights for the most severe OSA cases consist mainly of segments with apnoea-hypopnea events. So, the percentage values used for screening are more discriminative between severe and non-severe nights. However, it is important to take into account that a limited number of nights from participants with severe OSA were available. Although sensitivity and specificity were very similar for the AHI cut-off of 30 events/hour when using 60-second segments, for example, these indicate that 1 out of 8 severe nights was incorrectly classified as non-severe, whereas 8 out of 52 non-severe nights were incorrectly classified as severe.

Screening for OSA with Different Acoustic Features

After having examined the performance of the screening system when using the same features — rate maps — computed on distinct segment lengths, we investigated its performance when employing different acoustic features extracted from segments with the length that was found to consistently provide enough context. In the following experiments 60-second segments were used. Table 5.2 presents the results when using oxygen saturation deltas (baseline), logSTFT (baseline), rate maps, and bottleneck features to screen for OSA at different AHI cut-off points. As expected, oxygen saturation deltas achieved the best AUC at all AHI cut-off points, as this feature is not affected by background noise and directly measures the physiological effects of OSA. But its measurement requires specialised hardware and is obtrusive.

Table 5.2 Screening for OSA with Different Features

Features	AHI cut-off	5	10	15	20	25	30
	Nights below	16	31	22	43	49	52
	Nights above	44	29	38	17	11	8
Oxygen saturation deltas	**Sensitivity**	0.86	0.76	0.86	0.88	1.00	1.00
	Specificity	0.94	0.68	0.66	0.72	0.69	0.65
	AUC	0.90	0.87	0.90	0.89	0.94	0.94
logSTFT	**Sensitivity**	0.70	0.66	0.68	0.76	0.73	0.62
	Specificity	0.38	0.48	0.47	0.49	0.53	0.63
	AUC	0.55	0.63	0.64	0.70	0.65	0.75
Rate maps	**Sensitivity**	0.73	0.66	0.64	0.71	0.64	0.88
	Specificity	0.81	0.74	0.74	0.72	0.80	0.85
	AUC	0.75	0.74	0.72	0.75	0.87	0.94
Bottleneck features	**Sensitivity**	0.70	0.69	0.68	0.76	0.82	1.00
	Specificity	0.81	0.68	0.61	0.63	0.59	0.65
	AUC	0.83	0.80	0.76	0.80	0.81	0.90

Regarding acoustic features, using rate maps or bottleneck features derived from them resulted in a better performance than using logSTFT. As previously discussed, rate maps provide an expanded low-frequency resolution (Wang and Brown, 2006), which has proven useful in the acoustic analysis of sleep-disordered breathing, while logSTFT has a linear frequency resolution. Looking at AUC, bottleneck features performed better than rate maps for the AHI cut-off points below 25 events/hour, whereas rate maps did better than bottleneck features for the AHI cut-off points of 25 and 30 events/hour. This was caused by the lower specificity obtained using bottleneck features in comparison with rate maps, as more false positives were generated by the former, which made the distinction between severe and non-severe nights difficult.

One downside of screening for OSA with bottleneck features is the use of two different neural networks optimised separately: one for extraction of bottleneck features from rate maps, and the other one for classification based on bottleneck features. This differs from screening with rate maps, which uses a single network that was optimised simultaneously for feature extraction and classification as a whole. The best performance was obtained using rate maps to screen for severe OSA (i.e., AHI $\geq$ 30 events/hour). In fact, a better balance between sensitivity and specificity was allowed by rate maps in comparison with oxygen saturation deltas.

As introduced in Chapter 2, the ROC curve plots the false positive rate (1 – specificity) on the x-axis, and the true positive rate (sensitivity) on the y-axis at different thresholds. Here, these are different percentages of apnoea-hypopnea segments. Random performance is shown as a dashed diagonal line. If the curve is above this line, the screening

test does better than random performance. Otherwise, the test does worse than it, and has no diagnostic capability. The perfect screening test would result in a curve made by two perpendicular lines: $x = 0$ and $y = 1$. The AUC measures the area underneath the ROC curve, and provides an accumulated measure of the performance at all possible thresholds. The perfect screening test would have an AUC of 1.0, whereas a test equivalent to random performance would have an AUC of 0.5.

Figure 5.6 presents the mean ROC curve (black line) along with the standard deviation (grey area) of all AHI cut-off points when using oxygen saturation deltas, logSTFT, rate maps, and bottleneck features to screen for OSA. These give a consolidated visual impression of what was discussed above. All ROC curves were above random performance, which demonstrates that the proposed features and network architecture allowed the DNN to effectively learn from data. Consistent with the previous results, the ROC curve when using logSTFT was the closest to random performance (dashed line). That is, it had the lowest diagnostic capability among all the features considered. The ROC curve when using oxygen saturation deltas was the closest to the upper left corner (i.e., the perfect screening test), as it directly measures the physiological effects of OSA. Lastly, the other ROC curves were very similar. However, using bottleneck features resulted in a more consistent performance across all AHI cut-off points. This was reflected in a narrower grey area around the curve (i.e., a smaller standard deviation) in comparison with rate maps.

5.2 Screening for OSA Integrating Acoustic Features with Physiological Information

Having investigated how far could we get using acoustic evidence alone to screen for OSA, we will now look into the integration of acoustic features with one direct measurement of physiological data — oxygen saturation. As mentioned at the beginning of this chapter, this will be done with the aim of examining whether the inconvenience of wearing one sensor results in an improved performance with respect to employing acoustic evidence on its own. Oxygen desaturations will be used — as these characterise apnoea-hypopnea events by definition (American Academy of Sleep Medicine, 2020) — to lead the classifier to the segments that are likely to contain apnoea-hypopnea events. Following this, the classifier will confirm the presence of such events, since variations in oxygen saturation may also be caused by other factors, for example, body movement during sleep (Al-Mardini et al., 2014; Rofouei et al., 2011).

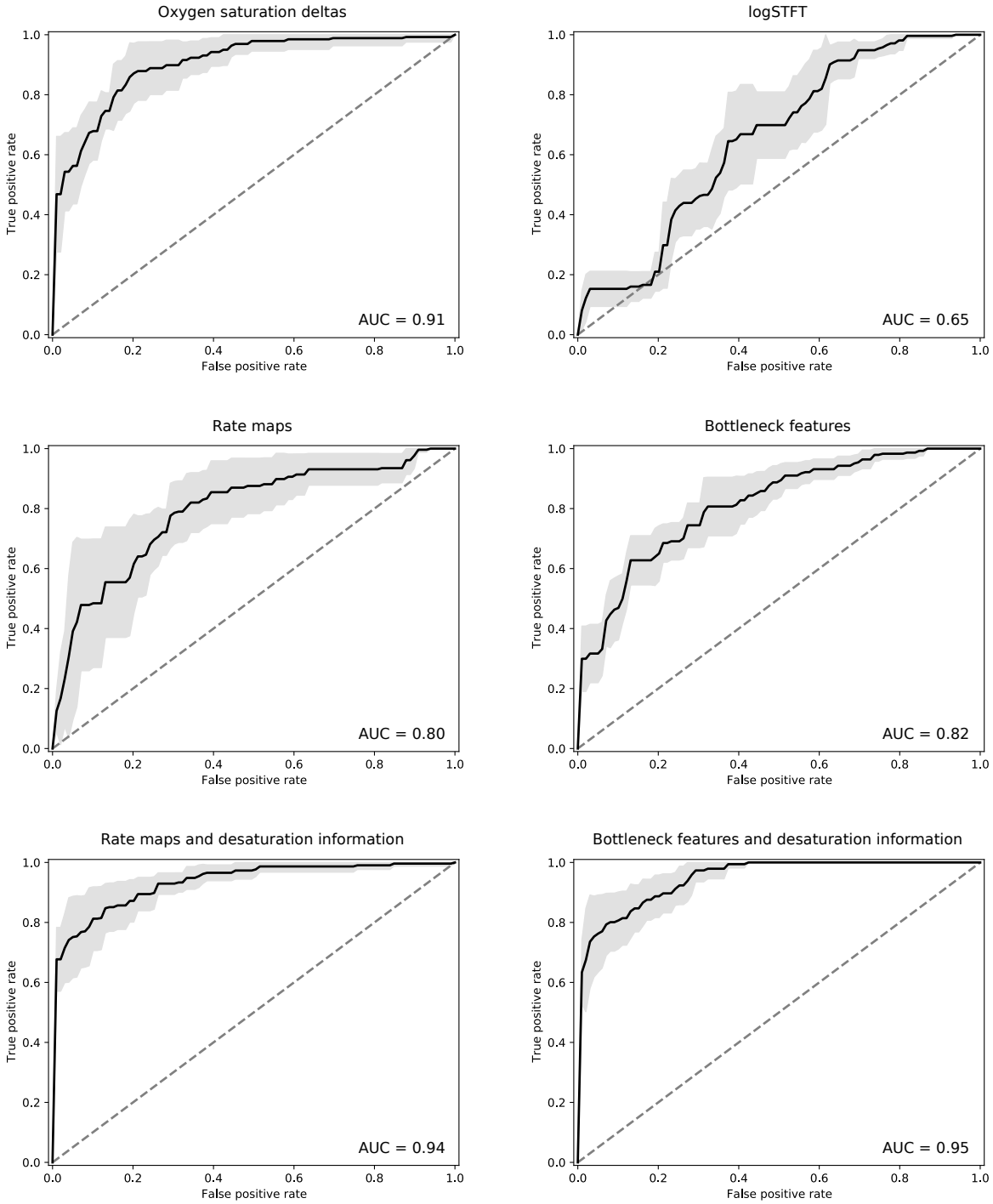


Fig. 5.6 Mean ROC curves for OSA screening using oxygen saturation deltas, logSTFT, rat maps, bottleneck features, rate maps and desaturation information, and bottleneck features and desaturation information

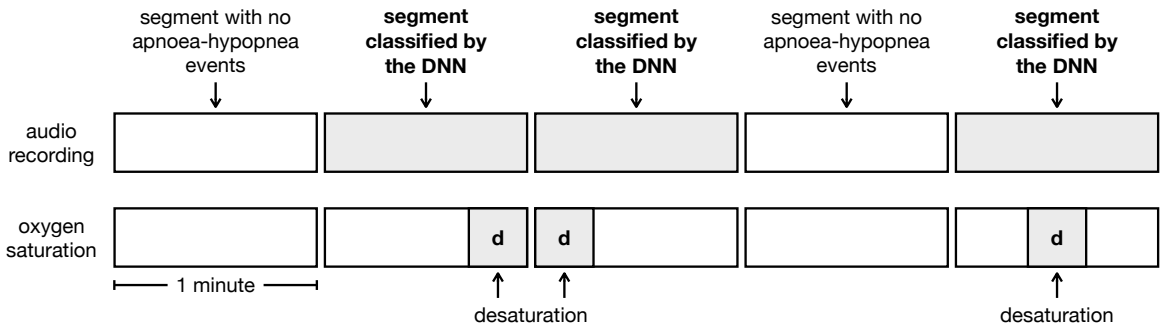


Fig. 5.7 Integration of acoustic features with oxygen saturations to screen for OSA. Only the audio recording segments associated with a desaturation are classified by the DNN. The remaining segments are regarded as 'segments with no apnoea-hypopnea events' without any further processing.

5.2.1 Experiments

To investigate the integration of acoustic features with physiological information, rate maps and bottleneck features were integrated with oxygen desaturation data. It was obtained from the HSAT data, and used for selecting the audio recording segments to be classified by the network: those with desaturations. That is, only the segments associated with a desaturation were classified by the DNN while the remaining segments were regarded as 'segments with no apnoea-hypopnea events' without any further processing. This is illustrated in Figure 5.7. The amount of segments that were classified by the network depended on the number of desaturations. As before, 60-second segments were provided as input to the DNN (Figure 5.5), leave-one-participant-out cross-validation was performed, and the same AHI cut-off points were considered to screen for OSA based on the percentage of predicted segments with apnoea-hypopnea events. Also, sensitivity, specificity, ROC curve, and AUC were evaluated.

The sample entropy method by Castillo-Escario et al. (2019) was implemented as a baseline. Using an approach that was similar to what was described in Section 3.2, Castillo-Escario et al. collected audio recordings with an Android smartphone from 13 participants — 8 men and 5 women — during HSAT with a ResMed ApneaLink Air. However, the smartphone was placed on the subject's chest, and the audio recordings were manually synchronised with the HSAT data. Their rule-based approach detects apnoeas and hypopneas by finding silent regions with a duration of at least 6 seconds in audio recording segments close to a desaturation. Desaturations are obtained from the HSAT data, and silent regions are detected using sample entropy: a measure of the predictability or complexity of a signal. For instance, silent regions will have a low sample entropy value (e.g., close to 0), as there is great similarity between silence frames, whereas snoring regions will have a high sample entropy value (e.g., close to 1) given the high variabil-

Table 5.3 Screening for OSA Integrating Oxygen Desaturations

Features	AHI cut-off	5	10	15	20	25	30
	Nights below	16	31	22	43	49	52
	Nights above	44	29	38	17	11	8
Sample entropy	**Sensitivity**	0.52	0.41	0.32	0.35	0.55	0.75
	Specificity	1.00	1.00	1.00	1.00	1.00	1.00
Rate maps	**Sensitivity**	0.77	0.90	0.91	0.94	0.91	1.00
	Specificity	0.81	0.74	0.74	0.67	0.84	0.96
	AUC	0.90	0.91	0.93	0.92	0.96	1.00
Bottleneck	**Sensitivity**	0.84	0.93	0.86	0.94	0.82	1.00
	Specificity	0.88	0.81	0.76	0.72	0.84	0.96
	AUC	0.96	0.94	0.93	0.93	0.94	1.00

ity between snoring frames. The number of silent regions and the duration of the audio recording are used for directly estimating the AHI as:

$$\mathrm{AHI}_{\mathrm{estimated}} = \frac{\text{number of silent regions}}{\text{audio recording duration}}$$

5.2.2 Results and Discussion

Table 5.3 presents the results when using sample entropy (baseline), rate maps, and bottleneck features integrated with oxygen desaturations to screen for OSA at different AHI cut-off points. AUC is not reported for the sample entropy approach, as it directly calculates the AHI and screening is based on this index. A specificity of 1.00 was reported for the baseline at all AHI cut-off points, since it avoids false positives by design: only the silent regions close to a desaturation are considered (Castillo-Escario et al., 2019). But its sensitivity was low. The differences in data collection between our study and the one by Castillo-Escario et al. probably contributed to this. The audio recordings in the OSA Sound Corpus were not carried out with a smartphone placed on the subject's chest as they did. So, our audio recordings are less sensitive to quiet breathing than theirs, and using sample entropy for silence detection is not very suitable for our data. In other words, an OSA screening system simply based on detecting silence might not be robust enough.

As hypothesised, incorporating physiological information improved the screening performance when compared to using acoustic features alone. Considering rate maps, the main improvement was a reduction in the number of false negatives reflected in a better sensitivity, since oxygen desaturation information led the classifier to segments that very likely contain apnoea-hypopnea events. Considering bottleneck features, sensitivity, specificity and AUC were improved at all AHI cut-off points. AUCs were above

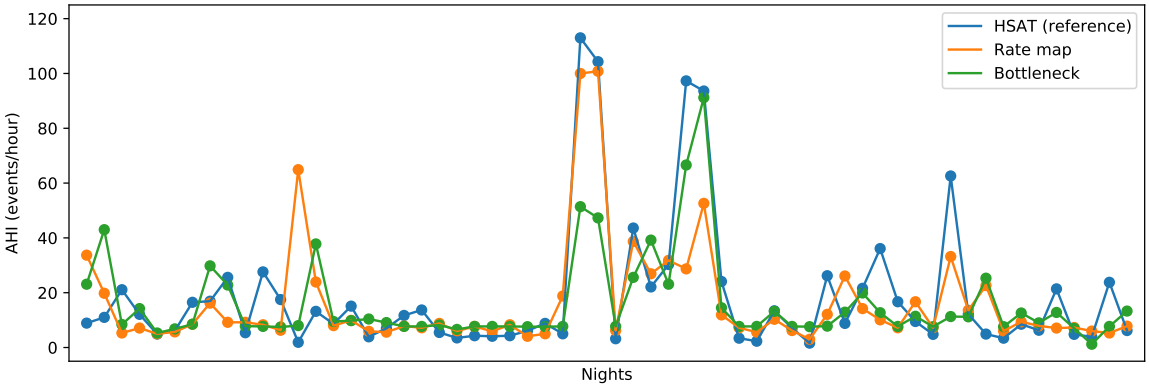


Fig. 5.8 AHI estimation using acoustic features

0.93, which evidences the encouraging screening capability of the proposed system. However, it is important to bear in mind that oxygen desaturations were obtained from HSAT, which limits the application of the screening system, as specialised hardware is required. An alternative might be using readily available hardware capable of measuring oxygen saturation, like a smartwatch (Phillips et al., 2019), in addition to a smartphone.

Integrating acoustics with oxygen saturation information also resulted in a better performance than using this physiological parameter on its own (Table 5.2), since factors different from apnoeas and hypopneas may cause desaturations as well. The results suggest that the acoustic evidence contributed to the differentiation between segments associated with a desaturation caused by apnoea-hypopnea events from those caused by body movement or other factors.

Figure 5.6 displays the mean ROC curve along with the standard deviation of all AHI cut-off points when integrating oxygen desaturations with rate maps and bottleneck features to screen for OSA. The ROC curve is not displayed for the sample entropy approach, as it directly computes the AHI. Both curves provide a synthesised visual intuition of the previously discussed results. These were the closest to the upper left corner (i.e., the perfect screening test) of all the different features and configurations considered. Integrating acoustic features with physiological information also resulted in a consistent performance across all AHI cut-off points evidenced by a small standard deviation or a narrow grey area around the ROC curves.

5.3 AHI Estimation

As discussed in Chapter 1, it is normal clinical practice to diagnose OSA based on the AHI. It is calculated from the number of apnoea-hypopnea events in a night and the time spent sleeping. Therefore, sleep status or sleep stages are required for its calculation. Related

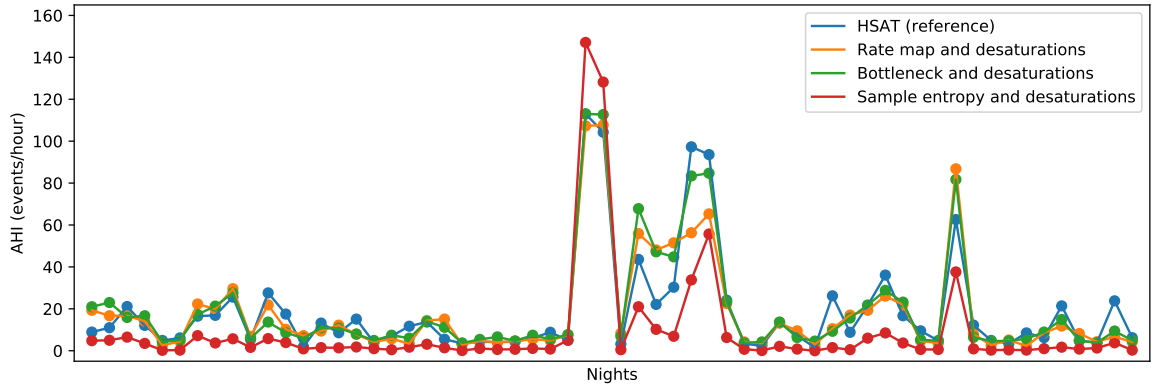


Fig. 5.9 AHI estimation using acoustic features and desaturation information

studies — like the one by Castillo-Escario et al. (2019) — have estimated the AHI from sleep audio recordings using the duration of the recordings instead of the time spent sleeping, as this information requires specialised hardware (e.g., an electroencephalograph), which they were (and we are) trying to avoid.

Since our screening system neither detects individual apnoea-hypopnea events nor considers sleep status, direct AHI calculation from the output of the DNN was not possible. For this reason, we estimated the AHI from the percentage of predicted segments with apnoea-hypopnea events in a night using regression. A linear regression model was optimised on the validation set, and used for mapping the percentage of predicted segments with apnoea-hypopnea events in a night to an AHI value. Specifically, a polynomial of degree 3 was fitted by minimising the squared error. The same features and configurations considered to screen for OSA were examined for AHI estimation.

5.3.1 Results and Discussion

The performance on the AHI estimation task was evaluated with Bland-Altman plots, which are commonly used in medical literature to analyse the agreement between two different measurement techniques (Bland and Altman, 1986). Here, these plot for each night the average of the estimated and reference AHIs against the difference between the estimated and reference AHIs as a scatter diagram. The mean difference between all reference and estimated AHIs is displayed as a solid line. The agreement limits, ±1.96 times the standard deviation of the difference between all reference and estimated AHIs, are shown as dashed lines.

Before considering Bland-Altman plots, Figure 5.8 shows the reference AHI from HSAT along with the estimated AHIs using rate maps and bottleneck features for each night in the OSA Sound Corpus. It can be seen that, in general, the reference AHI pattern was followed by the estimated AHIs. Although the AHI for some nights was incor-

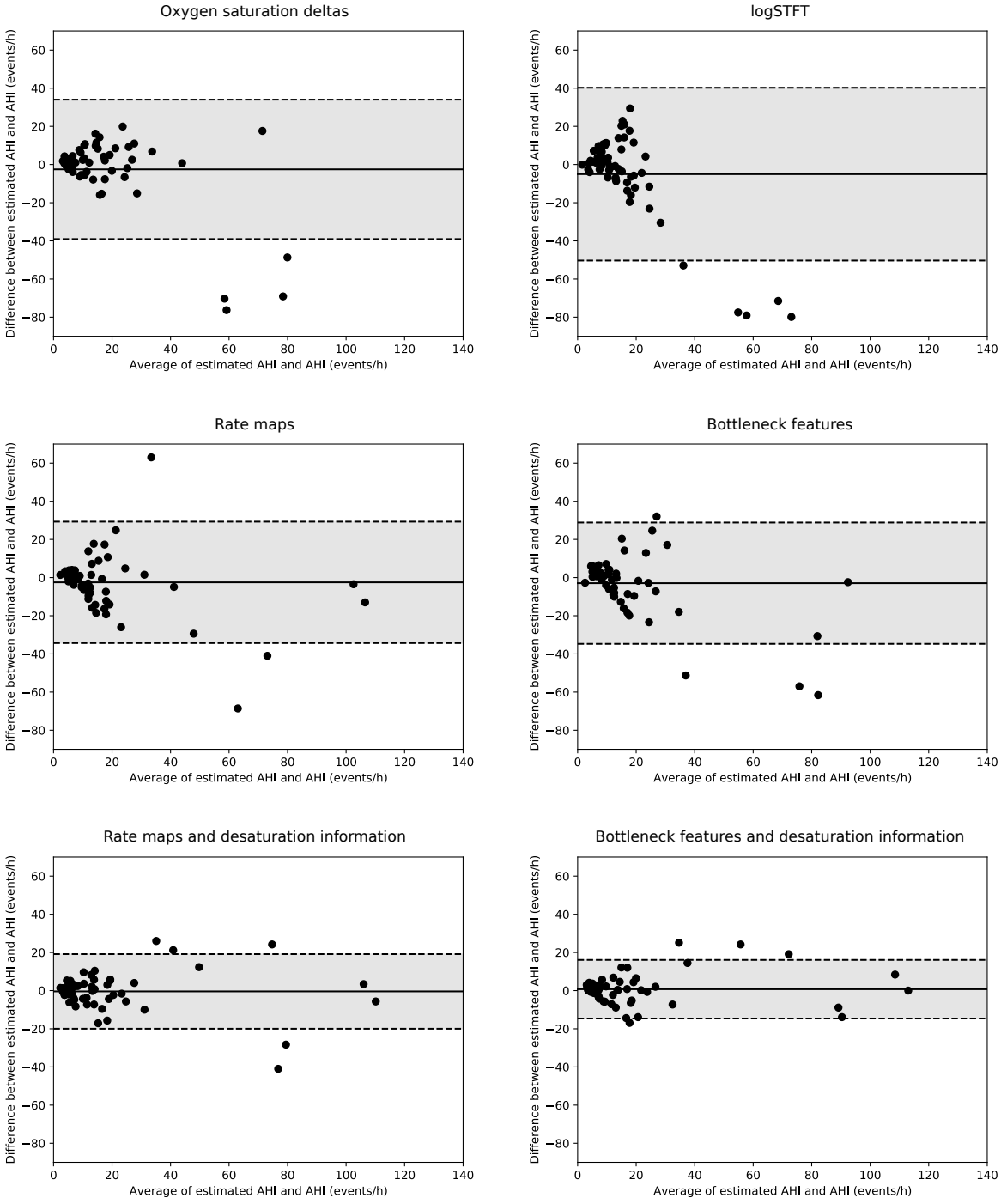


Fig. 5.10 Bland-Altman plots for AHI estimation using oxygen saturation deltas, logSTFT, rat maps, bottleneck features, rate maps and desaturation information, and bottleneck features and desaturation information

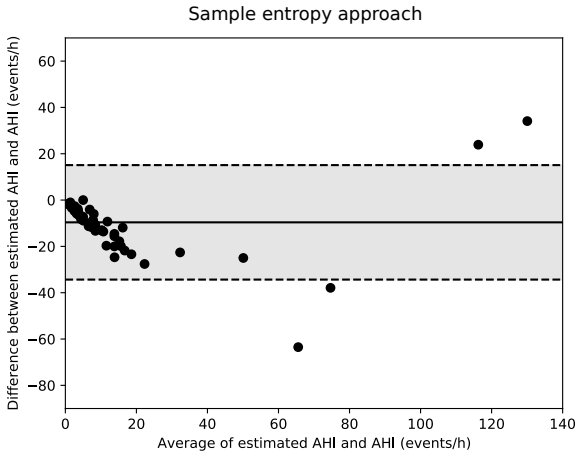


Fig. 5.11 Bland-Altman plot for AHI estimation with the sample entropy approach

rectly estimated, the performance — as will be reflected in the Bland-Altman plots — was reasonably good taking into account that acoustic evidence alone was used for estimating the AHI, which otherwise would be calculated from the manual scoring of HSAT or polysomnography data.

Likewise, Figure 5.9 presents the reference AHI along with the estimated AHIs integrating physiological information with acoustic features: sample entropy (baseline), rate maps, and bottleneck features. Oxygen desaturations directed the classifier to likely apnoea-hypopnea segments, and it confirmed the presence of such events, since desaturations can be caused by other factors as well. In this case, the reference AHI pattern was more closely followed by the estimated AHIs with the exception of those from sample entropy. Consistent with the screening results, the integration of acoustic features with desaturations resulted in the best performance on the AHI estimation task. This will be more clearly evidenced in the Bland-Altman plots.

Figure 5.10 shows the Bland-Altman plots for AHI estimation using oxygen saturation deltas (baseline), logSTFT (baseline), rate maps, and bottleneck features. The nights outside the agreement limits were from participants with severe OSA, since the screening system tends to underestimate the AHI on average: the mean differences (solid lines) were below zero. Related studies have reported this behaviour. For instance, Nakano et al. (2019) attributed it to using the total recording time, instead of the total sleep time, to calculate the AHI. The total recording time is always longer than the total sleep time, which results in consistently smaller AHI values in comparison with the reference AHI. In our screening system, this was probably caused by the use of large analysis windows. For those participants with very high AHIs (e.g., AHI > 100 events/hour), 60-second segments may contain more than one apnoea-hypopnea event, but such segments do not make a difference in the percentage of predicted segments with apnoea-hypopnea events,

Table 5.4 Classification of 9.6 hours of precomputed rate maps with a DNN

Mobile Device	Processing Time
Apple iPhone SE (2020)	4 seconds
Apple iPhone XS	6 seconds
Apple iPhone 7 Plus	100 seconds
Apple iPhone 6	313 seconds

which we used for AHI estimation. That is, a segment with multiple apnoea-hypopnea events contributes the same as a segment with only one apnoea-hypopnea event. Any AHI ≥ 30 events/hour is nonetheless regarded as severe.

The Bland-Altman plots for AHI estimation integrating physiological information with rate maps, and bottleneck features are presented in Figure 5.10. Figure 5.11 displays the Bland-Altman plot for AHI estimation with the sample entropy approach (baseline). Consistent with the performance on OSA screening, the best agreement between the reference and estimated AHIs was obtained using bottleneck features along with desaturation information. The narrowest agreement limits along with a mean difference very close to zero were obtained with this configuration, as desaturations led the classifier to segments that likely contained apnoea-hypopnea events. Similar performance was obtained when using rate maps along with desaturations. The sample entropy approach consistently underestimated most AHIs with the exception of those from two participants with considerably high AHIs (i.e., > 110 events/hour), which it overestimated. It was reflected in a mean difference far from zero. This was likely caused by the differences in data collection between the study by Castillo-Escario et al. (2019) and ours: they collected sleep audio recordings with a smartphone attached to the participant's chest, which made their recordings more sensitive to quiet breathing than ours. So, longer and less frequent silent regions were detected on average in our data with their approach.

5.4 Mobile Deployment

Since improving access to diagnosis is one of the main challenges that sleep medicine is still facing (Dafna et al., 2013; Phillips, 2007), novel approaches to screening for sleep-disordered breathing should ideally be robust enough to be applied in typical sleep conditions, and make use of readily available hardware — for example, smartphones. Although the sleep audio recordings in the OSA Sound Corpus were made with smartphones, the experiments presented in this chapter have been performed on a computer server. With the aim of investigating the feasibility of on-device processing (Kaissis et al., 2020), the proposed DNN was additionally deployed on smartphones.

The DNN model developed in TensorFlow was converted to a TensorFlow Lite model, and run on different iOS smartphones. TensorFlow Lite is a TensorFlow version optimised for mobile devices. The TensorFlow Lite model had a size of 22.9 MB, a very reasonable size considering the usually limited storage capacity of mobile devices. The measured processing times are presented in Table 5.4. Using precomputed rate maps, 576 60-second segments amounting to 9.6 hours of data were classified by the neural network in 4 seconds on an Apple iPhone SE (2020), in 6 seconds on an Apple iPhone XS, in 100 seconds on an Apple iPhone 7 Plus, and in 313 seconds on an Apple iPhone 6. Better than real-time performance was possible thanks to the small neural network size (around 6 million parameters), and the neural processing units and machine learning accelerators commonly found in relatively recent smartphones. This suggests that it is possible to run the entire screening system on a mobile device: audio recording, feature extraction, classification, results reporting, etc. However, the feature extraction step — which on a computer server takes approximately 5 seconds for each 60-second audio recording segment — was not investigated on mobile devices, and might considerably increase the processing time.

Although the DNN was not deployed on Android smartphones, the very same TensorFlow Lite model can be used both in iOS and Android apps. It is worth noting that the data collected to develop the screening system included recordings from both iOS and Android devices, and the DNN models were trained independently of the operating system.

5.5 Summary

In this chapter, the temporal pattern of respiration was exploited to screen for OSA from sleep audio recordings. The OSA Sound Corpus introduced in Chapter 3 was used to train and evaluate the screening system. The presence or absence of apnoea-hypopnea events in 1-minute audio recording segments was predicted with a DNN from acoustic features — rate maps and bottleneck features — and screening was based on the percentage of predicted segments with apnoea-hypopnea events in a night. 1-minute segments were used for effectively capturing the temporal pattern of respiration, since apnoea-hypopnea events usually last for 20–30 seconds. Each 1-minute segment was labeled as either having apnoea-hypopnea events in it or not. Various AHI cut-off points were considered to screen for OSA: 5, 10, 15, 20, 25, and 30 events/hour. That is, the screening system output whether a night was positive — above an AHI value — or negative — below an AHI value.

Bottleneck features performed better than rate maps for the AHI cut-off points below 25 events/hour, whereas rate maps did better than bottleneck features for the AHI cut-off points of 25 and 30 events/hour. This was caused by the lower specificity obtained using

bottleneck features in comparison with rate maps, since more false positives were generated by the former, which made the distinction between severe and non-severe nights difficult. The best performance was obtained using rate maps to screen for severe OSA (i.e., AHI $\geq$ 30 events/hour): a sensitivity of 0.88, a specificity of 0.85, and an AUC of 0.94 were reported. However, using bottleneck features resulted in a more consistent performance across all AHI cut-off points. Given that the developed system can be conveniently applied at home and does not require specialised hardware, OSA screening might be based on data from multiple nights, which would potentially provide a more representative assessment of the condition. Following diagnosis, the proposed system might be also employed for long-term monitoring of the progression of OSA. This would be useful to evaluate treatment efficacy or the impact of lifestyle changes (e.g., weight loss) on the disorder.

With the aim of further assessing the capability of the screening system using acoustic evidence on its own, we investigated the integration of acoustic features with physiological information. Rate maps and bottleneck features were integrated with oxygen desaturations. These were obtained from the HSAT data, and used for selecting the audio recording segments to be classified by the network: those with desaturations. The sample entropy method by Castillo-Escario et al. (2019) was implemented as a baseline. Their rule-based approach detects apnoeas and hypopneas by finding silent regions with a duration of at least 6 seconds in audio recording segments close to a desaturation using sample entropy. The number of silent regions and the duration of the audio recording are used for directly calculating the AHI.

Incorporating physiological information improved the screening performance compared to using acoustic features alone. Regarding rate maps, the main improvement was a reduction in the number of false negatives, since oxygen desaturation information led the classifier to segments that likely contain apnoea-hypopnea events. Considering bottleneck features, AUCs were above 0.93 at all AHI cut-off points, which evidences the encouraging screening capability of the proposed system. However, it is important to take into account that oxygen desaturations were obtained from HSAT. This is a limitation of the present study, as specialised hardware was required. The baseline achieved a specificity of 1.00 at all AHI cut-off points, since it avoid false positives by design. But its sensitivity was low. The differences in data collection between our study and the one by Castillo-Escario et al. (2019) probably contributed to this. They collected sleep audio recordings with a smartphone attached to the participant's chest, which made their recordings more sensitive to quiet breathing than ours. Our approach highlights the importance of learning from data — rather than producing rule-based methods — to develop robust systems. The high variability of breathing sounds during sleep, room acous-

tic characteristics, and microphone responses hinder proper generalisation of rule-based approaches.

AHI estimation was investigated as well. We estimated the AHI from the percentage of predicted segments with apnoea-hypopnea events in a night using a linear regression model. The performance was evaluated with Bland-Altman plots. Consistent with the performance on OSA screening, the best agreement between the estimated and reference AHIs was obtained using bottleneck features along with desaturation information. A mean difference very close to zero, and the narrowest agreement limits were obtained with this configuration.

Lastly, the proposed DNN was deployed on smartphones with the aim of investigating the feasibility of on-device processing. Better than real-time performance was achieved thanks to the small neural network size, and the neural processing units commonly found in relatively recent smartphones. On-device machine learning preserves user privacy, since the user's data strictly remains on their device. It might play a role in reducing the psychological effects of being under observation during a sleep test, and contribute to improving access to sleep-disordered breathing diagnosis, which has been negatively affected by the COVID-19 pandemic (Voulgaris et al., 2020; Williamson, 2020).

Future work might include collecting more data from participants with moderate or severe OSA to balance the amount of data for each OSA severity group in the OSA Sound Corpus. The feasibility of measuring oxygen saturation with readily available hardware might also be investigated in the future. For example, it could be measured with a smartwatch to avoid the need for specialised hardware that currently limits the application of the proposed screening system.

Chapter 6

Robustness to Bed Partner Breathing Sounds

"Sleep is a skilled magician. It changes the proportions of things, the distances between them. It separates people and they are lying next to each other, brings them together and they can barely see one another."
– J. Saramago

Up to this point, we have developed systems to screen for sleep-disordered breathing that assume the subject under study is sleeping on their own. However, typical sleep conditions — a bedroom in the home — commonly include a bed partner, who can negatively impact upon the performance of such systems. From a historical perspective, human beings have slept together because of fear of the dark. The American historian A. R. Ekirch (1950–) in his book 'At Day's Close: Night in Times Past' writes:

> Night, man's first necessary evil, inspired widespread fear before the Industrial Revolution... Never did families feel more vulnerable than when they retired at night. Bedmates afforded a strong sense of security, given the prevalence of perils, real and imagined — from thieves and arsonists to ghosts, witches, and the prince of darkness himself. (Ekirch, 2006)

Co-sleeping can indeed improve our perceived physical and emotional security, which results in enhanced sleep quality and decreased arousal levels. However, from a healthcare perspective, sleep is mainly considered as an individual phenomenon, and the complications arising from the presence of a bed partner remain an overlooked topic. Co-sleeping should be taken into account when dealing with conditions like snoring or OSA, as the sleep problems of one subject may be a problem for the other as well. For example, those who sleep with a snorer are typically affected by increased sleep fragmentation and decreased sleep quality. On the other hand, treatment of sleep conditions can positively

impact both the subject with these and their bed partner. For instance, the use of continuous positive airway pressure (CPAP), the main treatment for OSA, has been found to improve the quality of life of both the individual with OSA and their bed partner (Richter et al., 2006).

Ideally, systems to screen for sleep-disordered breathing should be robust to bed partner's breathing sounds, so they can be applied in typical sleep conditions. Furthermore, it would be useful to screen a participant and their bed partner for sleep-disordered breathing in the same session. Two different approaches to this are considered in the present chapter. We first investigate the use of dual audio recordings for source separation. In the second approach we examine a deep learning alternative to source separation that uses single-channel audio recordings: *snorer* diarisation. Some of the work presented in this chapter has previously appeared in (Romero et al., 2020b).

This chapter is organised into four sections. We first lay out the general problem of sound event detection in multisource environments in Section 6.1. Then the use of multichannel sleep audio recordings for source separation is examined in Section 6.2, whereas Section 6.3 investigates a single-channel approach that draws inspiration from speaker diarisation research. Finally, Section 6.4 gives a summary of the chapter.

6.1 Sound Event Detection in Multisource Environments

Distinguishing individual sound sources from a complex mixture of sounds is a challenging problem, as they commonly overlap in time and frequency, which hinders the isolation and characterisation of individual sources, for instance, differentiating each snorer's breathing sounds from the mixture of breathing sounds of both. One possible approach to this problem is using multichannel audio recordings to separate individual sources, and then detect sound events in the separated signals. Methods that follow this approach include different forms of beamforming (Anguera et al., 2007; Feng and Jones, 2006; Lockwood et al., 2004), and blind source separation (Yatabe, 2020; Yu et al., 2014), for example. An alternative approach is making use of single-channel audio recordings to separate individual sound sources or to directly detect sound events in the mixture without separating individual signals. We will now consider some of these techniques, and their suitability for screening a participant and their bed partner for sleep-disordered breathing in the same session. As illustrated in Figure 6.1, this is envisioned as an easy-to-use method that only requires readily available hardware, for example, two smartphones — one placed at each side of the bed on the bedside table.

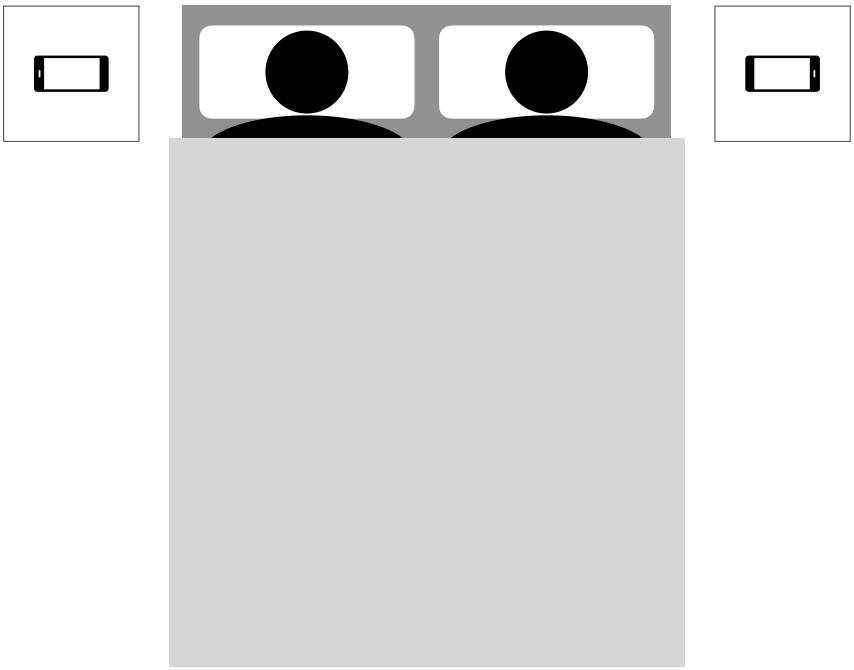


Fig. 6.1 Expected arrangement of sound sources (snorers) and sensors (smartphone microphones)

Beamforming

Beamforming is the process of using multiple synchronised microphones (i.e., ≥ 2 sensors) arranged in a known geometrical pattern to extract sound sources from a particular direction in a multisource environment. Specifically, an audio signal from a particular direction is enhanced by delaying the signals on each microphone to constructively sum them as follows:

$$y(n) = \sum_{m=1}^{M} W_m(n)\, x_m(n - \mathrm{TDOA}_{(m,\mathrm{ref})}(n)) \tag{6.1}$$

where M is the number of microphones, $W_m(n)$ is the weight for microphone m at instant n, $x_m(n)$ is the signal on microphone m, $\mathrm{TDOA}_{(m,\mathrm{ref})}$ is the time difference of arrival (TDOA) between microphone m and the reference microphone 'ref' that aligns all signals at instant n, and $y(n)$ is the enhanced signal. The TDOA is the result of sound reaching the microphones at a slightly different time, since sound travels a different distance to arrive at each microphone. As we will see later in this chapter, $\mathrm{TDOA}_{(m,\mathrm{ref})}$ is estimated with cross-correlation techniques, such as the generalised crosscorrelation with phase transform (GCC-PHAT). Sound sources are located by beamforming in every direction and searching for energy peaks in the output signal $y(n)$ (Anguera et al., 2007; Feng and Jones, 2006). One of the limitations of beamforming is the need for specialised hardware: an array of

physically connected microphones, as their arrangement is required to be known and they must be synchronised to effectively exploit the TDOAs.

Beamforming techniques assume that the different sound sources are located in front of the microphone array. This might limit their applicability for screening a participant and their bed partner for sleep-disordered breathing in the same session, since the expected arrangement (Figure 6.1) would result in the sound sources (i.e., snorers) being in the middle of the microphones rather than in front of them. We will investigate this in the next section.

Blind Source Separation

Blind source separation is another technique for separating individual signals using multichannel audio recordings. It assumes that each microphone receives different combinations of the sound sources, and the observed mixed signal is a series of microphone outputs. This can be mathematically defined as:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{6.2}$$

where $\mathbf{x}$ is the observed mixed signal: a length-M vector of the signals on the M microphones, $\mathbf{s}$ are the unobserved source signals: a length-N vector of the N source signals, and $\mathbf{A}$ is the mixing matrix: a $M \times N$ matrix that represents the gain of each source at each microphone. The objective of blind source separation is to recover the unobserved source signals — hence, 'blind' — from the observed set of mixtures without knowing the mixing process. It requires at least as many microphones as sources: $M \geq N$, and $\mathbf{A}$ to be an invertible matrix. Then the task consists in finding a demixing matrix that is the inverse of the mixing matrix $\mathbf{A}$ to recover the source signals $\mathbf{s}$. This can be expressed as:

$$\mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} \approx \mathbf{s} \tag{6.3}$$

where $\mathbf{W}$ is the $N \times M$ demixing matrix (Yatabe, 2020; Yu et al., 2014). Independent component analysis (ICA), for instance, is one method for finding such a matrix by exploiting statistical characteristics of the mixed signal (Comon, 1994). Screening a subject and their bed partner for sleep-disordered breathing in the same session using blind source separation would require at least two microphones. However, some studies have proposed blind source separation approaches using fewer microphones than sound sources (Olsson and Hansen, 2006).

Single-channel Speaker Diarisation

One could also make use of single-channel audio recordings to segregate individual sound sources. Although obtaining spatial information (e.g., TDOAs) commonly requires at least two microphones, some studies have proposed the use of other acoustic cues derived from single-channel audio recordings. For instance, Hu et al. (2015) investigated the use of direct-to-reverberant ratio (DRR) to obtain information about the location of speakers from single-channel audio recordings. Specifically, they used DRR for speaker diarisation, which aims to identify who spoke when in multi-speaker audio recordings. DRR was defined as:

$$\mathrm{DRR}_i = 10\log_{10}\left(\frac{\sum_{n=n_d-n_0}^{n_d+n_0} h_i^2(n)}{\sum_{n=0}^{n_d-n_0} h_i^2(n) + \sum_{n=n_d+n_0}^{\infty} h_i^2(n)}\right) \tag{6.4}$$

where DRR_i is the spatial feature for speaker i, n_0 is the number of samples in the analysis window (e.g., 8-ms window), n_d is the time index of the direct-path arrival, and $h_i^2(n)$ is the room impulse response. A bidirectional LSTM network was trained to estimate the DRR from time and frequency features (e.g., zero-crossing rate, pitch period, MFCCs, etc.) with good performance on simulated meeting data. In the proposed method, speakers were assumed to be stationary, which might limit its application to *snorer* diarisation, as snorers move throughout the night. Extrapolating the concept of speaker diarisation to snorer diarisation using single-channel audio recordings in a deep learning framework is nonetheless a promising direction for research that will be considered later in this chapter.

6.2 Dual Audio Recordings for Source Separation

A possible approach to screening a participant and their bed partner for sleep-disordered breathing using acoustics is separating each subject's breathing sounds and running our systems presented in Chapters 4 and 5 on the separated signals. As discussed in the previous section, many methods for separating one sound source from a mixture of signals — like blind source separation and beamforming — require multiple sensors or microphones (Thiemann and Vincent, 2013). However, the data collected for the classification of sleep-disordered breathing events (Chapter 4) and screening for OSA (Chapter 5) consists of single-channel audio recordings collected using a smartphone at home from participants sleeping on their own. For this reason, the required data to examine multichannel signal separation approaches for screening a participant and their bed partner for sleep-

disordered breathing had to be simulated using the available data. This will be described next.

6.2.1 Collection of Simulated Data

Since improving accessibility to diagnosis is one of the main challenges that sleep medicine is still facing (Dafna et al., 2013; Phillips, 2007), we used readily available hardware — smartphones — to collect simulated data with characteristics that were as close as possible to the expected use (Figure 6.1). Pairs of audio recordings from the Snoring Sound Corpus were played on loudspeakers placed on a bed at home. Each loudspeaker was placed on a pillow. The playback was recorded simultaneously with two smartphones, one at each side of the bed. A bespoke iOS app was developed to perform dual audio recordings. This simultaneously runs on two devices paired via Bluetooth and works as follows:

1. The app is opened in both devices. One is set to 'Me' mode and the other, to 'Bed Partner' mode.
2. The devices are placed by the corresponding side of the bed at head level, for example, on bedside tables.
3. From the device in 'Bed Partner' mode the 'Scan and Connect' button is tapped to automatically scan for an available device running the app in 'Me' mode and connect to it. A message informs when both devices are paired.
4. From the device in 'Me' mode, the 'Record' button is tapped to start recording audio on both devices.
5. Three 500-ms white noise bursts spaced 500 ms are played by the device in 'Bed Partner' mode to allow the synchronisation of the audio recordings.
6. A pair of audio recordings from the Snoring Sound Corpus is played on the loudspeakers. Each loudspeaker plays an audio recording from a different snorer.

Before investigating strategies for multichannel source separation, we will first examine the synchronisation between the audio recordings made on different devices, as this is crucial for such strategies.

6.2.2 Synchronisation Between Audio Recordings

Simulated data — sleep audio recordings played on loudspeakers — was collected with the developed app. Five pairs of 2-minute sleep audio recordings were randomly selected

from the Snoring Sound Corpus. Each pair consisted of data from two different snorers, and was played as a stereo file: the left snorer on the left loudspeaker, and the right snorer on the right loudspeaker. This playback was recorded simultaneously on two paired Apple iPhones running the custom app.

Initial analysis showed that accurate synchronisation between audio recordings made on different devices, and source separation from this kind of recordings are challenging tasks. To begin with, the measured TDOAs — those caused by the sound signal reaching each recording device at slightly different times due to their distinct spatial locations — did not always match the theoretical values. Theoretical TDOAs do not take into account the internal delay of the recording devices (Gaubitch et al., 2013), and the sampling frequency mismatch between them (Miyabe et al., 2015; Ochi et al., 2016). Also, reverberation should be considered, as it may lead to a wrong estimation of the TDOAs.

Dealing with the internal delay of the recording devices, and their sampling frequency mismatch were beyond the scope of this study, as app development was a useful and necessary tool (e.g., data collection) rather than a fundamental component of it. An estimation of TDOAs closer to the expected theoretical values was achieved to a limited extent by applying pre-emphasis, and estimating them with the generalised crosscorrelation with phase transform (GCC-PHAT) (Knapp and Carter, 1976) instead of the standard crosscorrelation. Pre-emphasis (Equation A.1) attenuates the lower frequencies, where the estimation of TDOAs is less reliable because of the presence of low-frequency noise (Himawan et al., 2011). GCC-PHAT is commonly used to estimate TDOAs from multichannel audio recordings, as it has been proven to de-emphasise the time lags caused by noise and reverberation, and emphasise the peak in the crosscorrelation function observed at the time lag that corresponds to the true source location (Adavanne et al., 2016). GCC-PHAT is computed as follows:

$$\hat{G}_{ij}(f) = \frac{x_i(f)x_j^*(f)}{|x_i(f)x_j^*(f)|} \tag{6.5}$$

where $x_i(f)$ is the Fourier transform of the audio signal on microphone i, and $x_j^*(f)$ is the complex conjugate of the Fourier transform of the audio signal on microphone j. A global maximum is observed in the crosscorrelation function at the TDOA between microphones i and j:

$$\tau(i,j) = \underset{\tau}{\operatorname{argmax}} \left(\hat{R}_{ij}(\tau) \right) \tag{6.6}$$

where $\hat{R}_{ij}(\tau)$ is the inverse Fourier transform of $\hat{G}_{ij}(f)$ (Himawan et al., 2011). The synchronisation experiments will be reported next. These were repeated multiple times to ensure the consistency of the results, different pairs of 2-minute sleep audio recordings

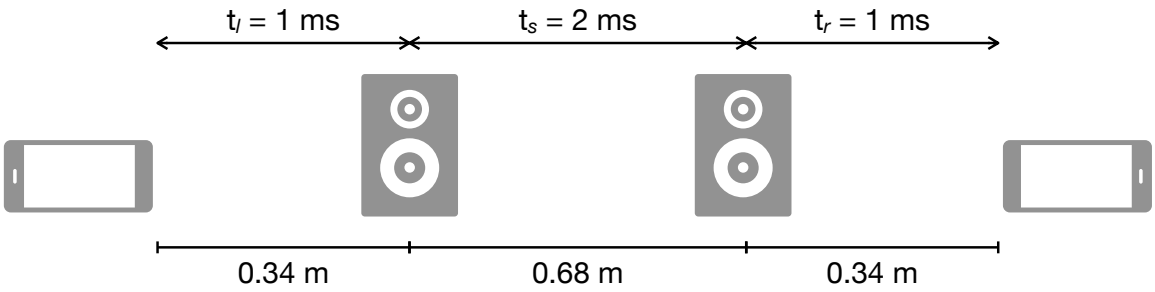


Fig. 6.2 Dual audio recording configuration

were used, and the TDOAs were measured throughout the 2 minutes, as will be explained below.

Configuration 1: phones at each side of the bed, and initial noise bursts played from one phone

The configuration in which dual audio recordings took place is depicted in Figure 6.2. The distances approximately correspond to those of two individuals (loudspeakers) sleeping on a double bed with their phones placed at each side of the bed. The loudspeakers were calibrated to match their output levels. To obtain a reference, each snorer was recorded on their own first. Then, both snorers were recorded together. This was done for five pairs of 2-minute sleep audio recordings.

Before measuring the actual TDOAs, we will first calculate the expected values using the nomenclature employed in Figure 6.2. Let t_l be the time of arrival from the left loudspeaker to the left phone; t_s, the time of arrival between both loudspeakers; and t_r, the time of arrival from the right loudspeaker to the right phone. A signal from the left loudspeaker will arrive at the left microphone after time t_l, and at the right microphone after time $t_s + t_r$. So, the TDOA would be:

$$\mathrm{TDOA}_{\mathrm{left}} = t_l - (t_s + t_r)$$

But, since the left smartphone plays the noise bursts, the left microphone is shifted forward in time by $t_l + t_s + t_r$ for synchronisation. Then the TDOA between the left and right microphones would be:

$$\mathrm{TDOA}_{\mathrm{left}} = t_l + (t_l + t_s + t_r) - (t_s + t_r) = 2 \cdot t_l$$

Likewise, a signal from the right loudspeaker will arrive after time $t_l + t_s$ at the left microphone, and after time t_r at the right microphone. Again, given that the left microphone is shifted forward in time by $t_l + t_s + t_r$ for synchronisation, the TDOA between the left and right microphones would be:

$$\text{TDOA}_{\text{right}} = (t_l + t_s) + (t_l + t_s + t_r) - t_r = 2 \cdot (t_l + t_s)$$

The speed of sound in dry air at 20 °C is 343 m/s (Risoud et al., 2018). Then, from the measured distances, $t_l = 1$ ms, $t_s = 2$ ms and $t_r = 1$ ms, and the theoretical TDOAs would be $\text{TDOA}_{\text{left}} = 2$ ms and $\text{TDOA}_{\text{right}} = 6$ ms.

Figures 6.3 (a), (b) and (c) show the TDOA over time between the left and right recordings of the left snorer on their own, the right snorer alone, and both snorers together, respectively. The TDOAs were estimated with the GCC-PHAT function, which was computed on 500-ms frames with an overlap of 10 ms. The maximum value or peak of the GCC-PHAT function corresponded to the TDOA. TDOAs of approximately 2 and 10 ms were observed when playing the left snorer on their own and the right snorer alone, respectively. The same TDOAs were noted when playing both snorers together. However, given the configuration in which recordings took place, TDOAs around 2 and 6 ms were expected for the recordings of the left and right snorers, respectively. So, only one of the measured TDOAs agreed with the expected values, because of the several factors previously described: internal delay of the devices, sampling frequency mismatch, and reverberation.

Configuration 2: phones at each side of the bed, and initial noise bursts played from both phones

We just saw in the first experiment that estimating TDOAs in this scenario is challenging. For this reason, in the present experiment we examined an alternative synchronisation procedure, and recorded spectrally contrasting audio signals. This was done with the aim of ruling out the possible effect of masking on the estimation of TDOAs: a weaker signal is masked by a stronger one in the same frequency band (Roman et al., 2003), which might lead to a wrong estimation of TDOAs. For this and the following experiments, 3-second segments of band noise were played every 5 seconds for 2 minutes from the loudspeakers instead of snoring. One loudspeaker played white noise between 0–4 kHz, and the other one, white noise between 4–8 kHz in order to have a mixture of spectrally contrasting signals. Dual audio recordings were made with the same configuration as before (Figure 6.2). But this time the initial noise bursts for synchronisation were played from both phones: first from the device in 'Bed Partner' mode, then from the device in 'Me' mode.

Figures 6.3 (d), (e) and (f) present the obtained results. A TDOA around 2 ms was observed when the left band noise was played alone. When the right band noise was played on its own, a TDOA was noted at about 6 ms. The same TDOAs that were measured individually were seen when both loudspeakers played the signals. As expected, spectrally contrasting audio signals facilitated the correct estimation of TDOAs. In practice,

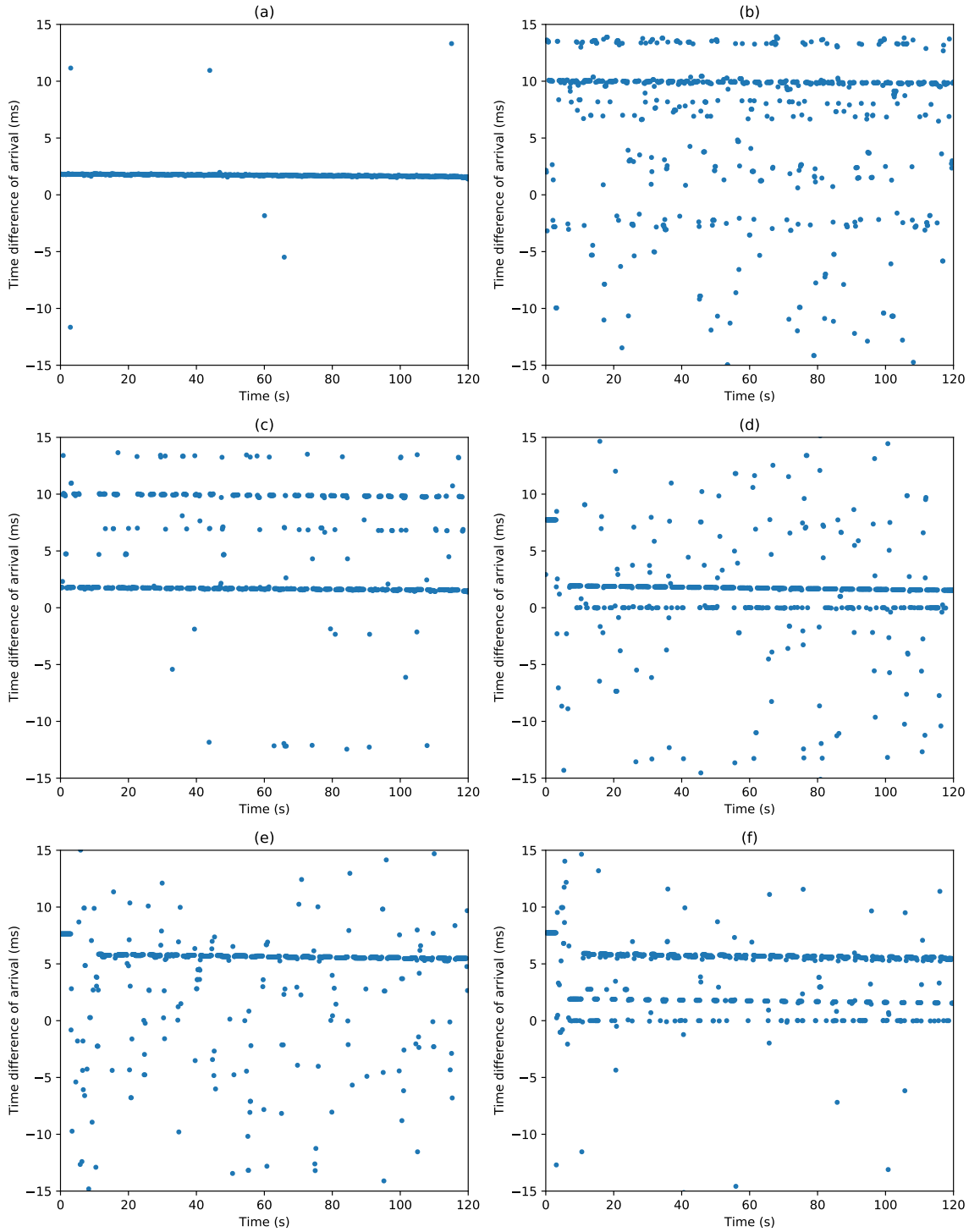


Fig. 6.3 TDOA between left and right recordings derived from the GCC-PHAT function. Configuration 1. (a) Left snorer played alone, TDOA: 2 ms. (b) Right snorer played alone, TDOA: 10 ms. (c) Both snorers played together, TDOAs: 2 and 10 ms. Configuration 2. (d) Left noise played alone, TDOA: 2 ms. (e) Right noise played alone, TDOA: 6 ms. (f) Both noises played together, TDOAs: 2 and 6 ms.

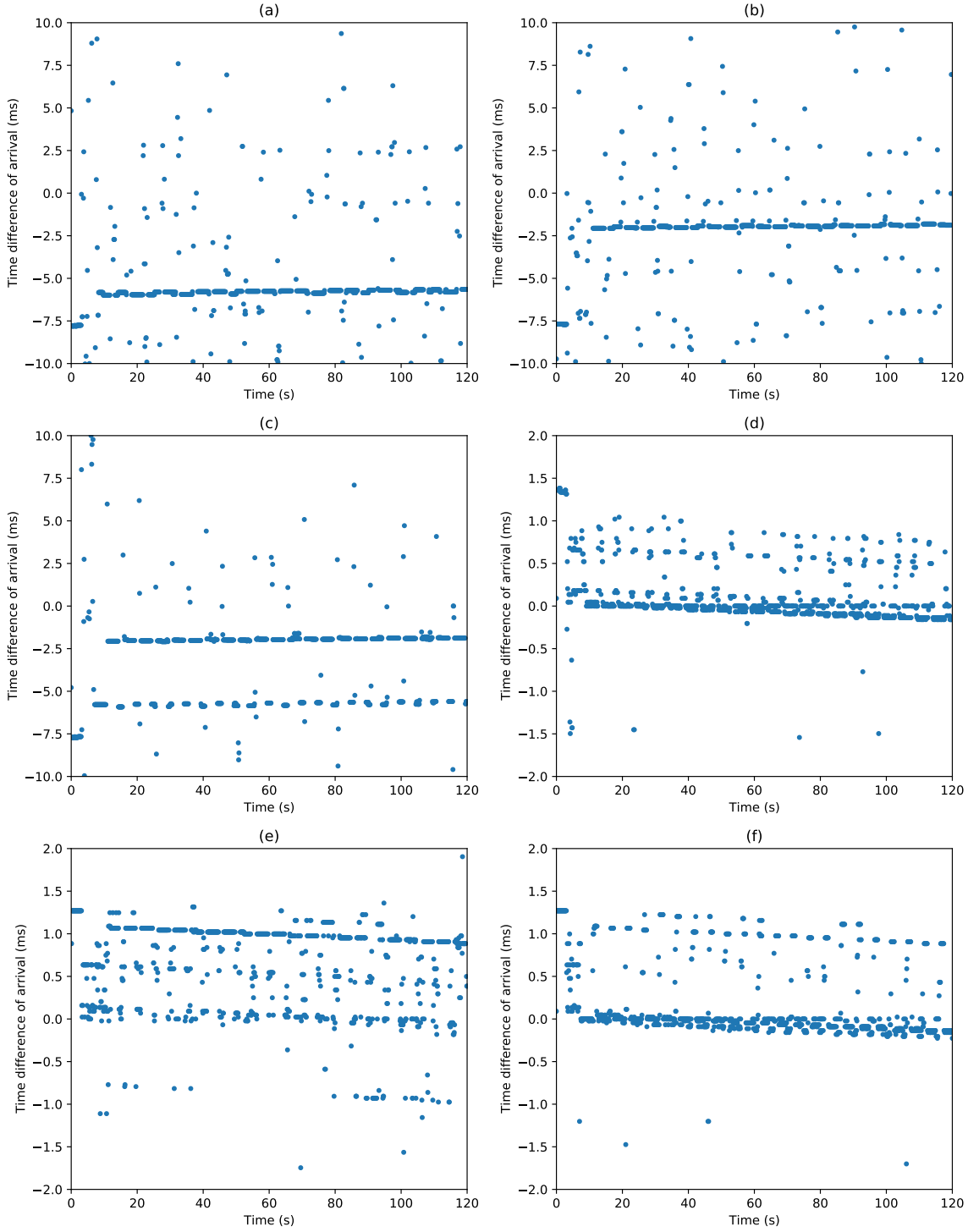


Fig. 6.4 TDOA between left and right recordings derived from the GCC-PHAT function. Configuration 3. (a) Left noise played alone, TDOA: -6 ms. (b) Right noise played alone, TDOA: -2 ms. (c) Both noises played together, TDOAs: -2 and -6 ms. Configuration 4. (d) Left noise played alone, TDOA: 0 ms. (e) Right noise played alone, TDOA: 1.2 ms. (f) Both noises played together, TDOAs: 0 and 1.2 ms.

however, two snoring signals would not be spectrally contrasting (Configuration 1), since these would exhibit similar frequency bands, as we saw in Chapter 3. On the other hand, the obtained results suggest that playing noise bursts from both phones instead of only one might be a more appropriate synchronisation procedure. Although the measured TDOAs did agree with the expected ones, the time differences drifted approximately 2 ms every 20 minutes, which limits the use of this kind of recordings for source separation, since precise synchronisation is critical for source separation techniques.

Configuration 3: swapped phones at each side of the bed, and initial noise bursts played from both phones

For this experiment the phones were swapped to examine the effect on the measured TDOAs when the location of the phone in 'Me' mode — the one controlling the Bluetooth connection and the audio recordings in both devices — is interchanged with that of the phone in 'Bed Partner' mode.

From the plots presented in Figures 6.4 (a), (b) and (c), it can be seen that the TDOAs estimated for the signals played individually were consistent with those measured for both signals played simultaneously. An absolute TDOA around 6 ms was noted when the left noise was played on its own. When the right noise was played alone, an absolute time difference of about 2 ms was observed. As expected, the location of the phone that controls the audio recordings and the Bluetooth connection did not have an effect on the estimated TDOAs. However, similar to what was reported in the previous experiment, the time differences drifted over time, which greatly limits the use of this type of recording for source separation.

Configuration 4: phones in the middle, and initial noise bursts played from both phones

Techniques for separating one audio signal from a mixture of signals — like beamforming — commonly make use of physically connected microphones separated by a few centimetres (Thiemann and Vincent, 2013). The experiments carried out so far suggest that, when the sensors are not physically connected and are far apart (e.g., phones at each side of the bed: 1.36 m), their functioning and the room acoustic characteristics are considerably different between them. So, in this last experiment, the phones were placed with a separation of 20 cm — simulating the average distance between the human ears (Risoud et al., 2018) — in the middle of the loudspeakers. The configuration in which recordings were made is shown in Figure 6.5.

Once more, before measuring the actual TDOAs, we will first work out the expected values using the nomenclature employed in Figure 6.5. Let t_l be the time of arrival from

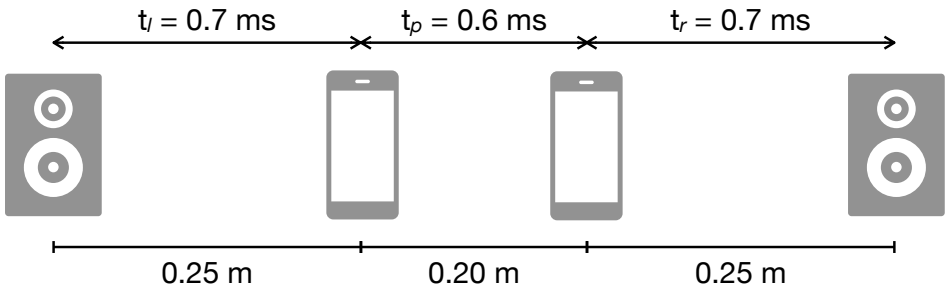


Fig. 6.5 Dual audio recording configuration

the left loudspeaker to the left phone; t_p, the time of arrival between both phones; and t_r, the time of arrival from the right loudspeaker to the right phone. A signal from the left loudspeaker will arrive at the left microphone after time t_l, and at the right microphone after time $t_l + t_p$. Then the TDOA would be:

$$\text{TDOA}_{\text{left}} = t_l - (t_l + t_p)$$

However, given that the left smartphone plays the noise bursts, the left microphone is shifted forward in time by t_p for synchronisation. Consequently, the TDOA between the left and right microphones would be:

$$\text{TDOA}_{\text{left}} = t_l + t_p - (t_l + t_p) = 0$$

In the same way, a signal from the right loudspeaker will arrive after time $t_p + t_r$ at the left microphone, and after time t_r at the right microphone. Again, since the left microphone is shifted forward in time by t_p for synchronisation, the TDOA between the left and right microphones would be:

$$\text{TDOA}_{\text{right}} = (t_p + t_r) + t_p - t_r = 2 \cdot t_p$$

Therefore, from the measured distances, $t_1 = 0.7$ ms, $t_2 = 0.6$ ms and $t_3 = 0.7$ ms, and the theoretical TDOAs would be $\text{TDOA}_{\text{left}} = 0$ ms and $\text{TDOA}_{\text{right}} = 1.2$ ms.

The estimated TDOAs over time are presented in Figures 6.4 (d), (e) and (f). It is worth noting that, due to reverberation, multiple tracks can be observed in the plots. When only the left loudspeaker was playing its signal, a TDOA was noted around 0.1 ms. A TDOA of about 1.2 ms was observed when only the right loudspeaker was playing its signal. These values were also noted when both loudspeakers were playing their signals simultaneously. However, the TDOAs were not sufficiently clear, and drifted over time. As discussed before, this is very likely caused by the internal delay of the devices (Gaubitch et al., 2013), and the sampling frequency mismatch between these (Miyabe et al., 2015; Ochi et al., 2016), which has been reported by related studies.

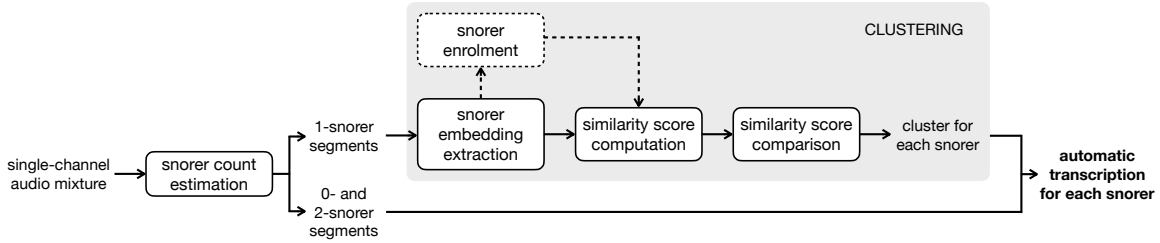


Fig. 6.6 Snorer diarisation system

Given the unreliability of dual audio recordings, we will now consider other strategies for screening a participant and their bed partner for sleep-disordered breathing in the same session. An alternative approach to using microphones in different devices for source separation might be using multiple microphones in one device. However, at the time of performing these experiments, no iOS application programming interface (API) that would allow access to multiple microphones was available. As discussed at the beginning of the present chapter, another possible approach might be using single-channel audio recordings made with one device, and leveraging deep learning techniques that have proven useful in related tasks, for instance, speaker diarisation. This will be our focus in the next section.

6.3 Snorer Diarisation

We have seen that using dual audio recordings made on different devices is challenging. Reverberation, sampling frequency mismatch, and internal delay of the devices result in unreliable TDOAs, which do not always match the expected theoretical values. This considerably limits the use of source separation techniques that rely on TDOAs. For this reason, an alternative approach was considered. The concept of *speaker* diarisation, which aims to identify who *spoke* when in multi-speaker audio recordings (Anguera et al., 2012), was extrapolated to *snorer* diarisation — who *snored* when — in a deep learning framework. It detects each subject's snore events in the same session using one single-channel sleep audio recording.

A diagram of the proposed snorer diarisation system is presented in Figure 6.6. It takes as input a single-channel sleep audio recording containing two snorers, and outputs an automatic snore transcription for each of them. This works on non-overlapping 250-ms segments, and is achieved by two subsystems: (1) snorer count estimation, and (2) clustering of 1-snorer events. The first subsystem estimates the number of active snorers in a 250-ms segment: 0 if there is no snoring, 1 if there is snoring from only one snorer or 2 if there is snoring from both snorers. 0- and 2-snorer segments are directly transcribed without any further processing, whereas 1-snorer segments are passed on to the

clustering subsystem. This assigns 1-snorer segments to a particular snorer based on a similarity score computed from snorer embeddings, and transcribes them for the corresponding snorer. Each subsystem will be considered in detail next.

6.3.1 Snorer Count Estimation

Snorer count estimation is approached as a classification task: a given audio segment is classified as containing breathing sounds from 0 snorers, 1 snorer or 2 snorers. 250-ms segments are used, since multiple frames are required for capturing enough temporal context to effectively estimate the number of active snorers and cluster snore events. Additionally, considering that snore events have an average duration of 1 second, as we discussed in Chapter 3, non-overlapping 250-ms segments provide a reasonable temporal resolution to determine the onset and offset of a snore event, while keeping the computation time low, taking into account that there are different subsystems.

To develop the snorer count estimation subsystem, the Snoring Sound Corpus presented in Chapter 3 was used to simulate 2-snorer mixtures. Before generating these, we employed an adaptive noise suppression algorithm (Dafna et al., 2013), since the signal-to-noise ratio (SNR) of the recordings was relatively low. Additionally, a bandpass filter (20–6000 Hz) was applied to attenuate low and high frequency noise that might be present, and all sleep audio recordings were normalised to the same root mean square (RMS) level before being mixed. The mixtures were generated at different SNRs: 5, 10 and 20 dB, to simulate the differences in amplitude arising from the fact that, in real conditions, one of the snorers is closer to the microphone, as the phone is expected to be placed on a bedside table. The rerecordings of snoring played back on loudspeakers performed in the previous section were used for calculating the expected SNR when there are breathing sounds from two snorers sleeping in a double bed. Assuming that the breathing sounds of the snorer next to the microphone are the signal, those of the other snorer are the noise, and both are equally loud and lying on their backs, an average SNR of 5 dB was measured. However, one snorer may be louder than the other, and people move during sleep, which results in different configurations, for example, one snorer facing the microphone and the other facing away from it. For these reasons, other SNRs were additionally considered.

Data from 4 subjects was used for training, and data from the remaining 2 subjects, for validation and testing. 2-fold cross-validation was performed: the participants used for validation and testing were rotated twice. We will refer to these as ‘pair 1’ and ‘pair 2’. Every possible pair of recordings from different subjects was mixed. 20 hours of data were produced. For the training dataset, 76,800 250-ms segments were generated for each of the three classes considered. For the validation and testing datasets, 9,600 segments

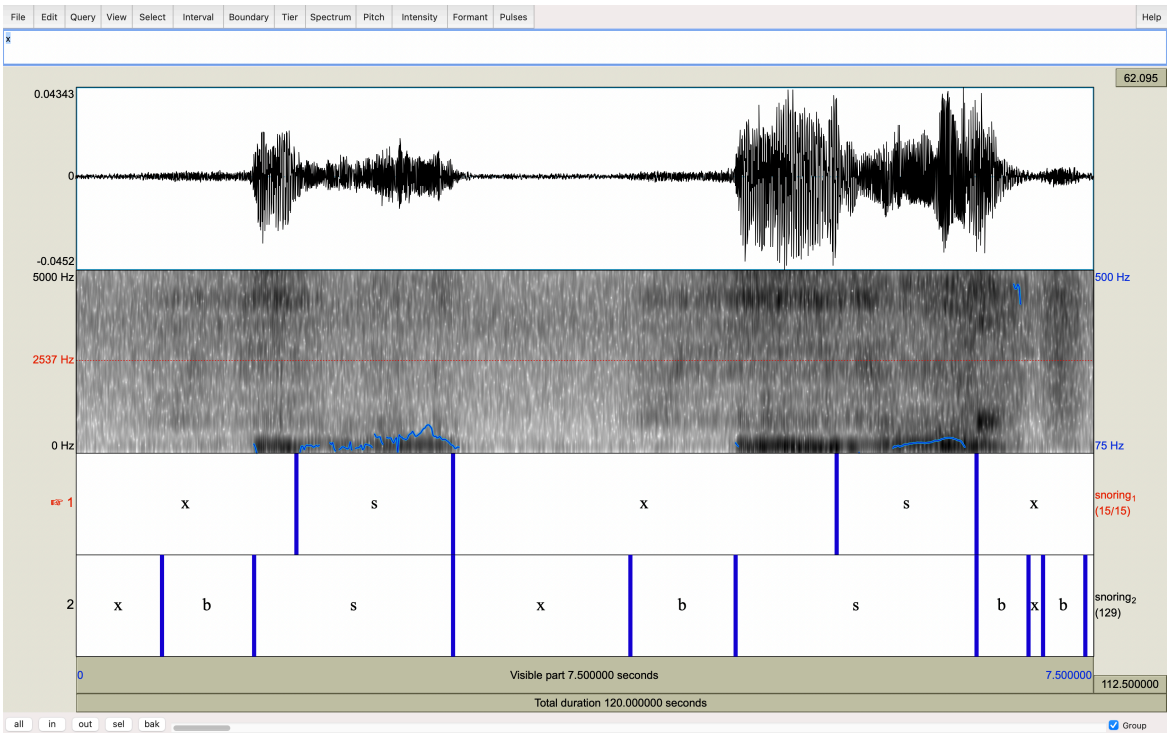


Fig. 6.7 Annotation of a 2-snorer sleep audio recording. The upper two panels display the waveform and spectrogram. Snore (s), breath (b), and silence (x) events are marked for each snorer in the lower two panels.

were generated for each class, respectively. The labels were automatically derived from the manual annotation taking into account the snore events.

In addition to the simulated 2-snorer mixtures, real 2-snorer single-channel audio recordings were collected from one pair of snorers with a smartphone in a bedroom at home. To manually annotate the recorded data, 2-channel audio recordings were made at the same time with binaural microphones placed on the bedhead close to each snorer, as the breathing events had to be assigned to a subject. An example is provided in Figure 6.7. Separate labels were manually generated for each snorer following the annotation scheme introduced in Chapter 3. Although sleep audio recordings were performed for two nights, a limited amount of data was usable, since the subjects were light snorers: $< 20\%$ of the recorded data contained snoring. Then, only 8 minutes of real 2-snorer audio recordings were manually annotated and used to evaluate the snorer diarisation system.

The snorer count estimation subsystem was built upon a deep neural network architecture that has been successfully used in speaker count estimation (Stoter et al., 2018). The log-compressed short-time Fourier transform (logSTFT) was provided as input to the network, since good frequency resolution is required to distinguish between the breathing sounds or speech produced by different individuals (Stoter et al., 2018). It was extracted from 250-ms segments using 25-ms frames with 10-ms overlap. A Hann window

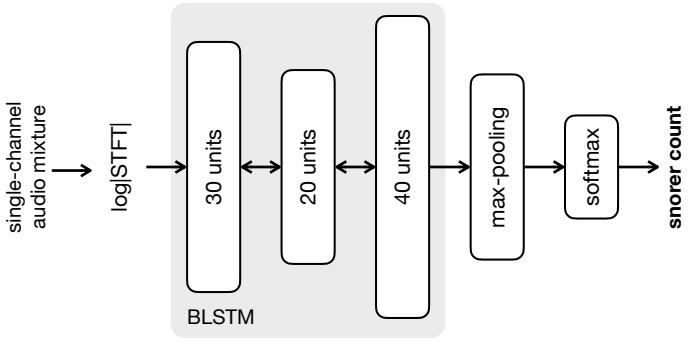


Fig. 6.8 Snorer count estimation

(Equation A.3) was applied each frame, and Z-score normalisation (i.e., zero mean and unit variance) with respect to the training dataset was performed for each logSTFT bin. The network consists of three layers of 30, 20 and 40 bidirectional long short-term memory (BLSTM) units, a max-pooling layer, and a fully connected layer with three softmax units, one for each class: 0 snorers, 1 snorer, and 2 snorers. This is illustrated in Figure 6.8. Given that no suitable baseline was available, a standard neural network was implemented as baseline: dense layers replaced those with BLSTM units. The network was trained in 50 epochs with a learning rate of 0.001, a batch size of 128, 50% dropout, and categorical crossentropy as the loss function.[1] It was developed using TensorFlow (Abadi et al., 2016), and the baseline was trained in the same way as the system.

As shown in Figure 6.6, the segments classified as 1-snorer segments are passed on to the clustering subsystem, which will be introduced in the following section, whereas those classified as 0-snorer segments do not need to be assigned to a particular snorer and those classified as 2-snorer segments are included for both snorers.

6.3.2 Clustering of Snore Events

Clustering of snore events is based on snorer embeddings: a learned feature representation that distinguishes one snorer from another. The sleep audio recordings collected to create the Snoring Sound Corpus that were not selected to be manually annotated — recordings from 41 different participants — were used to train the snorer embedding extraction system. 250-ms snore segments were extracted based on the automatic annotation generated by the best performing system for the classification of sleep-disordered breathing events presented in Chapter 4. 2,400 segments from each participant were used for training, and 300 segments from the same subjects were used for validation and

[1]Preliminary experiments conducted to determine the key hyperparameters of the DNNs developed in this chapter are reported in Appendix B.

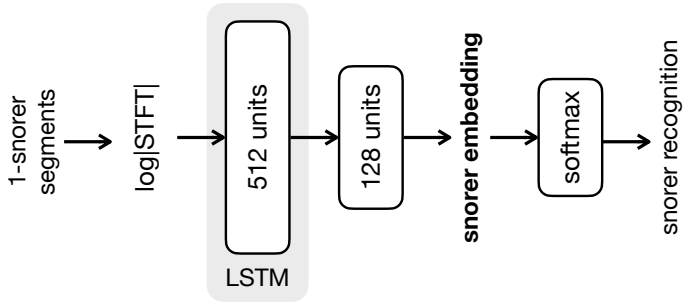


Fig. 6.9 Snorer embedding extraction (snorer recognition)

testing, respectively. These 41 subjects were different from those used to test the whole snorer diarisation system (i.e., it is snorer-independent).

Snorer embeddings are extracted with a deep neural network that has proven successful in speaker embedding extraction (Marchi et al., 2018). It consists of one layer of 512 long short-term memory (LSTM) units, a dense layer of 128 linear units, and a fully connected layer of 41 softmax units, one for every snorer in the dataset. This is illustrated in Figure 6.9. The same logSTFT features used for snorer count estimation are provided as input to the network, and the output of the second layer is the snorer embedding. Recognising each snorer is the objective during training. So, the last layer is discarded after training. The network was trained in 50 epochs with a learning rate of 0.001, a batch size of 128, batch normalisation, and categorical crossentropy as the loss function. It was developed using TensorFlow (Abadi et al., 2016). A standard neural network was implemented as baseline: a dense layer replaced that with LSTM units.

As described at the beginning of this section, the snorer diarisation system outputs an automatic snore transcription for each subject, which can be used to compute the time spent snoring or the snoring index (i.e., the number of snores per hour), for instance. Examples of these transcriptions are provided in Figures 6.10 and 6.11. It can be noted that the proposed snorer diarisation system performs reasonably well, although a few insertions and deletions can be observed.

Snorer Enrolment

The clustering subsystem requires snorer enrolment: the subject and their bed partner must provide one or two snores beforehand to extract five snorer embeddings for each of them. These are used as reference for clustering. The average cosine similarity is calculated between a given 1-snorer segment embedding and the reference embeddings of each snorer. Then, the average cosine similarities are compared, and the 1-snorer segment is assigned to the most similar snorer and transcribed. An alternative to this approach is extracting snorer embeddings from only one snorer, computing the average cosine simi-

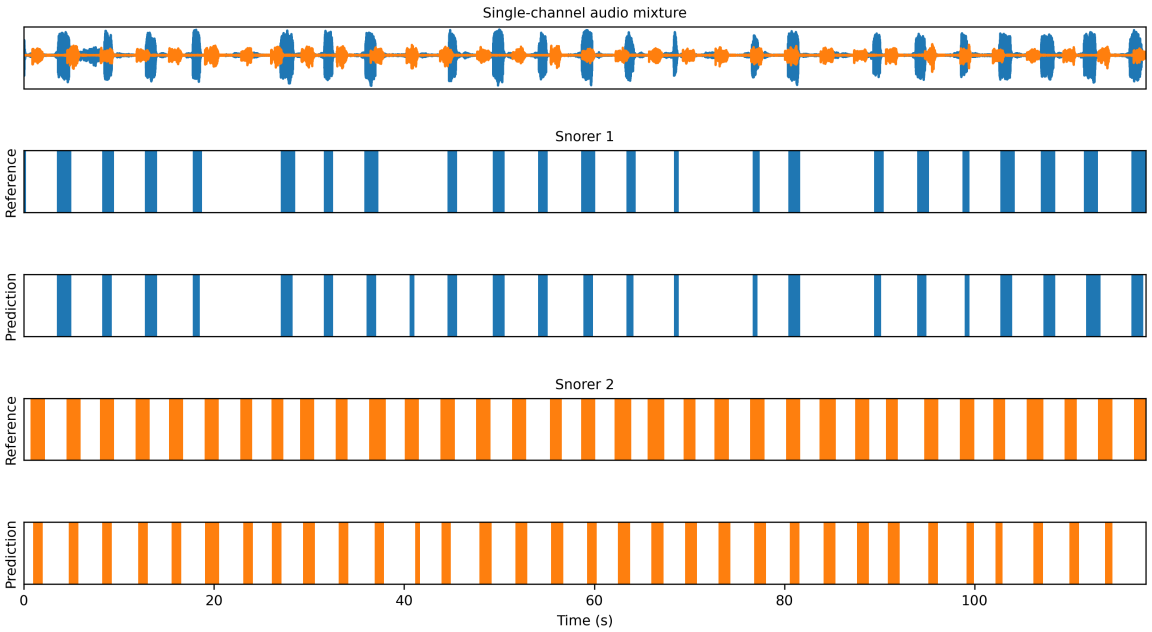


Fig. 6.10 2-minute snorer diarisation example. Snorer 1 is shown in blue, and snorer 2, in orange. The reference and prediction panels display the manually annotated and predicted snore events in the 2-snorer sleep audio recording, respectively.

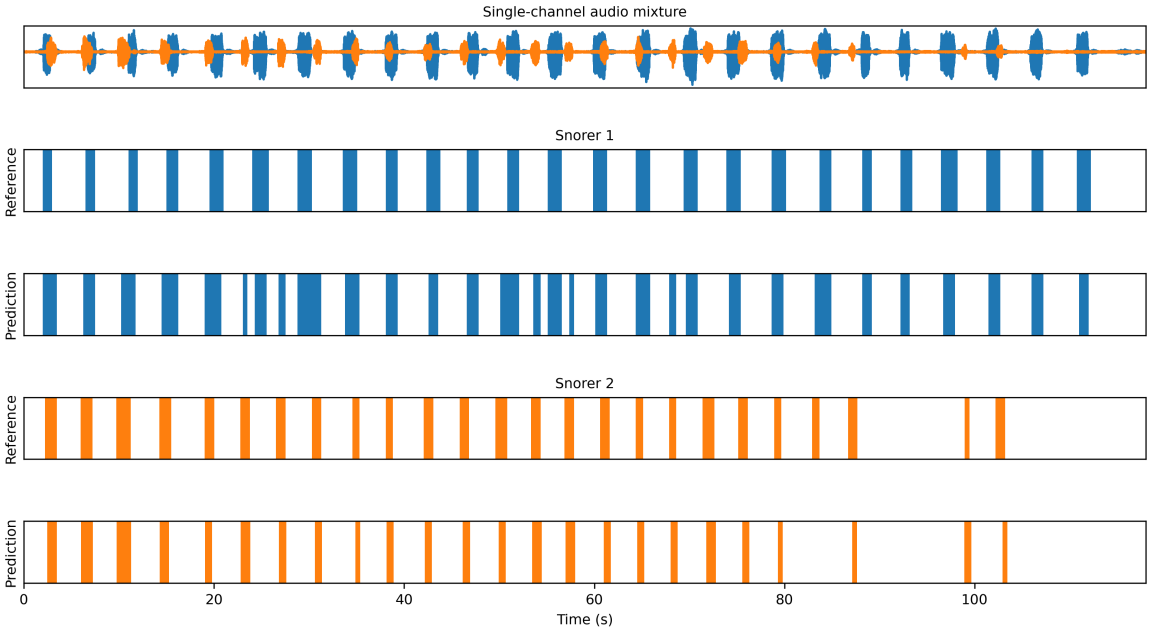


Fig. 6.11 2-minute snorer diarisation example. Snorer 1 is shown in blue, and snorer 2, in orange. The reference and prediction panels display the manually annotated and predicted snore events in the 2-snorer sleep audio recording, respectively.

larity as before, and using a threshold. If the average cosine similarity is above a defined threshold, a particular 1-snorer segment is assigned to the enrolled snorer. Otherwise, that segment is assigned to the other snorer. In the first experiment, such embeddings were extracted from snores in the mixtures.

Enrolling snorers with simulated or real snores is not practical. Firstly, simulated and real snoring differ, since vocal tract muscle tone is dependent on whether the subject is awake or not, and people usually produce a 'cartoonish' snore when asked to simulate one. Analysis of real and simulated snoring produced by one of the participants in the Snoring Sound Corpus showed that the mean pitch was decreased by 53%, the mean spectral centroid was increased by 62%, and the mean duration of simulated snores was increased by 42% with respect to real snores. Secondly, obtaining real snoring in advance is not possible because it would require the subjects to be asleep. However, after one night of use, real snores from the sleep audio recording might be possibly used for enrolment if the participants are able to reliably identify their snores in it. This remains to be investigated.

A practical alternative might be enrolling snorers with speech, as it can be easily obtained. The correlation between speech and sleep-disordered breathing has been investigated in previous studies, as both are generated in the vocal tract, and its physiology and tissue properties affect the production of sound. Fiz et al. (1993) analysed the harmonics of vowels uttered by OSA and healthy individuals, and described differences in vocalisation between the two groups. Robb et al. (1997) studied the formant frequencies and bandwidth of elongated vowels vocalised by OSA and healthy subjects, and found wider formant bandwidth and lower formant values for the OSA group in comparison with the healthy one. Formants are affected by the length of the vocal tract, which is believed to be longer for subjects with sleep-disordered breathing in comparison with healthy individuals. Fernandez et al. (2009) put together a speech corpus of OSA and healthy Spanish-speaking participants, and reported differences in nasalisation between both groups. Elisha et al. (2012) made use of features extracted from vowel and nasal phonemes to classify OSA and healthy Hebrew-speaking individuals. Botelho et al. (2019) used features computed from free speech, prolonged vowels /a/ and /i/, and read speech to distinguish between OSA and healthy Portuguese-speaking subjects.

These studies suggest that enrolling snorers with speech is plausible, since speech and sleep-disordered breathing are indeed correlated. For example, using a read sentence that highlights vowel and nasal phonemes:

> Why women and men are on my main ammonium moon.
>
> /wʌɪ ˈwɪmɪn ən(d) mɛn ɑː ɒn mʌɪ meɪn əˈməʊnjəm muːn/

Analysis of these phonemes and snoring produced by one of the participants in the Snoring Sound Corpus showed that the first and second formants of the /ʌ/ phoneme: 784 and 2,390 Hz, were similar to those of snoring: 783 and 2,299 Hz. This phoneme is produced by the vocal tract in neutral position (Story, 2016), the one in which breathing takes place. Then, other alternatives for snorer enrolment with speech might be uttering an elongated /ʌ/ or a read sentence with this phoneme, for instance:

> One Monday or Sunday we had so much fun having a honey bun for lunch with a young duck, a blood-coloured monkey and a dove in front of a London bus.
>
> /wʌn ˈmʌndeɪ ɔː ˈsʌndeɪ wiː had səʊ mʌtʃ fʌn haviŋ ə ˈhʌnɪ bʌn fɔː lʌn(t)ʃ wið ə jʌŋ dʌk ə blʌd ˈkʌləd ˈmʌŋki ənd ə dʌv ɪn frʌnt ɒv ə ˈlʌndən bʌs/

6.3.3 Evaluation

Diarisation error rate (DER) is typically reported for speaker diarisation systems. It considers as errors the false positives, confused, overlapped, and missed events with respect to the reference. However, the standard DER definition cannot be directly applied here, as there are important differences between the proposed snorer diarisation system and speaker diarisation systems. One of the main differences is that overlapped events are seen as errors in speaker diarisation systems, since such events are simply ignored, whereas in the proposed system they are transcribed for both snorers. Then, these are correctly transcribed events rather than errors. Another difference is that the snorer diarisation system generates an automatic transcription for each snorer, whereas a single transcription is generated for all speakers in speaker diarisation systems.

We evaluate the proposed snorer diarisation system at segment and event level. At segment level, the missed segments are false negatives, the false alarm segments are false positives, the overlapped segments are true positives if correctly detected, and the confused segments are false negatives for one snorer and false positives for the other. At event level, the missed events are deletions, the false alarm events are insertions, the overlapped segments are hits if correctly detected, and the confused events are insertions for one snorer and deletions for the other. Both forms of evaluation are done with respect to the reference (i.e., manual annotation). The DER is defined at segment and event level as follows:

$$\mathrm{DER}_{\mathrm{segment}} = \frac{\mathrm{FP} + \mathrm{FN}}{\text{snore segments in annotation}} \tag{6.7}$$

$$\mathrm{DER}_{\mathrm{event}} = \frac{\text{insertions} + \text{deletions}}{\text{snore events in annotation}} \tag{6.8}$$

Table 6.1 Results for the snorer diarisation baseline

Enrolment	Real snore				Simulated snore		Speech	
Test data	**Pair 1**	**Pair 2**	**Aggregated**	**Real**	**Pair 1**	**Real**	**Pair 1**	**Real**
Precision	66.06%	83.48%	74.86%	54.80%	57.64%	55.04%	66.96%	49.38%
Sensitivity	63.80%	70.93%	67.63%	66.81%	57.17%	67.46%	64.26%	60.52%
Specificity	86.99%	93.10%	89.95%	92.47%	83.33%	92.47%	87.42%	91.52%
Accuracy	80.41%	85.80%	83.10%	89.38%	75.90%	89.46%	80.84%	87.79%
F-measure	64.91%	76.69%	71.06%	60.22%	57.40%	60.62%	65.58%	54.39%
Segment DER	19.59%	14.20%	16.90%	10.62%	24.10%	10.54%	19.16%	12.21%
Event DER	26.68%	14.33%	20.52%	51.67%	31.45%	49.17%	25.12%	56.67%

Table 6.2 Results for the snorer diarisation system

Enrolment	Real snore				Simulated snore		Speech	
Test data	**Pair 1**	**Pair 2**	**Aggregated**	**Real**	**Pair 1**	**Real**	**Pair 1**	**Real**
Precision	66.41%	75.60%	71.30%	52.11%	61.56%	51.34%	66.10%	51.44%
Sensitivity	63.74%	71.07%	67.68%	59.00%	58.38%	58.35%	62.94%	58.13%
Specificity	87.21%	88.73%	87.94%	92.58%	85.54%	92.44%	87.20%	92.49%
Accuracy	80.54%	82.91%	81.73%	88.54%	77.83%	88.34%	80.31%	88.36%
F-measure	65.05%	73.27%	69.44%	55.34%	59.93%	54.62%	64.48%	54.58%
Segment DER	19.46%	17.09%	18.27%	11.46%	22.17%	11.66%	19.69%	11.64%
Event DER	23.04%	18.93%	20.99%	49.17%	26.78%	50.83%	21.34%	50.83%

Additionally, similar to the evaluation framework used in Chapter 4 for the classification of sleep-disordered breathing events, precision, sensitivity, specificity, accuracy and F-measure are evaluated at frame level. Confusion matrices are presented as well.

6.3.4 Results and Discussion

The results obtained for the baseline and the proposed system are shown in Tables 6.1 and 6.2, respectively. These are reported for three enrolment approaches: real snoring, simulated snoring, and speech. Additionally, the results are detailed for each snorer pair (Pair 1 and Pair 2), both pairs (Aggregated) where possible, and real 2-snorer sleep audio recordings (Real). When evaluated on simulated 2-snorer mixtures (Aggregated), the baseline and system performed reasonably well using real snoring for enrolment. These achieved a specificity above 87%, a sensitivity above 67% and a DER below 21%, which evidences that both neural network architectures were able to effectively detect snoring, and successfully cluster snore events. The standard DNN baseline performed slightly better than the LSTM system due to the limited amount of data available. More complex neural network architectures usually require a larger amount of data to properly generalise in

comparison with standard architectures, as a greater number of parameters have to be learned. For real 2-snorer audio recordings, the baseline and system using real snoring for enrolment attained a lower precision and a higher event DER with respect to simulated mixtures, since many false positives were generated by the classifiers. This was caused by a difference in the amount of snore events in the data: simulated 2-snorer recordings contained a great amount of snoring, whereas real 2-snorer recordings contained fewer snore events.

The results show that it is possible to cluster 1-snorer events using alternative enrolment approaches in spite of having trained the network for extracting snorer embeddings only with real snoring data. Using a read sentence for enrolment resulted in a performance comparable to enrolling with real snoring: F-measures were about 65%, and segment DERs were around 20% in both cases. When evaluated on simulated 2-snorer mixtures, enrolling with simulated snoring degraded the performance, since, as we previously discussed, simulated and real snoring differ. However, this was not evident when evaluated on real 2-snorer mixtures. Although the DNN baseline performed slightly better in most cases, the LSTM system was better suited to deal with real data and snorer enrolment with speech. The performance of the clustering subsystem when using a read sentence might be further improved by training the network for extracting snorer embeddings with real snoring and speech from the same subjects. In this way, the network would learn vocal tract embeddings rather than snore embeddings.

Thus far, we have evaluated the whole snorer diarisation system. We will now evaluate each subsystem individually to assess its contribution to the whole system. In Figure 6.12 the confusion matrices for the snorer count estimation baseline and system tested on the snorer pair 1 and 2 are shown. It can be noted that the system and baseline had a similar performance on 0-snorer classification, as this class clearly differed from the other two classes, so the standard network architecture performed as good as the more complex one. The system performed better at 2-snorer classification in comparison with the baseline, but the latter performed better than the former at 1-snorer classification, since the baseline incorrectly classified more 2-snorer segments as 1-snorer, whereas the system incorrectly classified more 1-snorer segments as 2-snorer. This suggests that the BLSTM architecture was better suited than the standard DNN for classifying the most complex class, as, in addition to detecting snoring sounds, the network had to determine if these came from two snorers. The BLSTM network effectively did this by exploiting the temporal information in the features. Previously, in Chapter 4, standard DNNs were shown to successfully classify 1-snorer events in a related task, then the standard network architecture was properly suited for classifying such events in the snorer diarisation task, and the additional temporal information that the BLSTM architecture was able to exploit did not prove beneficial in this case.

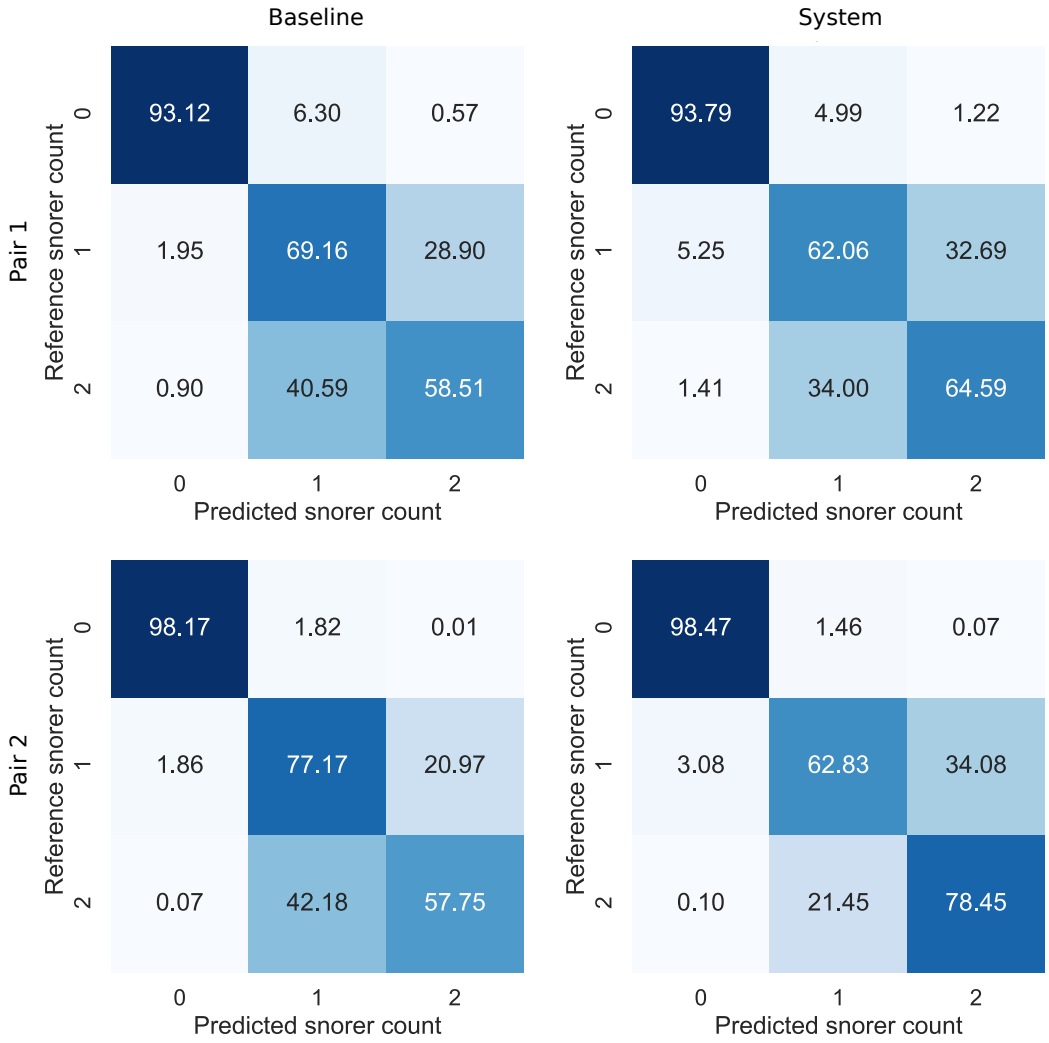


Fig. 6.12 Confusion matrices for the snorer count estimation baseline (left panels) and system (right panels) tested on the snorer pair 1 (top panels) and 2 (bottom panels)

The snorer embedding extraction (clustering) system when employed for recognising the 41 snorers used for training excluding the snorer pair 1 achieved an average recall of 78.99%, which outperformed the baseline average recall of 73.49%. All snorers had the same amount of snore segments. The same behaviour was noted when excluding the snorer pair 2: the baseline obtained an average recall of 73.62%, whereas the system, an average recall of 79.68%. It indicates that the LSTM architecture was better suited for this task than the standard dense architecture, since, once more, the LSTM network exploited the temporal information in the features to effectively recognise snorers.

To ensure that the network did not learn microphone or room embeddings, we enrolled snorers using 'other' events in the sleep audio recordings, for instance, background noise or a car passing by. This was evaluated on the snorer pair 1. The confusion matrices for the clustering of 1-snorer segments with the LSTM system using 'other' events,

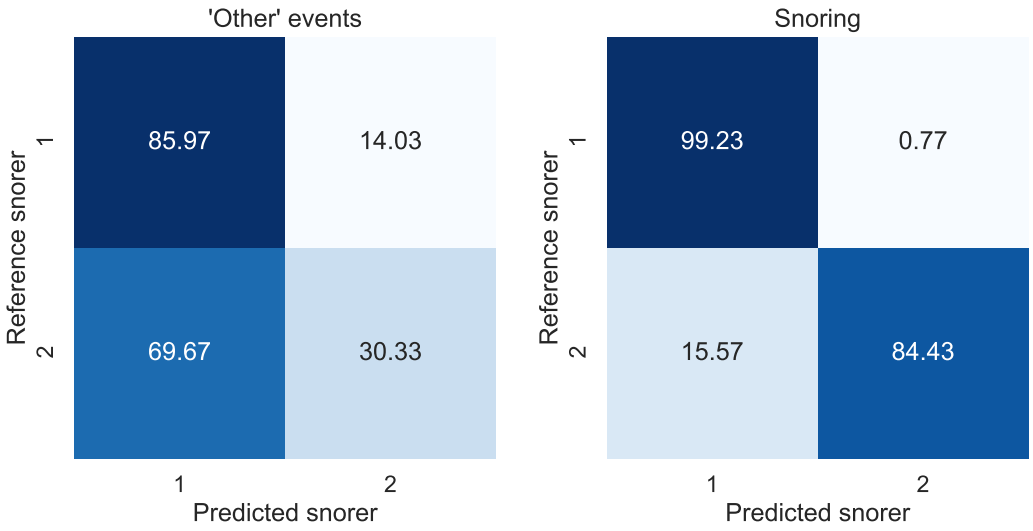


Fig. 6.13 Confusion matrices for the clustering of 1-snorer segments with the proposed system using 'other' events and snoring for snorer enrolment. 1: snorer 1, 2: snorer 2.

and snoring for enrolment are presented in Figure 6.13. It can be seen that the network successfully learned snorer embeddings rather than room or microphone embeddings: enrolling with 'other' events resulted in an average recall of 58.15%, whereas enrolling with snoring yielded an average recall of 91.83%.

6.4 Summary

Although couples typically spend the night in the same bed in Western societies, sleep is mostly considered from a healthcare standpoint as an individual phenomenon, and co-sleeping is still an overlooked topic. Ideally, systems to screen for sleep-disordered breathing should be robust to bed partner breathing sounds. Even more, it would be useful to screen a subject and their bed partner for sleep-disordered breathing in the same session. For this purpose, two different approaches were considered in this chapter: (1) dual audio recordings for source separation, and (2) snorer diarisation. The first approach was conceived with the idea of separating each subject's breathing sounds and running our previously developed systems on the separated signals. Methods for separating one audio source from a mixture of signals usually require multiple microphones. However, the data collected for the classification of sleep-disordered breathing events and screening for OSA consists of single-channel audio recordings from participants sleeping on their own. For this reason, the required data — multichannel audio recordings from two snorers — had to be simulated using the available recordings. Pairs of audio recordings from the Snoring Sound Corpus were played on loudspeakers placed on a bed

at home, and the playback was recorded simultaneously with two smartphones — one at each side of the bed — using a custom app.

By artificially mixing two audio signals in that way, three room acoustics were mixed as well: one from each room in which the original recordings took place, and one from the room in which the original recordings were played back and rerecorded. This was a limitation, as it hindered the synchronisation between the audio rerecordings. Another limitation was that the loudspeakers did not epitomise human sound emission (Farnsworth, 1942; Flanagan, 1960) and sleeping patterns. For example, the loudspeakers were stationary, whereas people move during sleep. We found that accurate synchronisation between audio recordings made on different devices and, therefore, source separation from this kind of recordings are challenging tasks. In most cases, the measured TDOAs drifted over time, and did not always match the theoretical values, since the internal delay of the recording devices, the sampling frequency mismatch between these, and reverberation resulted in a wrong estimation of the TDOAs. It considerably limited the use of this kind of recordings for source separation, as precise synchronisation is critical for source separation techniques.

Given the unreliability of dual audio recordings, we considered an alternative approach. The concept of speaker diarisation was extrapolated to snorer diarisation in a deep learning framework. It takes as input a single-channel sleep audio recording containing two snorers, and outputs an automatic snore transcription for each of them. This is achieved by two subsystems: (1) snorer count estimation, and (2) clustering of 1-snorer events. Snorer count estimation is approached as a classification task: a given audio segment is classified as containing breathing sounds from 0 snorers, 1 snorer or 2 snorers. This is done with a deep neural network using logSTFT features. The segments classified as 1-snorer segments are passed on to the clustering subsystem, whereas those classified as 0- and 2-snorer segments are transcribed for both snorers without any further processing. Clustering of 1-snorer events is based on snorer embeddings: a learned feature representation that distinguishes one snorer from another. Snorer embeddings are extracted with a deep neural network trained to recognise snorers using the same logSTFT features. The clustering subsystem requires snorer enrolment: the subjects must provide snoring or speech beforehand to extract snorer embeddings for each of them, which are used as reference for clustering. The average cosine similarity is calculated between the embedding of a given 1-snorer segment and the reference embeddings of each snorer. Then, the average cosine similarities are compared, and the 1-snorer segment is assigned to the most similar snorer and transcribed.

A key strength of the proposed snorer diarisation system is that it only requires a single-channel audio recording made with readily available hardware (i.e., a smartphone), which might contribute to improving accessibility to sleep-disordered breathing diagno-

sis. Another strength is that it exploits the relationship between breathing sounds during sleep and speech to conveniently enrol snorers. However, one of the limitations of the proposed method is that it consists of two separate subsystems that were optimised individually for their particular task instead of jointly for the objective of the whole snorer diarisation system. Because of the limited amount of data, an end-to-end approach, for example, could not be implemented. Furthermore, errors from the first subsystem are carried to the second subsystem with no possibility of recovery, since both output discrete decisions. A probabilistic approach might allow both subsystems to better interact, which would potentially result in a more robust system. For instance, a 2-snorer segment incorrectly classified by the first subsystem as 1-snorer will likely have very low similarity scores with the embeddings of both snorers. Instead of assigning such a segment to the snorer with the highest similarity score, it could be rectified as a 2-snorer segment based on the very low similarity scores.

We have now come to the end of the experimental work in this thesis. In the final chapter, we will revisit the research questions that were put forward at the beginning of the thesis to reflect on the contributions and limitations of this study, and consider future research directions.

Chapter 7

Summary and Scope for Future Work

"In theory, sleep is a negative thing, a mere cessation of life. But nothing will persuade me that sleep is not really quite positive, some mysterious pleasure which is too perfect to be remembered. It must be some drawing on our divine energies, some forgotten refreshment at the ancient fountains of life. If this is not so, why do we cling to sleep when we have already had enough of it; why does waking up always seem like descending from heaven upon earth?"

– G. K. Chesterton

The present thesis has investigated the application of deep learning methods for the acoustic analysis of sleep-disordered breathing. This problem was tackled using acoustics because it has benefits in terms of physical comfort, cost, and keeping the contact to a minimum, which is increasingly relevant in the context of the COVID-19 pandemic. Whereas most related studies have made use of data collected in controlled conditions with specialised sensors that must be attached to the body, the systems proposed here use sleep breathing sounds to conveniently screen for sleep-disordered breathing in a bedroom at home using a smartphone.

The tools of speech technology, which have been extensively researched and developed, were a good starting point for the acoustic analysis of sleep-disordered breathing, since there are some similarities between speech and breathing sounds. Both are time-series signals, have a similar frequency range, and are produced in the vocal tract. However, there are some differences between them. For example, pitch has been traditionally disregarded in speech technology systems, as it does not provide relevant information for recognising speech (with the exception of tonal languages), whereas pitch does contribute useful information for classifying sleep-disordered breathing events. Speech production is a conscious process, unlike breathing during sleep. Therefore, speakers

are aware of each other in a conversation and do not overlap frequently (Kurtic et al., 2013), whereas co-sleepers' breathing events commonly overlap. This makes the analysis of breathings sounds in typical sleep conditions, which might include a bed partner, a challenging task.

When compared to related studies, a key strength of the proposed systems is that they were developed by applying deep learning to exploit complex patterns in data rather than by handcrafting sets of rules, which are often proposed in related studies (Al-Mardini et al., 2014; Castillo-Escario et al., 2019; Narayan et al., 2018; Saha et al., 2020). Rule-based methods are unlikely to be robust enough in practice to effectively deal with the great variability of breathing sounds and room acoustic characteristics. For instance, an approach that detects snores by only using sound energy (Alakuijala and Salmi, 2016) would very likely detect other loud events such as speech or aircraft sound incorrectly as snores, although it is true that most people sleep in a reasonable quiet environment and somniloquy is rare (Caylak, 2013). Detecting apnoeas by simply looking for silence (Almazaydeh et al., 2013) would probably mistake a period of quiet healthy breathing for an apnoea in ambient sound recordings, for example. However, this would not be an issue when using obtrusive tracheal sound recordings, as they pick up quiet breathing.

Each of the research questions put forward in Section 1.3.1 will be restated now along with a summary of how this thesis addressed it.

(RQ1) What are the desirable characteristics for an acoustic corpus of sleep-disordered breathing?

An acoustic corpus of sleep-disordered breathing should be collected in a variety of realistic sleep environments with readily available hardware, and use a consistent and detailed annotation scheme. These characteristics might facilitate scaling up the technology developed in studies such as the one reported here in the general population. In Chapter 3, the Snoring Sound Corpus and the OSA Sound Corpus used in this study were introduced. Both were collected from participants who slept in their own bedrooms at home with a smartphone placed on a bedside table within arm's reach, as improving access to diagnosis is one of the main challenges that sleep medicine is still facing (Dafna et al., 2013; Phillips, 2007). This also allowed us to have a diversity of room acoustic characteristics, microphone responses and levels of background noise that epitomise realistic sleep conditions.

The Snoring Sound Corpus was manually annotated considering six acoustic events: 'snore', 'breath', 'noisy in-breath', 'wheezing', 'silence', and 'other'. These could be collapsed as needed. For example, 'noisy in-breath', 'wheezing', and 'snoring' were collapsed into the 'snoring' class for classification tasks. Inter-annotator agreement was evaluated with Cohen's kappa. Values between 0.43 and 0.85 were reported, which evidences that, although

some annotations were easily done, annotating sleep-disordered breathing data is not a straightforward task due to the variability of breathing sounds and the subjectivity involved in the process. The OSA Sound Corpus was collected during HSAT. HSAT data was scored by a Registered Polysomnographic Technologist who manually annotated obstructive apnoeas, central apnoeas, and hypopneas based on respiratory effort and airflow information. This provided the reference to label the acoustic data.

A reasonable volume of data is needed to effectively capture the great variability of breathing sounds and sleep-disordered breathing conditions. The Snoring Sound Corpus is made up of 6 hours of data from 6 subjects, and the OSA Sound Corpus comprises 416 hours of data from 45 participants. While it is true that sleep-disordered breathing is more prevalent in men than in women (Chuang et al., 2017; Yaggi and Strohl, 2010), an acoustic corpus of sleep-disordered breathing should ideally be gender-balanced. However, this is not the case in our corpora. The Snoring Sound Corpus consists of data from male participants only, and 42% of the participants in the OSA Sound Corpus are female.

(RQ2) What is the acoustic definition of snoring?

Using the Snoring Sound Corpus and the OSA Sound Corpus, the principal forms of sleep-disordered breathing were characterised from an acoustic point of view in Chapter 3. Although there is a commonly understood perception of snoring, there is no widely accepted acoustic definition of it. The duration, pitch, waveform, and spectral shape of the snore events in the Snoring Sound Corpus were examined. Based on this examination, the following definition of snoring was proposed:

> *Snoring is a pitched breathing sound during sleep generated by the vibration of at least one of the following upper airway structures: soft palate, epiglottis, pharyngeal walls or tongue. It is a periodic loud event with a duration of about 1 second, occurs approximately every 3 seconds, its pitch is near 85 Hz, and its spectral centroid is around 1,700 Hz.*

(RQ3) To what extent is it possible to robustly screen for sleep-disordered breathing using acoustics?

In general, breathing events follow a temporal pattern (i.e., breathing is a sequence of events), but this is not usually taken into account when analysing breathing sounds in related studies. In Chapter 4, such a pattern was exploited for the classification of sleep-disordered breathing events with especial attention to snoring. Evaluated on the created Snoring Sound Corpus, it was shown that incorporating information on the sequence of events allowed a better classification performance than considering breathing events in isolation, as the former approach encourages realistic event durations.

Chapter 5 introduced a system to screen for OSA, the most severe form of sleep-disordered breathing. The screening system, rather than detecting individual apnoea-hypopnea events, exploits the temporal pattern of breathing sounds to predict the presence of apnoea-hypopnea events in large audio recording segments. A DNN architecture with convolutional layers was shown to effectively capture the temporal pattern of breathing sounds from time-frequency representations. Screening experiments on the created OSA Sound Corpus showed that the temporal pattern of breathing sounds can be exploited to accurately screen for OSA. The classifiers and feature representations — bottleneck features from auditory nerve firing rate maps (see RQ4) — in Chapters 4 and 5 were learned from data collected across a variety of acoustic conditions applying deep learning rather than based on sets of handcrafted rules, which are unlikely to successfully deal with the wide range of breathing sounds during sleep and room acoustic characteristics.

As noted above, acoustics were used for the analysis of sleep-disordered breathing in contrast to other approaches that use multiple sensors attached to the body. However, an interesting compromise between unobtrusiveness and physiological information being available is combining acoustic evidence with only one sensor rather than the whole set of sensors used in PSG or HSAT. This allowed us to further assess the screening capability of using acoustic evidence on its own. Chapter 5 described an approach that integrates acoustic features with only one physiological parameter, oxygen saturation, to screen for OSA. This approach singles out the audio recording segments associated with desaturations, which are classified by the proposed DNN. The system is able to screen for OSA with encouraging sensitivity and specificity, as oxygen saturation directly reflects the physiological effects of OSA. However, its measurement requires specialised hardware.

(RQ4) How can a large amount of unlabelled sleep-disordered breathing data be leveraged to achieve robustness to background noise in typical sleep conditions?

Even though large amounts of sleep audio recordings collected in typical sleep conditions were available in this study, only a subset of them were manually labelled, as annotating data is a time-consuming and expensive task. For this reason, unlabelled sleep audio recordings were leveraged by applying unsupervised learning to learn feature representations. In Chapters 4 and 5, bottleneck features from auditory nerve firing rate maps and the autocorrelation function were learned with autoencoder DNNs, and used for classifying sleep-disordered breathing events and screening for OSA, respectively. The first half of an autoencoder encodes the input into a compressed representation — otherwise known as bottleneck features — and its second half reconstructs the input by decoding the compressed representation. Successfully reconstructing the input requires the bottleneck features to capture the important characteristics. Therefore, the relevant infor-

mation is assumed to be encoded within such features. No labels are required to train an autoencoder DNN, as the expected output is the input itself.

Using bottleneck features, instead of conventional acoustic features like MFCCs or logSTFT, resulted in more robust systems. Pitch information was proven useful for classifying sleep-disordered breathing events. Rate maps and, by extension, bottleneck features derived from them better capture pitch information in comparison with MFCCs and logSTFT, since rate maps offer an expanded low-frequency representation of the spectrum of the audio signal (Wang and Brown, 2006). The bottleneck features derived from the autocorrelation function also capture pitch information effectively, as this function is the basis for many pitch determination algorithms (Cosi et al., 1984; De Cheveigne and Kawahara, 2002; Hess, 1983; McLeod and Geoff, 2005). By contrast, MFCCs purposely discard pitch information.

(RQ5) To what extent is it possible to achieve robustness to bed partner breathing sounds?

Ideally, systems to screen for sleep-disordered breathing should be robust enough to be applied in typical sleep conditions, which might include a bed partner who can negatively impact the performance of the screening systems. This is not usually considered in related studies, as participants are assumed to be sleeping on their own. Drawing inspiration from speaker diarisation research, Chapter 6 presented a snorer diarisation system that allows the differentiation of breathing sounds from two subjects using single-channel audio recordings. A DNN architecture with LSTM layers was shown to successfully use temporal information in the audio signal to estimate the number of active snorers and cluster snore events. Exploiting the relationship between speech and sleep-disordered breathing, snorers are effectively enrolled with a read sentence that highlights vowels and nasal phonemes. Evaluated on simulated 2-snorer mixtures generated from the Snoring Sound Corpus and a limited amount of real 2-snorer sleep audio recordings, the snorer diarisation system exhibits good performance at various SNRs, considering the challenging conditions of the task.

7.1 Contributions

An Acoustic Screening System for OSA

The system to screen for OSA from sleep breathing sounds that was developed contributes to improving accessibility to sleep-disordered breathing diagnosis. It can be applied to high-risk subjects to single out those who really need PSG to evaluate their condition, which might facilitate a better use of the limited PSG resources. Unlike PSG, the screen-

ing system uses readily available hardware, and can be applied in a typical sleep environment. This has practical importance, since it could help people to manage their condition at home. In addition to screening, it can be used for long-term monitoring, as PSG is usually carried out for only one night because of its high cost and the progression of the condition is, therefore, not assessed. A preliminary version of the acoustic screening system for OSA was published and presented in *Sleep 2020* (Romero et al., 2020c). At the time of writing this document, a paper presenting the final version of the screening system is under review for publication in the *IEEE Journal of Biomedical and Health Informatics*.

Novel Feature Representations for Detecting Sleep Breathing Sounds

Novel feature representations — bottleneck features from auditory nerve firing rate maps and the autocorrelation function — were proposed to robustly classify sleep-disordered breathing events, and screen for OSA. Given that labelled data was not easily available, as manual annotation is costly and time-consuming, large amounts of unlabelled data were leveraged by learning feature representations in an unsupervised manner. Bottleneck features were proven to perform better than standard features, like MFCCs or logSTFT, since these novel feature representations for sleep breathing sounds were learned from data rather than obtained from a handcrafted set of rules. A paper presenting the Snoring Sound Corpus and the proposed novel feature representations was published and presented in *ICASSP 2019* (Romero et al., 2019).

A System for Snorer Diarisation

The concept of snorer diarisation was introduced with the aim of developing a system to screen for sleep-disordered breathing robust to bed partner breathing sounds. Although typical sleep conditions commonly include a bed partner, who can negatively impact the performance of acoustic screening systems, most studies to date assume the potential users will be sleeping on their own. Snorer diarisation allows the screening of a subject and their bed partner for sleep-disordered breathing in the same session using single-channel sleep audio recordings. Exploiting the relationship between breathing sounds and speech, snorers can be conveniently enrolled with a read sentence that highlights vowels and nasal phonemes. A paper on the proposed snorer diarisation system was published and presented in *ICASSP 2020* (Romero et al., 2020b).

The Snoring Sound Corpus

The Snoring Sound Corpus was created to overcome data scarcity in the field of acoustic analysis of sleep-disordered breathing. It consisted of 354 minutes of audio recordings collected from 6 male participants in realistic sleep conditions with readily avail-

able hardware. That is, in a bedroom at home with a smartphone. This differs from other sleep-disordered breathing corpora that have been collected in laboratories (Emoto et al., 2018; Nonaka et al., 2016) with specialised hardware such as ceiling mounted microphones (Kim et al., 2018). The Snoring Sound Corpus was manually annotated using a detailed annotation scheme, which considered snore, breath, noisy in-breath, wheezing and silence events, unlike related studies that have simply annotated snore and non-snore events (Dafna et al., 2013; Xie et al., 2021). It was used for different acoustic analysis tasks, for example, characterisation, classification, etc. At the time of writing this document, there were no immediate plans to release the Snoring Sound Corpus, as it was collected in association with Passion for Life Healthcare, one of the PhD sponsors, and has particular commercial value to them.

An Acoustic Characterisation of Sleep-Disordered Breathing

The main forms of sleep-disordered breathing — snoring and OSA — were characterised from an acoustic perspective. This informed the design of the systems developed in this study. For example, the segment length for analysis in the system to screen for OSA was determined based on the statistics of the duration of apnoea-hypopnea events. The findings regarding the pitch and formants of snoring events inspired snorer enrolment with speech in the snorer diarisation system. The spectral shape of snoring informed the frequency range used for the computation of acoustic features. Lastly, given that snoring does not have a widely accepted acoustic definition, one was stablished from its characterisation. A preliminary version of the acoustic characterisation of sleep-disordered breathing was published and presented in *Acoustics 2020* (Romero et al., 2020a).

7.2 Limitations

Limited Data

A limited amount of data was used in this thesis in comparison with related studies. For example, Nakano et al. (2019) used data from 1,852 subjects collected over the course of 9 years, whereas our Snoring Sound Corpus consisted of 354 minutes of data from 6 participants, and our OSA Sound Corpus, of 416 hours of data from 45 participants — apnoeas are scarce events though. It poses limitations on the experiments reported, as the great variability of breathing sounds, room acoustic characteristics, and microphone responses might not have been fully considered in the developed systems. It also limited the machine learning techniques that could be used. For instance, there was not enough data to properly train end-to-end systems or attention models. Limited data is a limitation of many studies working on the acoustic analysis of sleep-disordered breath-

ing (Al-Mardini et al., 2014; Castillo-Escario et al., 2019; Duckitt et al., 2016; Emoto et al., 2018; Sun et al., 2015). In general, healthcare-related studies have this limitation because of privacy concerns, the high cost of diagnostic tests, and the time-consuming scoring of these (Meijerink et al., 2020). For example, a PSG test costs around £700 per night in the United States (Kim et al., 2015), and a polysomnography technologist spends up to two hours to manually score it (Malhotra et al., 2013).

Male Bias

The Snoring Sound Corpus consisted of data from male snorers only, which made the systems for the classification of sleep-disordered breathing events male-biased. This might probably limit their performance on data from female snorers due to known anatomical differences between the vocal tract of female and male participants (e.g., length), and the resulting acoustic dissimilarities (Simpson, 2001). However, the developed systems were run on a small amount of data from a female snorer, and they performed well: most snore events were correctly classified. While it is true that snoring is more prevalent in men than in women (Chuang et al., 2017), differences between male and female snoring remain to be studied. The OSA Sound Corpus consisted of data from male and female participants nonetheless.

Simulated Data

Simulated data was used for developing and testing the snorer diarisation system. This also poses a limitation, since the simulated data was generated by mixing pairs of audio recordings from the Snoring Sound Corpus. Although the differences in amplitude arising from the fact that one of the snorers is closer to the microphone were simulated, and it was shown that clustering of snore events is not based on microphone responses or room acoustic characteristics, the simulated data is unlikely to epitomise co-sleeping characteristics. For example, it does not account for the sleep stage synchronisation, increased REM sleep (Drews et al., 2020), and better sleep quality resulting from co-sleeping in comparison with individual sleeping (Drews et al., 2017). Related studies have also used simulated data due to the complexity of annotating 2-snorer sleep audio recordings: in addition to marking breathing events, they have to be assigned to a particular snorer (Mordoh and Zigel, 2021; Nigam and Priemer, 2006). To the best of our knowledge, the snorer diarisation system proposed here is the only one that has been nevertheless tested on a limited amount of real 2-snorer data.

Specialised Hardware

The system that integrates acoustic features with oxygen saturation to screen for OSA requires specialised hardware, as this physiological parameter is obtained from a pulse oximeter. This hinders wide use of such an approach. However, the screening system that relies on acoustic evidence alone can be used as a fallback when oxygen saturation information is not available. Alternatively, readily available hardware capable of measuring oxygen saturation, like a smartwatch (Phillips et al., 2019), might be used for obtaining the required information. The feasibility and reliability of this kind of hardware remains to be investigated.

7.3 Scope for Future Work

More Data

The developed systems might be further improved with more data, as the performance of deep learning applications greatly depends on the quality and quantity of data available for their development (Wallis, 2019). For example, collecting female snoring would allow the differences between male and female snoring to be investigated, which might further inform the development of systems for the classification of sleep-disordered breathing events. Collecting 2-snorer sleep audio recordings would improve the performance of the proposed snorer diarisation system on real data. Although several considerations were taken to approximate the simulated data to the expected real conditions, there was an inevitable mismatch between the simulated data used for training and the real data used for testing. Retraining the snorer diarisation system with real 2-snorer data would overcome this mismatch. At the time of writing this thesis, more audio recordings have been collected during HSAT. However, the data collection effort was significantly delayed due to the COVID-19 pandemic and not used here.

Other Deep Learning Techniques

More data would also allow the application of state-of-the-art deep learning techniques. These can model complex relations between the input and output but commonly require much more data and computational resources than long-established techniques (Yu et al., 2018). For instance, attention models (Vaswani et al., 2017) might be used for efficiently exploiting the temporal pattern of respiration to screen for OSA, as they would learn to attend to the relevant parts of the audio signal. End-to-end systems (Zinemanas et al., 2019) would couple the feature extraction and classification steps into a single framework. This can be optimised as a whole instead of part by part, which is currently a limitation, and would dispense with the feature selection process.

Source Separation Techniques

Source separation techniques for snorer diarisation might be revisited. An iOS API that allows the use of multiple microphones in one device for audio recording is now available (Apple Inc., 2021). Although its capabilities remain to be investigated, it would potentially facilitate the collection of sleep audio recordings with spatially distributed microphones using readily available hardware. Recent Apple iPhones have four microphones: one on the front, one on the back, and two on the bottom of the device separated by approximately 20 mm. The latter could possibly be used to collect multichannel sleep audio recordings. Making use of the collected data, multichannel source separation techniques like blind source separation (Ochi et al., 2016) could be reconsidered to develop a more robust snorer diarisation system, since there would be as many sensors as snorers, and more information would be available to exploit. Given that the microphones are physically connected, accurate synchronisation would not be an issue. Other techniques such as beamforming (Anguera et al., 2007), and localisation-based grouping (Feng and Jones, 2006) could also be reconsidered. However, placing the smartphone on a bedside table would result in the sensors and the snorers being on the same axis, which might be problematic for the last two techniques, as they expect the sound sources to be in front of the sensors. Additional considerations should be taken for the placement of the smartphone, and the fact that people move during sleep.

On-device Deep Learning and Sensor Integration

Although the sleep audio recordings used in this study were made with smartphones, those were processed and analysed in computers. On-device deep learning was briefly considered. It has been insisted throughout this thesis on the challenge of improving accessibility to sleep-disordered breathing diagnosis (Dafna et al., 2013; Phillips, 2007), and on-device processing might potentially play a role in overcoming it, since only readily available hardware (e.g., a smartphone) would be needed to screen for sleep-disordered breathing. On-device deep learning also preserves user privacy, since the user's data strictly remains on their device (Kaissis et al., 2020).

As briefly mentioned before, smartphones can be integrated with other readily available hardware, for example, smartwatches. They commonly feature sensors to measure physiological parameters — like oxygen saturation and heart rate — and track body movement for general fitness purposes. Such additional parameters might be used to further inform the developed systems, as it was done with oxygen saturation in Chapter 5. For instance, *sleep* time can be derived from these, and be used instead of *recording* time to estimate the AHI, which would potentially result in more accurate figures. Furthermore, the developed systems can be used as part of a health monitoring ecosystem,

e.g., bringing together activity monitoring during the day and sleep as a broader picture of health.

7.4 Epilogue

This thesis was produced during the COVID-19 pandemic. COVID-19 has spread around the world since December 2019, and brought unparalleled challenges to healthcare services. Since SARS-CoV-2, the coronavirus that causes COVID-19, is transmitted through close contact, and airborne aerosols and droplets, non-urgent procedures and diagnostic assessments have been cancelled or postponed. This has been done to protect patients and healthcare workers by avoiding potential contamination of laboratories and hospitals as well as unnecessary exposure. Sleep laboratories have deferred in-laboratory sleep studies and follow-up examinations. On the other hand, there is an important overlap between the risk factors for COVID-19 and OSA, for instance, male sex, age, excess weight, hypertension, asthma and diabetes. It is likely that the COVID-19 pandemic will bring about a significant and lasting change to sleep medicine. Medical care in sleep medicine should be provided with minimum contact, and move from sleep laboratories to the home (Voulgaris et al., 2020; Williamson, 2020). Therefore, the technology developed here has real practical importance. By allowing monitoring over extended time scales, it could also help to differentiate medical issues related to long COVID-19 from those due to sleep disorders.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *2016 Symposium on Operating Systems Design and Implementation (OSDI)*, pages 265–283, Savannah, United States. USENIX.

Adavanne, S., Parascandolo, G., Pertilä, P., Heittola, T., and Virtanen, T. (2016). Sound event detection in multichannel audio using spatial and harmonic features. In *2016 Detection and Classification of Acoustic Scenes and Events (DCASE)*, Budapest, Hungary. DCASE.

Al-Mardini, M., Aloul, F., Sagahyroon, A., and Al-Husseini, L. (2014). Classifying obstructive sleep apnea using smartphones. *Journal of Biomedical Informatics*, 52:251–259.

Alakuijala, A. and Salmi, T. (2016). Predicting obstructive sleep apnea with periodic snoring sound recorded at home. *Journal of Clinical Sleep Medicine*, 12(7):953–958.

Albawi, S., Mohammed, T. A., and Al-Azawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, Antalaya, Turkey. IEEE.

Almazaydeh, L., Elleithy, K., Faezipour, M., and Abushakra, A. (2013). Apnea detection based on respiratory signal classification. *Procedia Computer Science*, 21:310–316.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., and Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8(292):1–67.

American Academy of Sleep Medicine (2020). *AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications (Version 2.6)*.

Andres, E., Gass, R., Charloux, A., Brandt, C., and Hentzler, A. (2018). Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0. *Journal of Medicine and Life*, 11(2):89–106.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: a review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

Anguera, X., Wooters, C., and Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022.

Apple Inc. (2021). AVAudioSession – Capturing stereo audio from built-in microphones. https://developer.apple.com/documentation/avfaudio/avaudiosession/capturing_stereo_audio_from_built-in_microphones.

Bengio, Y. (2012). *Neural Networks: Tricks of the Trade*, chapter Practical Recommendations for Gradient-based Training of Deep Architectures, pages 437–478. Springer.

Bishop, C. M. (2006a). *Pattern Recognition and Machine Learning*, chapter Mixture of Gaussians, pages 110–113. Springer.

Bishop, C. M. (2006b). *Pattern Recognition and Machine Learning*, chapter Linear Models for Classification, pages 179–224. Springer.

Blackman, R. B. and Tukey, J. W. (1958). The measurement of power spectra from the point of view of communications engineering. *The Bell System Technical Journal*, 37(1):185–282.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310.

Botelho, M. C., Trancoso, I., Abad, A., and Paiva, T. (2019). Speech as a biomarker for obstructive sleep apnea detection. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5851–5855, Brighton, United Kingdom. IEEE.

Breasted, J. H. (1980). *The Edwin Smith Surgical Papyrus*. The University of Chicago Press.

Bridle, J. S. and Brown, M. D. (1974). An experimental automatic word recognition system. *Joint Speech Research Unit Report*, 1003(5):33.

Brown, G. J. and Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336.

Brunese, L., Martinelli, F., Mercaldo, F., and Santone, A. (2020). Deep learning for heart disease detection through cardiac sounds. *Procedia Computer Science*, 176:2202–2211.

Campbell, S. (2013). A short history of sonography in obstetrics and gynaecology. *Facts, Views and Vision in ObGyn*, 5(3):213–229.

Castaneda, A., Jauregui-Maldonado, E., Ratnani, I., Varon, J., and Surani, S. (2018). Correlation between metabolic syndrome and sleep apnea. *World Journal of Diabetes*, 9(4):66–71.

Castillo-Escario, Y., Ferrer-Lluis, I., Montserrat, J. M., and Jane, R. (2019). Entropy analysis of acoustic signals recorded with a smartphone for detecting apneas and hypopneas: a comparison with a commercial system for home sleep apnea diagnosis. *IEEE Access*, 7:128224–128241.

Caylak, E. (2013). *Encyclopedia of Sleep*, chapter Familial and Genetic Factors, pages 179–183. Academic Press.

Chan, J., Raju, S., Nandakumar, R., Bly, R., and Gollakota, S. (2019). Detecting middle ear fluid using smartphones. *Science Translational Medicine*, 11.

Chiba, T. and Kajiyama, M. (1941). *The vowel. Its Nature and Structure*. Phonetic Society of Japan, Tokyo, Japan.

Cho, E. R., Kim, H., Seo, H. S., Suh, S., Lee, S. K., and Shin, C. (2013). Obstructive sleep apnea as a risk factor for silent cerebral infarction. *Journal of Sleep Research*, 22:452–458.

Chokroverty, S. and Avidan, A. Y. (2016). *Bradley's Neurology in Clinical Practice*, chapter Sleep and Its Disorders, pages 1615–1685. Number 102. Elsevier, seventh edition.

Chuang, L. P., Lin, S. W., Lee, L. A., Li, H. Y., Chang, C. H., Kao, K. C., Li, L. F., Huang, C. C., Yang, C. T., and Chen, N. H. (2017). The gender difference of snore distribution and increased tendency to snore in women with menopausal syndrome: a general population study. *Sleep and Breathing*, 21(2):543–547.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Colten, H. R. and Altevogt, B. M., editors (2006a). *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*, chapter Sleep Physiology, pages 33–53. Institute of Medicine (US) Committee on Sleep Medicine and Research.

Colten, H. R. and Altevogt, B. M., editors (2006b). *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*, chapter Extent and Health Consequences of Chronic Sleep Loss and Sleep Disorders, pages 55–135. Institute of Medicine (US) Committee on Sleep Medicine and Research.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36:287–314.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Cosi, P., Frigo, L., Mian, G. A., and Sinigallia, T. (1984). On the use of autocorrelation for pitch extraction: some statistical considerations and their application to the SIFT algorithm. *Speech Communication*, 3:309–334.

Coy, A. and Barker, J. (2007). An automatic speech recognition system based on the scene analysis account of auditory perception. *Speech Communication*, 49(5):384–401.

Creswell, A., Arulkumaran, K., and Bharath, A. A. (2017). On denoising autoencoders trained to minimise binary cross-entropy. *ArXiv*.

Cummins, N., Baird, A., and Schuller, B. W. (2018). Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54.

Dafna, E., Tarasiuk, A., and Zigel, Y. (2013). Automatic detection of whole night snoring events using non-contact microphone. *PLoS ONE*, 8(12).

De Cheveigne, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930.

Dean, J. (2020). The deep learning revolution and its implications for computer architecture and chip design. In *2020 International Solid-State Circuits Conference (ISSCC)*, San Francisco, United States. IEEE.

Demir, F., Sengur, A., Cummins, N., Amiriparian, S., and Schuller, B. (2018). Low level texture features for snore sound discrimination. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 413–416, Honolulu, United States. IEEE.

Donald, I., MacVicar, J., and Brown, T. G. (1958). Investigation of abdominal masses by pulsed ultrasound. *The Lancet*, 271(7032):1188–1195.

Drews, H. J., Wallot, S., Brysch, P., Berger-Johannsen, H., Weinhold, S. L., Mitkidis, P., Baier, P. C., Lechinger, J., Roepstorff, A., and Göder, R. (2020). Bed-sharing in couples is associated with increased and stabilized REM sleep and sleep-stage synchronization. *Frontiers in Psychiatry*, 11(583):1–12.

Drews, H. J., Wallot, S., Weinhold, S. L., Mitkidis, P., Baier, P. C., Roepstorff, A., and Göder, R. (2017). Are we in sync with each other? Exploring the effects of cosleeping on heterosexual couples' sleep using simultaneous polysomnography: a pilot study. *Sleep Disorders*, 2017:1–5.

Duckitt, W. D., Tuomi, S. K., and Niesler, T. R. (2016). Automatic detection, segmentation and assessment of snoring from ambient acoustic data. *Physiological Measurement*, 27:1047–1056.

Durning, S. J., Capaldi, V. F., Artino, A. R., Graner, J., van der Vleuten, C., Beckman, T. J., Costanzo, M., Holmboe, E., and Schuwirth, L. (2014). A pilot study exploring the relationship between internists' self-reported sleepiness, performance on multiple-choice exam items and prefrontal cortex activity. *Medical Teacher*, 36(5):434–440.

Ekirch, A. R. (2006). *At Day's Close: Night in Times Past*. W. W. Norton and Company, London, United Kingdom.

Elisha, O., Tarasiuk, A., and Zigel, Y. (2012). Automatic detection of obstructive sleep apnea using speech signal analysis. In *Speech Processing Conference 2012*, pages 20–23, Tel Aviv, Israel. ISCA.

Emoto, T., Abeyratne, U. R., Kawano, K., Okada, T., Jinnouchi, O., and Kawata, I. (2018). Detection of sleep breathing sound based on artificial neural network analysis. *Biomedical Signal Processing and Control*, 41:81–89.

Fant, G. (1960). *Acoustic Theory of Speech Production*, chapter Source-Filter Description of Speech Production, pages 15–20. Mouton and Co. N. V., The Hague, The Netherlands.

Farnsworth, D. W. (1942). Radiation pattern of the human voice. *The Scientific Monthly*, 55(2):139–143.

Feng, A. S. and Jones, D. L. (2006). *Computational Auditory Scene Analysis*, chapter Localization-Based Grouping, pages 187–208. IEEE.

Fernandez, R., Blanco, J. L., Hernandez, L., Lopez, E., Alcazar, J., and Toledano, D. T. (2009). Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker Recognition Techniques. *EURASIP Journal on Advances in Signal Processing*.

Fiz, J. A., Abad, J., Jané, R., Riera, M., Mañanas, M. A., Caminal, P., Rodenstein, D., and Morera, J. (1996). Acoustic analysis of snoring in patients with simple snoring and obstructive sleep apnoea. *European Respiratory Journal*, 9:2365–2370.

Fiz, J. A., Morera, J., Abad, J., Belsunces, A., Haro, M., Fiz, J. I., Jane, R., Caminal, P., and Rodenstein, D. (1993). Acoustic Analysis of Vowel Emission on Obstructive Sleep Apnea. *Chest*, 104(4):1093–1096.

Flanagan, J. L. (1960). Analog measurements of sound radiation from the mouth. *Journal of the Acoustical Society of America*, 32(12):1613–1620.

Freiherr, G. (1980). *The Seeds of Artificial Intelligence. SUMEX-AIM*. National Institutes of Health.

Freund, Y. and Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780.

Gales, M. and Young, S. (2007). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304.

Gaubitch, N., Kleijn, W., and Heusdens, R. (2013). Auto-localization in ad-hoc microphone arrays. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 106–110, Vancouver, Canada. IEEE.

Giannakopoulos, T. and Pikrakis, A. (2014). *Introduction to Audio Analysis*, chapter Audio Features, pages 59–103. Academic Press, Oxford, United Kingdom.

Girithari, G., Santos, I. C., Claro, E., Belykh, S., Matias, D., and Santos, O. (2018). The hippocratic splash. *European Journal of Case Reports in Internal Medicine*, 5(11).

Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*, chapter Deep Feedforward Networks, pages 164–223. MIT Press, Cambridge, United States.

Gutierrez, J., Fraile, R., Camacho, A., Durand, T., Jarrin, J., and Mendoza, S. (2016). Synthetic sound event detection based on MFCC. In *2016 Detection and Classification of Acoustic Scenes and Events (DCASE)*, Budapest, Hungary. DCASE.

Habibzadeh, F., Habibzadeh, P., and Yadollahie, M. (2016). On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia Medica*, 26(3):297–307.

Hajar, R. (2012). The art of listening. *Heart Views*, 13(1):24–25.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*, chapter Data Preprocessing, pages 83–124. Elsevier, Waltham, United States.

Hess, W. (1983). *Pitch Determination of Speech Signals: Algorithms and Devices*, chapter Short-term Analysis Pitch Determination, pages 343–470. Springer-Verlag, Berlin, Germany.

Hess, W. J. (2008). *Springer Handbook of Speech Processing*, chapter Pitch and Voicing Determination of Speech with an Extension Toward Music Signals, pages 181–212. Springer, Berlin, Heidelberg, Germany.

Himawan, I., McCowan, I., and Sridharan, S. (2011). Clustered blind beamforming from ad-hoc microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):661–676.

Hochreiter, S. and Schmidhüber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hoffstein, V., Mateika, J. H., and Mateika, S. (1991). Snoring and sleep architecture. *American Review of Respiratory Disease*, 143(1):92–96.

Hope, K. (2016). Sleep deprivation "costs UK £40bn a year". https://www.bbc.co.uk/news/business-38151180.

Hossain, J. L. and Shapiro, C. M. (2002). The prevalence, cost implications, and management of sleep disorders: an overview. *Sleep Breath*, 6:85–102.

Hu, M., Parada, P. P., Sharma, D., Doclo, S., van Waterschoot, T., Brookes, M., and Naylor, P. A. (2015). Single-channel speaker diarization based on spatial features. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, New Paltz, United States. IEEE.

Huang, L. (1995). Mechanical modeling of palatal snoring. *Journal of the Acoustical Society of America*, 97(6):3642–3648.

Huang, L., Quinn, S. J., Ellis, P. D. M., and Williams, J. E. F. (1995). Biomechanichs of snoring. *Endevour*, 96(3):96–100.

Hughes, J. (1988). The Edwin Smith Surgical Papyrus: an analysis of the first case reports of spinal cord injuries. *Spinal Cord*, 26:71–82.

Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, M. S., Ali, K., John, C. N., Iftikhar, M. D., and Nabeel, M. (2020). AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Informatics in Medicine Unlocked*, 20.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456, Lille, France.

Iriarte, J., Campo, A., Alegre, M., Fernández, S., and Urrestarazu, E. (2015). Catathrenia: respiratory disorder or parasomnia? *Sleep Medicine*, 16(7):827–830.

Janjua, Z. H., Vecchio, M., Antonini, M., and Antonelli, F. (2019). IRESE: An intelligent rare-event detection system using unsupervised learning on the IoT edge. *Engineering Applications of Artificial Intelligence*, 84:41–50.

Janott, C., Schmitt, M., Zhang, Y., Qian, K., Pandit, V., Zhang, Z., Heiser, C., Hohenhorst, W., Herzog, M., Hemmert, W., and Schuller, B. (2018). Snoring classified: the Munich-Passau Snore Sound Corpus. *Computers in Biology and Medicine*, 94:106–118.

Japkowicz, N. and Shah, M. (2011a). *Evaluating Learning Algorithms: a Classification Perspective*, chapter Machine Learning and Statistics Overview, pages 23–73. Cambridge University Press.

Japkowicz, N. and Shah, M. (2011b). *Evaluating Learning Algorithms: a Classification Perspective*, chapter Performance Measures I, pages 74–110. Cambridge University Press.

Japkowicz, N. and Shah, M. (2011c). *Evaluating Learning Algorithms: a Classification Perspective*, chapter Performance Measures II, pages 111–160. Cambridge University Press.

Jebara, T. (2002). *Discriminative, Generative and Imitative Learning*. PhD thesis, School of Architecture and Planning, Massachusetts Institute of Technology.

Jurafsky, D. and Martin, J. H. (2014a). *Speech and language processing*, chapter Automatic Speech Recognition, pages 291–342. Pearson Education.

Jurafsky, D. and Martin, J. H. (2014b). *Speech and language processing*, chapter N-Grams, pages 85–124. Pearson Education.

Kaissis, G. A., Makowski, M. R., and Ruckert, D. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2:305–311.

Kaur, J., Singh, A., and Kadyan, V. (2020). Automatic speech recognition system for tonal languages: state-of-the-art survey. *Archives of Computational Methods in Engineering*.

Kaya, H. and Karpov, A. A. (2017). Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: snoring, addressee and cold. In *Proceedings of INTERSPEECH 2017*, pages 3527–3531, Stockholm, Sweden. ISCA.

Kezirian, E. J., Hohenhorst, W., and de Vries, N. (2011). Drug-induced sleep endoscopy: the VOTE classification. *European Archives of Oto-Rhino-Laryngology*, 268:1233–1236.

Kim, R. D., Kapur, V. K., Redline-Bruch, J., Rueschman, M., Auckley, D. H., Benca, R. M., Foldvary-Schafer, N. R., Iber, C., Zee, P. C., Rosen, C. L., Redline, S., and Ramsey, S. D. (2015). An economic evaluation of home versus laboratory-based diagnosis of obstructive sleep apnea. *Sleep*, 38(7):1027–1037.

Kim, T., Kim, J. W., and Lee, K. (2018). Detection of sleep-disordered breathing severity using acoustic biomarker and machine learning techniques. *Biomedical Engineering Online*, 17.

Knapp, C. and Carter, G. (1976). The generalized correlation method for estimation of time delay. *Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23:89–109.

Krenker, A., Bester, J., and Kos, A. (2003). *Artificial neural networks. Methodological advances in biomedical applications*, chapter Introduction to the artificial neural networks, pages 3–18. Number 1. Elsevier Science.

Kulikowski, C. (2019). Beginnings of Artificial Intelligence in Medicine (AIM): Computational Artifice Assisting Scientific Inquiry and Clinical Art - with Reflections on Present AIM Challenges. *IMIA Yearbook of Medical Informatics*, 28(1):249–256.

Kumar, R. (2020). Cerebral LSTM: a better alternative for single- and multi-stacked LSTM cell-based RNNs. *Springer Nature Computer Science*, 1(85):1–8.

Kurtic, E., Brown, G. J., and Wells, B. (2013). Resources for turn competition in overlapping talk. *Speech Communication*, 55(5):721–743.

Kwong, S. and He, Q. (2001). The use of adaptive frame for speech recognition. *EURASIP Journal on Applied Signal Processing*, 2:82–88.

Laënnec, R. (1821). *A Treatise on the Diseases of the Chest in Which They Are Described According to Their Anatomical Characters and Their Diagnosis*. T. and G. Underwood, London.

Lal, C., Strange, C., and Bachman, D. (2012). Neurocognitive impairment in obstructive sleep apnea. *Chest*, 141(6):1601–1610.

Lasserre, J. A., Bishop, C. M., and Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 87–94, New York, United States. IEEE.

Latif, S., Qadir, J., Qayyum, A., Usama, M., and Younis, S. (2020). Speech technology for healthcare: opportunities, challenges, and state-of-the-art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K. R. (2012). *Neural Networks: Tricks of the Trade*, chapter Efficient BackProp, pages 9–48. Springer.

Lee, G., Lee, L., Wang, C., Chen, N., Fang, T., Huang, C., Cheng, W., and Li, H. (2016). The frequency and energy of snoring sounds are associated with common carotid artery intima-media thickness in obstructive sleep apnea patients. *Nature Scientific Reports*.

Lent, R., Azevedo, F. A. C., Andrade-Moraes, C. H., and Pinto, A. V. O. (2012). How many neurons do you have? Some dogmas of quantitative neuroscience under revision. *European Journal of Neuroscience*, 35:1–9.

Levartovsky, A., Dafna, E., Zigel, Y., and Tarasiuk, A. (2016). Breathing and snoring sound characteristics during sleep in adults. *Journal of Clinical Sleep Medicine*, 12(3):375–384.

Liu, Z. S., Luo, X. Y., Lee, H. P., and Lu, C. (2007). Snoring source identification and snoring noise prediction. *Journal of Biomechanics*, 40:861–870.

Lockwood, M. E., Jones, D. L., Bilger, R. C., Lansing, C. R., O'Brien, W. D., Wheeler, B. C., and Feng, A. S. (2004). Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. *Journal of the Acoustical Society of America*, 115(1):379–391.

Lu, W., Cantor, J., Aurora, R. N., Nguyen, M., Ashman, T., Spielman, L., Ambrose, A., Kerllman, J. W., and Gordon, W. (2014). Variability of respiration and sleep during polysomnography in individuals with TBI. *Neurorehabilitation*, 35:245–251.

Ma, Y., Peng, L., Kou, C., Hua, S., and Yuan, H. (2017). Associations of overweight, obesity and related factors with sleep-related breathing disorders and snoring in adolescents: a cross-sectional survey. *International Journal of Environmental Research and Public Health*, 14(2):1–10.

Maas, J. B., Robbins, R. S., Fortgang, R. G., and Driscoll, S. R. (2011). *Encyclopedia of Adolescence*, chapter Adolescent Sleep, pages 56–65. Academic Press.

Maimon, N. and Hanly, P. J. (2010). Does snoring intensity correlate with the severity of obstructive sleep apnea? *Journal of Clinical Sleep Medicine*, 6(5):475–478.

Malhotra, A., Younes, M., Kuna, S. T., Benca, R., Kushida, C. A., Walsh, J., Hanlon, A., Staley, B., Pack, A. I., and Pien, G. W. (2013). Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep*, 36(4):573–582.

Mansukhani, M. P., Kolla, B. P., Kent, E., and Morgenthaler, T. I. (2020). *Conn's Current Therapy*, chapter Sleep Disorders, pages 739–753. Elsevier, Philadelphia, PA.

Mantua, J. and Simonelli, G. (2019). Sleep duration and cognition: is there an ideal amount? *Sleep*, 42(3):1–3.

Marchi, E., Shum, S., Hwang, K., Kajarekar, S., Sigtia, S., Richards, H., Haynes, R., Kim, Y., and Bridle, J. (2018). Generalised discriminative transform via curriculum learning for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5324–5328. IEEE.

McLeod, P. and Geoff, W. (2005). A smarter way to find pitch. In *2005 International Computer Music Conference (ICMC)*.

Meijerink, L., Cina, G., and Tonutti, M. (2020). Uncertainty estimation for classification and risk prediction on medical tabular data. *ArXiv*, pages 1–15.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. In Chen, C. H., editor, *Pattern Recognition and Artificial Intelligence*, pages 374–388, New York, United States.

Messner, E., Fediuk, M., Swatek, P., Scheidl, S., Smolle-Jüttner, F. M., Olschewski, H., and Pernkopf, F. (2020). Multi-channel lung sound classification with convolutional recurrent neural networks. *Computers in Biology and Medicine*, 122.

Middendorp, J. J., Sanchez, G. M., and Burridege, A. L. (2010). The Edwin Smith papyrus: a clinical reappraisal of the oldest known document on spinal injuries. *European Spine Journal*, 19:1815–1823.

Miller, R. A., Pople, H. E., and Myers, J. D. (1985). *Computer-Assisted Medical Decision Making*, chapter INTERNIST-I. An Experimental Computer-based Diagnostic Consultant for General Internal Medicine, pages 139–158. Computers and Medicine. Springer, New York, NY.

Miyabe, S., Ono, N., and Makino, S. (2015). Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation. *Signal Processing*, 107:185–196.

Montinari, M. R. and Minelli, S. (2019). The first 200 years of cardiac auscultation and future perspectives. *Journal of Multidisciplinary Healthcare*, 12:183–189.

Moody, J. (2012). *Neural Networks: Tricks of the Trade*, chapter Forecasting the Economy with Neural Nets: a Survey of Challenges and Solutions, pages 343–367. Springer.

Mordoh, V. and Zigel, Y. (2021). Audio source separation to reduce sleeping partner sounds: a simulation study. *Physiological Measurement*.

Motamedi, K. K., McClary, A. C., and Amedee, R. G. (2009). Obstructive sleep apnea: a growing problem. *Ochsner Journal*, 9:149–153.

Murphy, K. P. (2012a). *Machine Learning: a Probabilistic Perspective*, chapter Logistic Regression, pages 245–279. MIT Press.

Murphy, K. P. (2012b). *Machine Learning: a Probabilistic Perspective*, chapter Deep Learning, pages 245–279. MIT Press.

Murphy, K. P. (2012c). *Machine Learning: a Probabilistic Perspective*, chapter Kernels, pages 245–279. MIT Press.

Muza, R. T. (2015). Central sleep apnoea – a clinical review. *Journal of Thoracic Disease*, 7(5):930–937.

Nakano, H., Furukawa, T., and Tanigawa, T. (2019). Tracheal sound analysis using a deep neural network to detect sleep apnea. *Journal of Clinical Sleep Medicine*, 15(8):1125–1133.

Narayan, S., Shivdare, P., Niranjan, T., Williams, K., Freudman, J., and Sehra, R. (2018). Noncontact identification of sleep-disturbed breathing from smartphone-recorded sounds validated by polysomnography. *Sleep and Breathing*, 23(1):269–279.

NHS (2020). 2020/21 Annex A: The national tariff workbook. https://www.england.nhs.uk/wp-content/uploads/2021/02/20-21NT_Annex_A_National_tariff_workbook.xlsx.

Nigam, V. and Priemer, R. (2006). A snore extraction method from mixed sound for a mobile snore recorder. *Journal of Medical Systems*, 30:91–99.

Nonaka, R., Emoto, T., Abeyratne, U. R., Jinnouchi, O., Kawata, I., Ohnishi, H., Akutagawa, M., Konaka, S., and Kinouchi, Y. (2016). Automatic snore sound extraction from sleep sound recordings via auditory image modeling. *Biomedical Signal Processing and Control*, 27:7–14.

Nuttall, F. Q. (2015). Body mass index. Obesity, BMI, and health: a critical review. *Nutrition Today*, 50(3):117–128.

Ochi, K., Ono, N., Miyabe, S., and Makino, S. (2016). Multi-talker speech recognition based on blind source separation with ad hoc microphone array using smartphones and cloud storage. In *Proceedings of INTERSPEECH 2016*, pages 3369–3373, San Francisco, United States. ISCA.

Olsson, R. K. and Hansen, L. K. (2006). Blind separation of more sources than sensors in convolutive mixtures. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 657–660, Toulouse, France. IEEE.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. *ArXiv*, pages 1–11.

Partinen, M., Guilleminault, C., Quera-Salva, M. A., and Jamieson, A. (1988). Obstructive sleep apnea and cephalometric roentgenograms: the role of anatomic upper airway abnormalities in the definition of abnormal breathing during sleep. *Chest*, 93(6):1199–1205.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992). Complex sounds and auditory images. In Cazals, Y., Demany, L., and Horner, K., editors, *Auditory physiology and perception, Proc. 9th International Symposium on Hearing*, pages 429–446.

Petersen-George, C. F. (1999). Diagnostic techniques in obstructive sleep apnea. *Progress in Cardiovascular Diseases*, 41(5):355–366.

Pevernagie, D., Aarts, R., and Meyer, M. D. (2010). The acoustics of snoring. *Sleep Medicine Reviews*, 14:131–144.

Phillips, B. (2007). Improving access to diagnosis and treatment of sleep-disordered breathing. *Chest*, 132(5):1418–1420.

Phillips, C., Liaqat, D., Gabel, M., and de Lara, E. (2019). Reliable peripheral oxygen saturation readings from wrist-worn pulse oximeters. *ArXiv*, pages 1–15.

Qaisar, S. M. (2019). Isolated speech recognition and its transformation in visual signs. *Journal of Electrical Engineering and Technology*, 14:955–964.

Rahali, H., Hajaiej, Z., and Ellouze, N. (2014). ASR systems in noisy environment: auditory features based on gammachirp filter using the AURORA database. In *22nd European Signal Processing Conference (EUSIPCO)*, pages 696–700, Lisbon, Portugal. IEEE.

Rao, K. S. and Vuppala, A. K. (2014). *Speech Processing in Mobile Environments*, chapter Appendix A, pages 103–106. Springer International Publishing, Switzerland.

Rao, M. V. A. and Ghosh, P. K. (2017). Pitch prediction from mel-frequency cepstral coefficients using sparse spectrum recovery. In *2017 Twenty-third National Conference on Communications (NCC)*, pages 1–6, Chennai, India. IEEE.

Rath, S. P., Knill, K. M., Ragni, A., and Gales, M. J. F. (2014). Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages. In *Proceedings of INTERSPEECH 2014*, pages 835–839, Singapore, Singapore. ISCA.

Rhodes, D. (2017). Sleep disorder testing carried out by NHS doubles. https://www.bbc.co.uk/news/uk-england-40122979.

Richter, K., Adam, S., Geiss, L., Peter, L., and Niklewski, G. (2006). Two in a bed: the influence of couple sleeping and chronotypes on relationship and sleep. An overview. *Chronobiology International*, 33(10):1464–1472.

Risoud, M., Hanson, J. N., Gauvrit, F., Renard, C., Lemesre, P. E., Bonne, N. X., and Vincent, C. (2018). Sound source localization. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 135(4):259–264.

Robb, M. P., Yates, J., and Morgan, E. J. (1997). Vocal tract resonance characteristics of adults with obstructive sleep apnea. *Acta Oto-Laryngologica*, 117(5):760–763.

Rofouei, M., Sinclair, M., Bittner, R., Blank, T., Saw, N., DeGean, G., and Heffron, J. (2011). A non-invasive wearable neck-cuff system for real-time sleep monitoring. In *2011 International Conference on Body Sensor Networks*, pages 159–161, Dallas, United States. IEEE.

Roman, N., Wang, D., and Brown, G. J. (2003). Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114(4):2236–2252.

Romero, H. E., Ma, N., and Brown, G. J. (2020a). Acoustic characterisation of inhalation and exhalation. In *Acoustics 2020*, Chester, United Kingdom. Institute of Acoustics.

Romero, H. E., Ma, N., and Brown, G. J. (2020b). Snorer diarisation based on deep neural network embeddings. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 876–880, Barcelona, Spain. IEEE.

Romero, H. E., Ma, N., Brown, G. J., Beeston, A. V., and Hasan, M. (2019). Deep learning features for robust detection of acoustic events in sleep-disordered breathing. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 810–814, Brighton, United Kingdom. IEEE.

Romero, H. E., Ma, N., Hill, E. A., and Brown, G. J. (2020c). Screening for obstructive sleep apnea at home based on deep learning features derived from respiration sounds. In *Sleep*, volume 43, pages A219–A220, Philadelphia, United States. Associated Professional Sleep Societies, Oxford University Press.

Rosen, I. M., Kirsch, D. B., Carden, K. A., Malhotra, R. K., Ramar, K., Aurora, R. N., Kristo, D. A., Martin, J. L., Olson, E. J., Rosen, C. L., Rowley, J. A., and Shelgikar, A. V. (2018). Clinical use of a home sleep apnea test: an updated american academy of sleep medicine position statement. *Journal of Clinical Sleep Medicine*, 14(12):2075–2077.

Rosenblatt, M. (1974). *Random Processes. Graduate Texts in Mathematics*, volume 17, chapter Markov Chains, pages 36–67. Springer, New York, United States.

Ruiz Villareal, M. (2007). Respiratory system. Public Domain.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Saha, S., Kabir, M., Ghahjaverestan, N. M., Hafezi, M., anf K. Zhu, B. G., Alshaer, H., and Yadollahi, A. (2020). Portable diagnosis of sleep apnea with the validation of individual event detection. *Sleep Medicine*, 69:51–57.

Schmidhüber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, 61:85–117.

Schröder, J., Anemüller, J., and Goetze, S. (2016). Peformance comparison of GMM, HMM and DNN based approaches for acoustic event detection within task 3 of the DCASE 2016 challenge. In *2016 Detection and Classification of Acoustic Scenes and Events (DCASE)*, Budapest, Hungary. DCASE.

Scott, M. J. J., Niranjan, M., Melvin, D. G., and Prager, R. W. (1998). Maximum realisable performance: a principled method for enhancing performance by using multiple classifiers in variable cost problem domains. In *1998 British Machine Vision Conference*, Southampton, United Kingdom. British Machine Vision Association.

Shakespeare, W. (2014). *The Complete Works of Shakespeare*, chapter Henry IV. Public Domain.

Shao, X. and Milner, B. (2005). Predicting fundamental frequency from mel-frequency cepstral coefficients to enable speech reconstruction. *Journal of the Acoustical Society of America*, 118(2):1134–1143.

Sherwood, L. (2016). *Human Physiology: from Cells to Systems*, chapter The Respiratory System. Cengage Learning, Australia, 9th edition.

Shokoueinejad, M., Fernandez, C., Carroll, E., Wang, F., Levin, J., Rusk, S., Glattard, N., Mulchrone, A., Zhang, X., Xie, A., Teodorescu, M., Dempsey, J., and Webster, J. (2017). Sleep apnea: a review of diagnostic sensors, algorithms, and therapies. *Physiological Measurement*, 38:204–252.

Shortliffe, E. H. (2019). Artificial intelligence in medicine: weighing the accomplishments, hype, and promise. *IMIA Yearbook of Medical Informatics*, 28(1):257–262.

Simpson, A. P. (2001). Dynamic consequences of differences in male and female vocal tract dimensions. *Journal of the Acoustical Society of America*, 109(5):2153–2164.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Starke, S., Leger, S., Zwanenburg, A., Leger, K., Lohaus, F., Linge, A., Schreiber, A., Kalinauskaite, G., Tinhofer, I., Guberina, N., Guberina, M., Balermpas, P., von der Grün, J., Ganswindt, U., Belka, C., Peeken, J. C., Combs, S. E., Boeke, S., Zips, D., Richter, C., Troost, E. G. C., Krause, M., Baumann, M., and Löck, S. (2020). 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma. *Nature Scientific Reports*, 10(15625).

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190.

Stiefel, M., Shaner, A., and Schaefer, S. D. (2006). The Edwin Smith Papyrus: the Birth of Analytical Thinking in Medicine and Otolaryngology. *The Laryngoscope*, 116:182–188.

Story, B. (2016). *The Oxford Handbook of Singing*, chapter The Vocal Tract in Singing. Oxford University Press.

Stoter, F. R., Chakrabarty, S., Edler, B., and Habets, E. A. P. (2018). Classification vs. regression in supervised learning for single channel speaker count estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada. IEEE.

Sun, X., Kim, J. Y., Won, Y., Kim, J. J., and Kim, K. A. (2015). Efficient snoring and breathing detection based on sub-band spectral statistics. *Bio-Medical Materials and Engineering*, 26:S787–S793.

Thiemann, J. and Vincent, E. (2013). An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–5, Southampton, United Kingdom. IEEE.

Tiron, R., Lyon, G., Kilroy, H., Osman, A., Kelly, N., O'Mahony, N., Lopes, C., Coffey, S., McMahon, S., Wren, M., Conway, K., Fox, N., Costello, J., Shouldice, R., Lederer, K., Fietze, I., and Penzel, T. (2020). Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology. *Journal of Thoracic Disease*, 12(8):4476–4495.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59:433–460.

Ursavas, A., Ercan, I., Arabaci, R., Sekir, U., Ozkaya, G., Demirdogen, E., Karadag, M., and Gozu, R. O. (2008). The association between snoring, daytime sleepiness and obesity in professional wrestlers. *European Journal of General Medicine*, 5(1):9–15.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the Association for Computing Machinery*, 27(11):1134–1142.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Kaiser, Ł. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, United States. Neural Information Processing Systems Foundation.

Viterbi, A. J. (2006). A personal history of the Viterbi algorithm. *IEEE Signal Processing Magazine*, 23(4):120–142.

Voulgaris, A., Ferini-Strambi, L., and Steiropoulos, P. (2020). Sleep medicine and COVID-19. Has a new era begun? *Sleep Medicine*, 73:170–176.

Wallis, C. (2019). How artificial intelligence will change medicine. *Nature*, 576:S48.

Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis*, chapter Fundamentals of Computational Auditory Scene Analysis, pages 1–44. IEEE.

Williamson, L. (2020). Obstructive sleep apnea: latest surgical advances and considerations during the COVID-19 pandemic. *The Lancet Respiratory Medicine*.

Wold, H. (1973). Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Proceedings of the Third International Symposium on Multivariate Analysis*, pages 383–407, Dayton, United States. Academic Press.

World Health Organization (2020). *Screening programmes: a short guide. Increase effectiveness, maximize benefits and minimize harm*. WHO Regional Office for Europe, Copenhagen, Denmark.

Wu, Y., Liu, L., Bae, J., Chow, K. H., Iyengar, A., Pu, C., Wei, W., Yu, L., and Zhang, Q. (2009). Demystifying learning rate policies for high accuracy training of deep neural networks. In *2019 IEEE International Conference on Big Data*, pages 1971–1980, Los Angeles, United States. IEEE.

Xie, J., Aubert, X., Long, X., Van Dijk, J., Arsenali, B., Fonseca, P., and Overeem, S. (2021). Audio-based snore detection using deep neural networks. *Computer Methods and Programs in Biomedicine*, 200:1–9.

Xu, M., Duan, L., Cai, J., Chia, L., Xu, C., and Tian, Q. (2004). HMM-based audio keyword generation. In Aizawa, K., Nakamura, Y., and Satoh, S., editors, *Advances in Multimedia Information Processing – Pacific-Rim Conference on Multimedia 2004*, volume 3333 of *Lecture Notes in Computer Science*, pages 566–574, Germany. Springer.

Yadollahi, A. and Moussavi, Z. (2009). Acoustic obstructive sleep apnea detection. In *Proceedings of the 31st Annual International Conference of the IEEE EMBS*, pages 7110–7113. IEEE.

Yaggi, H. K. and Strohl, K. P. (2010). Adult obstructive sleep apnea-hypopnea syndrome: definitions, risk factors, and pathogenesis. *Clinics in Chest Medicine*, 31(179–186).

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9:611–629.

Yatabe, K. (2020). Consistent ICA: determined BSS meets spectrogram consistency. *IEEE Signal Processing Letter*, 27:870–874.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book*. Cambridge University Engineering Department.

Young, T., Finn, L., Peppard, P. E., Szklo-Coxe, M., Austin, D., Nieto, F. J., Stubbs, R., and Hla, K. M. (2008). Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin sleep cohort. *Sleep*, 31(8):1071–1078.

Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., and Badr, S. (1993). The occurrence of sleep-disordered breathing among middle-aged adults. *The New England Journal of Medicine*, 328:1230–1235.

Young, T., Skatrud, J., and Peppard, P. E. (2004). Risk factors for obstructive sleep apnea in adults. *Journal of the American Medical Association*, 291(16):2013–2016.

Yu, K. H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2:719–731.

Yu, M., Wen, Y., Xu, L., Han, F., and Gao, X. (2020). Polysomnographic characteristics and acoustic analysis of catathrenia (nocturnal groaning). *Physiological Measurement*, 41(12):1–9.

Yu, X., Hu, D., and Xu, J. (2014). *Blind Source Separation: Theory and Applications*, chapter Introduction, pages 1–15. John Wiley & Sons Incorporated, Singapore, Singapore.

Zhou, L. and Hripcsak, G. (2007). Temporal reasoning with medical data — a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40:183–202.

Zinemanas, P., Cancela, P., and Rocamora, M. (2019). End-to-end convolutional neural networks for sound event detection in urban environments. In *24th Conference of Open Innovations Association (FRUCT)*, pages 533–539, Moscow, Russia. Finnish-Russian University Cooperation in Telecommunications.

Appendix A

Feature Extraction Techniques

A.1 Mel-frequency Cepstral Coefficients (MFCCs)

Speech representations for ASR are mainly founded on the source-filter theory of speech production put forward by the Swedish engineer Gunnar Fant in 1960 based on a previous study on the nature and structure of vowels by Chiba and Kajiyama (1941). Fant proposed that the speech signal is the response of the vocal tract *filter* to a sound *source*. This implies that the speech signal can be defined in terms of source and filter properties. The technical terms source and filter correspond to the phonetic terms phonation (i.e., the production of speech sounds) and articulation. The response of the vocal tract filter is the result of the position of the articulators (e.g., tongue, lips, teeth and palate), whereas the sound source is the modulation of airflow in the upper airway by the closing and opening movement of the vocal cords. The main characteristic of the sound source is its periodicity. That is, its fundamental frequency or pitch. These terms can often be employed interchangeably but, strictly speaking, fundamental frequency is a property of the sound stimulus, and pitch is a tonal sensation (Fant, 1960). Traditionally, the fundamental frequency has been ignored in ASR, as it does not carry relevant information to discriminate between phones (i.e., speech sounds). Therefore, only filter information is commonly used in ASR. It is worth noting that this is not true for tonal languages, like Mandarin, in which changes in pitch result in changes in meaning (Kaur et al., 2020). So, pitch is crucial to phoneme identification in those languages (Hess, 1983).

Cepstral analysis is one of the approaches to separate the source and filter in a speech signal. MFCCs (Bridle and Brown, 1974; Mermelstein, 1976) are highly effective in ASR and modelling the frequency content of audio signals (Xu et al., 2004), as they offer a very compact feature representation and are less correlated than spectral features, which facilitates their modelling. MFCCs are extracted from an audio signal as follows (Rao and Vuppala, 2014):

1. Pre-emphasis is applied to the audio signal with the aim of emphasising the higher frequencies, and improving the SNR. This is carried out by using a first order filter:

$$x'(n) = x(n) - \alpha x(n-1) \tag{A.1}$$

where $x(n)$ is the value n of the audio signal x, the filter coefficient α is typically 0.97, and x' is the pre-emphasised audio signal.

2. The pre-emphasised audio signal is split into overlapping frames and windowed. Typically, 25-ms frames with 10-ms overlap are used, since they allow a balance between time and frequency resolutions. The Hamming or Hann window is applied on the frames to narrow the signal towards the frame borders, and smooth the edges. The Hamming window is defined as:

$$\text{hamming}(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{A.2}$$

where $0 \leq n \leq N-1$, and N is the number of points in the window (Blackman and Tukey, 1958). Likewise, the Hann window is defined as:

$$\text{hann}(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) \tag{A.3}$$

3. The discrete Fourier transform (DFT) is applied to each windowed frame to compute its spectrum:

$$X(k) = \sum_{n=0}^{N-1} x(n)\exp\left(\frac{-j2\pi nk}{N}\right) \tag{A.4}$$

where $0 \leq k \leq N-1$, N is the number of points to compute the DFT, x is the audio frame, and X is its spectrum.

4. The power spectrum is passed through the Mel filter bank. It is a set of bandpass filters based on psychophysical models of the human auditory system (Stevens et al., 1937). Humans do not perceive pitch linearly, so the Mel scale is almost linearly spaced below 1 kHz, and logarithmically spaced above 1 kHz. The physical frequency in Hz is converted into the Mel scale as:

$$f_{Mel} = 2595\log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{A.5}$$

where f_{Hz} is the physical frequency in Hz, and f_{Mel} is the perceived frequency. A triangular Mel filter bank is commonly used for computing the Mel spectrum of the power spectrum by multiplying the latter by every triangular Mel weighting filter:

$$s(m) = \sum_{k=0}^{N-1} \left[|X(k)|^2 H_m(k) \right] \tag{A.6}$$

where $0 \leq m \leq M-1$, M is the number of weighting filters, $|X(k)|^2$ is the power spectrum, $H_m(k)$ is the weight of the bin k of the filter m defined as:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \text{ or } k > f(m+1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \leq f(m+1) \end{cases} \tag{A.7}$$

5. The Mel spectrum is represented on a logarithmic scale, and the discrete cosine transform (DCT) is applied to the Mel-frequency coefficients to produce a set of cepstral coefficients. The latter, unlike the former, are decorrelated, as we mentioned at the beginning of this section. MFCCs are computed as:

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(s(m)) \cos\left(\frac{\pi n (m-0.5)}{M} \right) \tag{A.8}$$

where $n = 0, 1, 2, ..., C-1$, C is the number of MFCCs, and $c(n)$ is the MFCC n. The 0th coefficient is frequently excluded, as it represents the average energy of the audio signal, which does not provide subject-specific information in the case of a speech or breathing signal. Also, given that the first few MFCCs contain most of the relevant information (i.e., the vocal tract filter), MFCC-based systems only use between 8 to 13 coefficients. As we previously noted, these offer a very compact representation of an audio frame.

6. MFCCs are static features, as they only incorporate information from a single frame. Additional information about the temporal dynamics of the audio signal is obtained by calculating the first and second derivatives of MFCCs, otherwise known as deltas and accelerations, respectively. These are computed as:

$$\Delta c_m(n) = \frac{\sum_{t=1}^{T} t \left(c_m(n+t) - c_m(n-t) \right)}{2 \sum_{t=1}^{T} t^2} \tag{A.9}$$

where $c_m(n)$ is the feature m for the frame n, and T is the number of consecutive frames used for computation. T is commonly taken as 2. Accelerations are calculated as the first derivative of delta features (Young et al., 2006).

Appendix B

Hyperparameter Tuning Experiments

In deep learning, parameters that control the learning process or define the DNN architecture but are not part of the network itself are known as hyperparameters (Moody, 2012). As discussed in Chapter 2, they include — amongst others — batch size, kernel size, number of layers, number of epochs, and learning rate. This appendix succinctly presents a set of results of the preliminary experiments conducted with the aim of tuning the key hyperparameters of the DNNs developed. Such experiments were carried out on a subset of the available data — the corpora introduced in Chapter 3. Sections B.1, B.2 and B.3 are to be read alongside Chapters 4, 5 and 6, respectively.

B.1 Classification of Sleep-disordered Breathing Events

B.1.1 Classification with Bottleneck Features from an Auditory Model

Figure B.1 presents the frame-level confusion matrices for the classification of sleep-disordered breathing events with an HMM using bottleneck features from rate maps. The matrices display the results obtained when extracting bottleneck features with rate map autoencoders trained in 10 and 60 epochs. Given that this experiment was carried out on a limited amount of sleep audio recordings, 10 epochs were enough to effectively learn from data. However, when learning from a larger dataset, a DNN typically requires a greater number of epochs to properly generalise. Keeping the number of epochs in 60, Figure B.2 shows the performance achieved when extracting bottleneck features with rate map autoencoders trained using a learning rate of 0.01 and 0.001. As expected, the best recall for all classes was obtained with the standard value commonly used in deep learning: 0.001. This learning rate was employed throughout the thesis. Lastly, Figure B.3 displays the frame-level confusion matrices for the classification of sleep-disordered breathing events

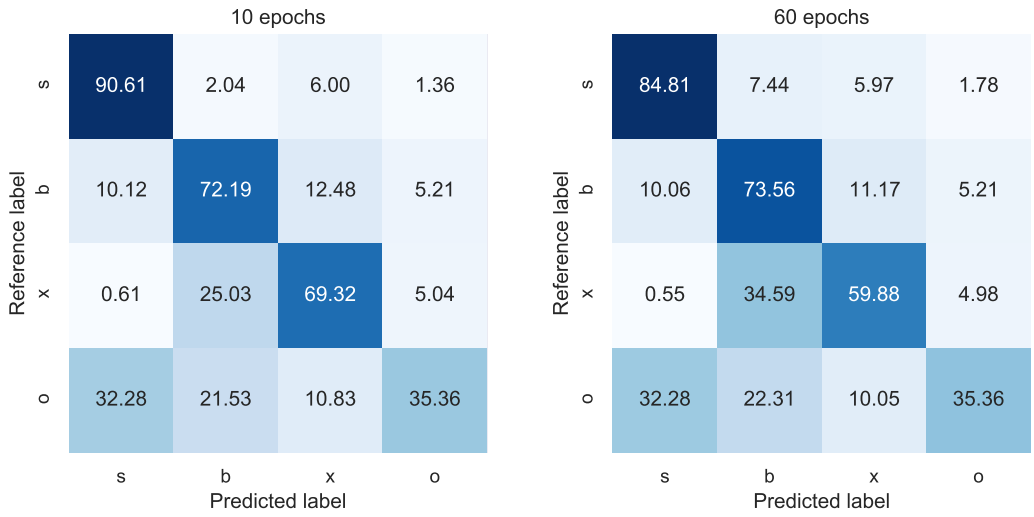


Fig. B.1 Frame-level confusion matrices for the classification of sleep-disordered breathing events with an HMM using bottleneck features from rate maps. The left and right panels display the results obtained when extracting bottleneck features with rate map autoencoders trained in 10 and 60 epochs, respectively. s: snore, b: breath, x: silence, o: other.

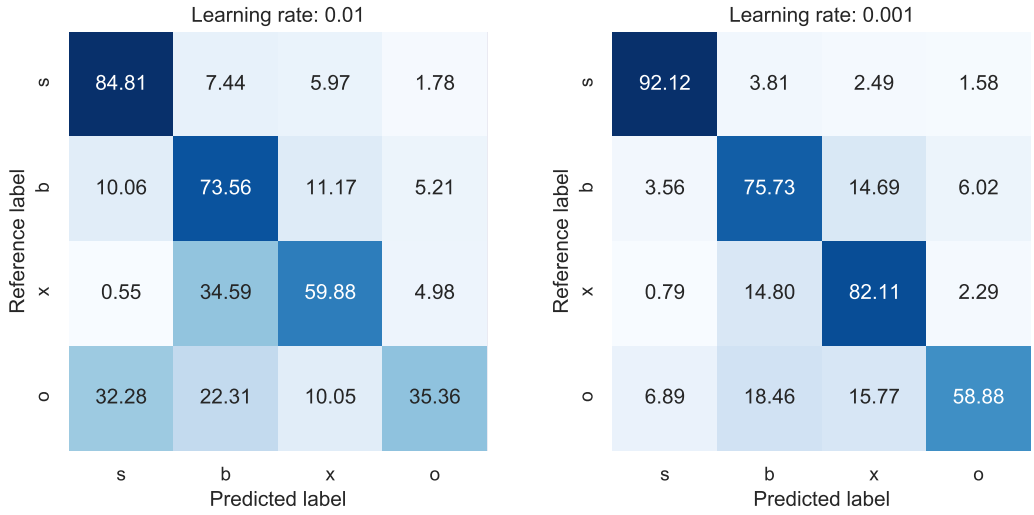


Fig. B.2 Frame-level confusion matrices for the classification of sleep-disordered breathing events with an HMM using bottleneck features from rate maps. The left and right panels display the results obtained when extracting bottleneck features with rate map autoencoders trained using a learning rate of 0.01 and 0.001, respectively. s: snore, b: breath, x: silence, o: other.

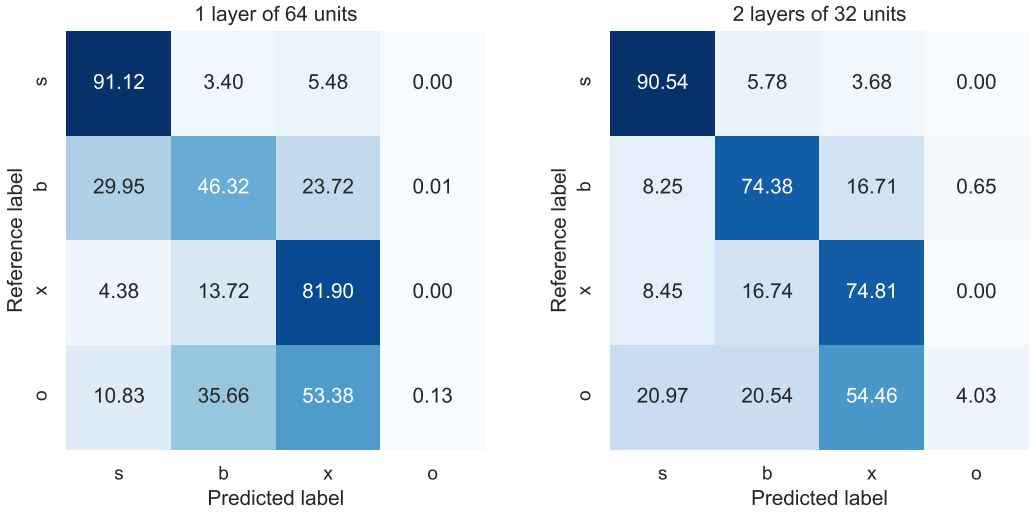


Fig. B.3 Frame-level confusion matrices for the classification of sleep-disordered breathing events with a DNN using bottleneck features from rate maps. The left and right panels display the results obtained when using a DNN with 1 hidden layer of 64 units and 2 hidden layers of 32 units, respectively. s: snore, b: breath, x: silence, o: other.

with a DNN using bottleneck features from rate maps. The matrices present the results attained when using a DNN with 1 hidden layer of 64 units and 2 hidden layers of 32 units. Although the performance of the DNN with 2 layers of 32 units on snore and silence classification was moderately lower than that of the DNN with 1 layer of 64 units, the performance on breath and 'other' classification was improved. This evidences that deep network architectures are better suited to learn complex patterns from data than shallow ones. For this reason, networks with multiple layers were developed throughout this study.

B.2 Screening for Obstructive Sleep Apnoea

B.2.1 Predicting the Presence of Apnoea-hypopnea Events

Table B.1 presents the sensitivity, specificity and AUC when screening for OSA from 1-minute rate maps making use of a CNN with a kernel size of 3×3 and 3×4. The former attained a better performance than the latter, which suggests that temporal and spectral information can be successfully captured for the screening task with a square kernel. Continuing with a kernel size of 3×3, the confusion matrices for the classification of 1-minute rate maps with no overlap and with 50% overlap using a CNN are displayed in Figure B.4. Extracting rate maps with 50% overlap achieved a slightly better performance on the classification of segments with hypopnea events in comparison to extracting them

Table B.1 Screening for OSA from 1-minute Rate Maps Using a CNN with a Kernel Size of 3×3 and 3×4

Kernel Size	**3×3**	**3×4**
Sensitivity	0.90	0.87
Specificity	0.90	0.80
AUC	0.93	0.89

with no overlap. However, the performance on the classification of the other two classes — 'segment with no apnoea-hypopnea events' and 'segment with apnoea events' — was degraded, as overlapping segments yields more testing samples, and therefore there is more room for errors. Accordingly, non-overlapping segments were employed to screen for OSA in Chapter 5. Finally, the confusion matrices for the classification of 1-minute rate maps into 2 and 3 classes with a CNN are shown in Figure B.5. Due to the acoustic similarities between some hypopneas and healthy breathing, and between some hypopneas and apnoeas, the recall of the segments with hypopnea events was low when considering 3 classes. Collapsing the classes 'segment with hypopnea events' and 'segment with apnoea events' into the class 'segment with apnoea-hypopnea events' resulted in a better overall performance thanks to a simpler learning objective (i.e., classifying data into 2 classes instead of 3). Then the systems to screen for OSA throughout Chapter 5 considered two classes.

B.3 Robustness to Bed Partner Breathing Sounds

B.3.1 Snorer Diarisation

Figure B.6 displays the confusion matrices for the snorer count estimation system tested on MFCCs using a segment length of 2 seconds and 250 ms. Even though a better performance was attained when using non-overlapping 2-second segments, as a wider temporal context was considered, making use of non-overlapping 250-ms segments allowed a better temporal resolution with a minor impact on performance. A good temporal resolution is relevant for accurately computing the time spent snoring, for example. Maintaining a segment length of 250 ms, Figure B.7 shows the confusion matrices for the snorer count estimation system tested on STFT features with no dropout and with dropout after each layer with BLSTM units. As expected, employing dropout achieved the best overall performance, since this regularisation technique successfully avoids overfitting (Srivastava et al., 2014). Lastly, Figure B.8 presents the confusion matrices for the snorer count estimation system tested on STFT features with no batch normalisation and with batch

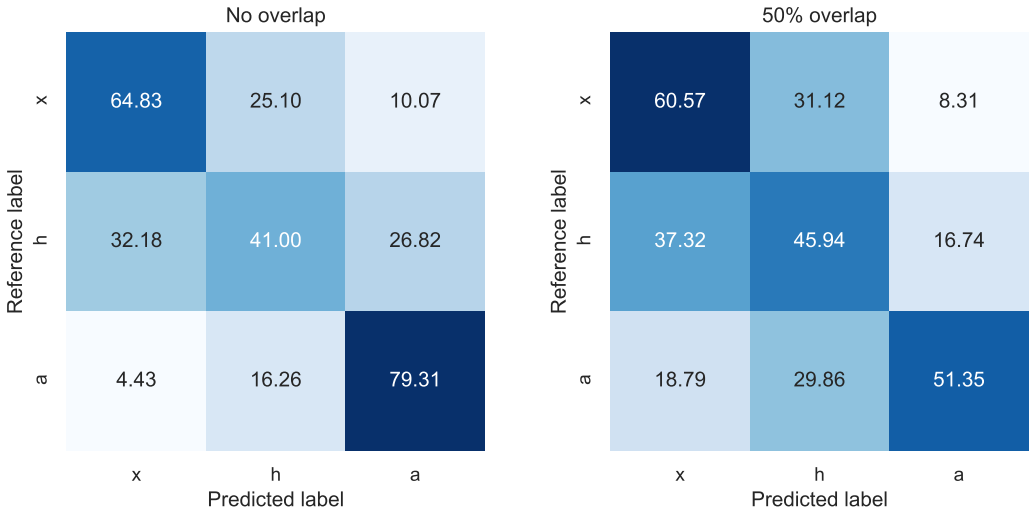


Fig. B.4 Confusion matrices for the classification of 1-minute rate maps with no overlap (left panel) and 50% overlap (right panel) using a CNN. x: segment with no apnoea-hypopnea events, h: segment with hypopnea events, a: segment with apnoea events.

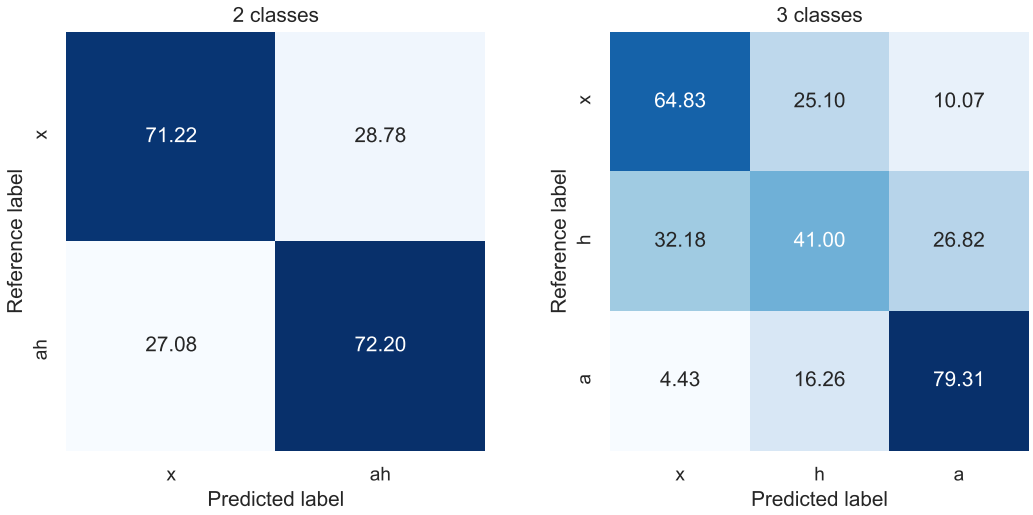


Fig. B.5 Confusion matrices for the classification of 1-minute rate maps into 2 (left panel) and 3 (right panel) classes using a CNN. x: segment with no apnoea-hypopnea events, ah: segment with apnoea-hypopnea events, h: segment with hypopnea events, a: segment with apnoea events.

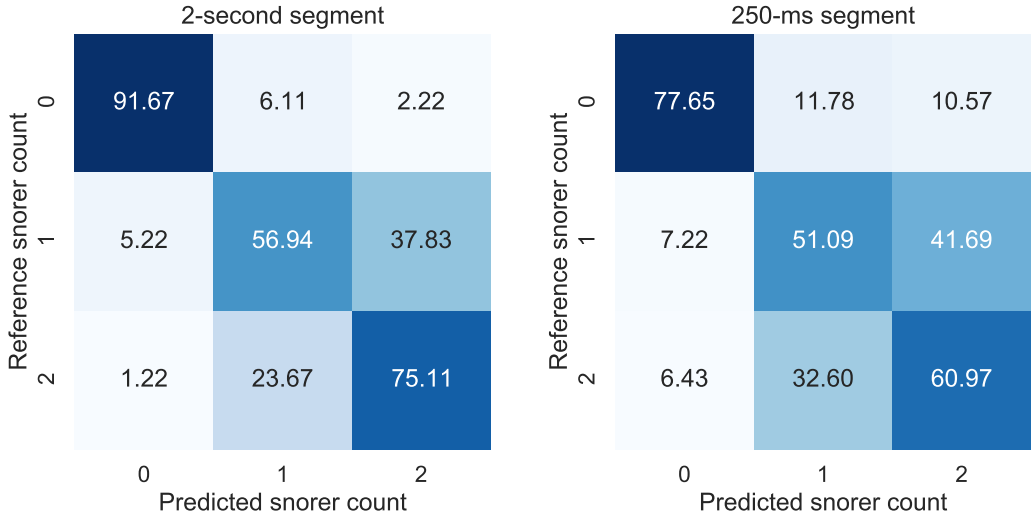


Fig. B.6 Confusion matrices for the snorer count estimation system tested on MFCCs using a segment length of 2 seconds (left panel) and 250 ms (right panel)

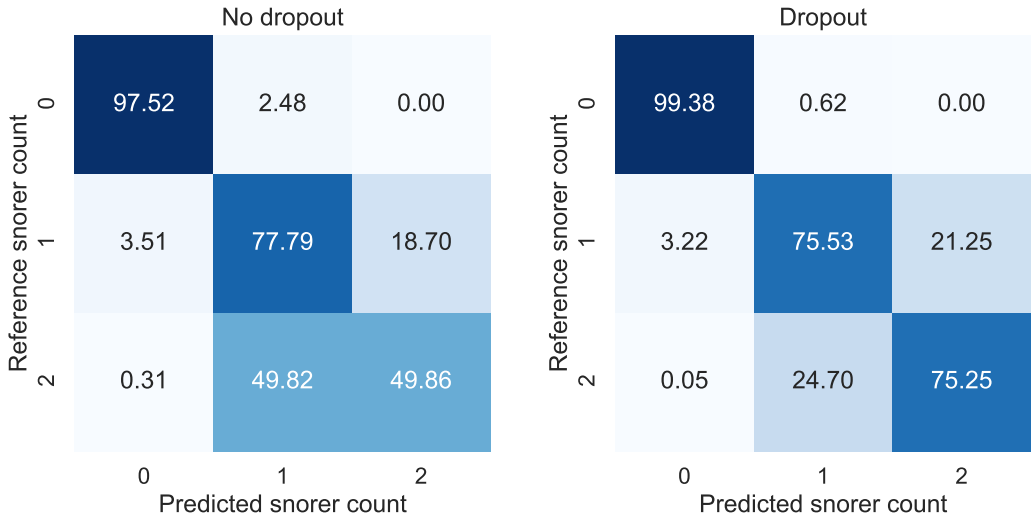


Fig. B.7 Confusion matrices for the snorer count estimation system tested on STFT features with no dropout (left panel) and with dropout (right panel) after each layer with BLSTM units

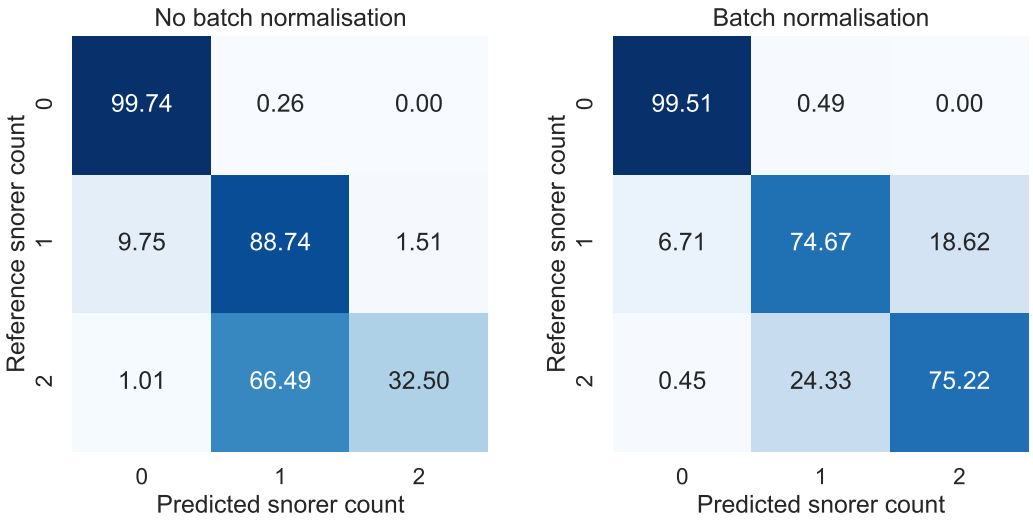


Fig. B.8 Confusion matrices for the snorer count estimation system tested on STFT features with no batch normalisation (left panel) and with batch normalisation (right panel) after each layer with BLSTM units

normalisation after each layer with BLSTM units. The performance on the classification of 2-snorer segments was considerably improved when applying batch normalisation — with a small impact on the performance on the remaining classes — in comparison to not applying it. Batch normalisation facilitates DNN training by normalising the distribution of each layer's input (Ioffe and Szegedy, 2015). For these reasons, dropout and batch normalisation were used to develop the networks in Chapter 6.